CONTROL DE VERSIONES

Elaborado por:	Carlos Alberto Gutiérrez García	**Fecha de Elaboración:**	30/05/2024	**Versión:**	**01**
Aprobado por:		**Fecha de Aprobación:**			

Historia de Modificaciones

No. de Versión	Fecha de Versión	Autor	Revisado	Aprobado	Descripción
1	30/05/2024	Carlos Alberto Gutiérrez García			Creación del Manual Técnico para el proceso de calidad de datos

Diagnóstico y Mejora Continua de la Calidad de Datos en Azure

CONTENIDO

1. INTRODUCCIÓN

El proyecto de Diagnóstico y Mejora Continua de la Calidad de Datos en Azure para Bancoomeva tiene como objetivo principal garantizar la integridad y precisión de los datos en un entorno en la nube. La solución desarrollada aprovecha las capacidades de Microsoft Azure, integrando componentes como Azure Machine Learning, Azure SQL Database, y Power BI para realizar un análisis exhaustivo y un proceso de remediación de datos.

La importancia de mantener altos estándares de calidad de datos radica en la capacidad de tomar decisiones informadas, mejorar la eficiencia operativa y cumplir con las normativas aplicables. El proceso se lleva a cabo en varias etapas, incluyendo la parametrización, análisis diagnóstico de calidad e implementación de estrategias de remediación, asegurando que los datos cumplan los criterios establecidos en las dimensiones de calidad.

2. Modelo de datos Relacional para la Gestión de Calidad de Datos en Azure

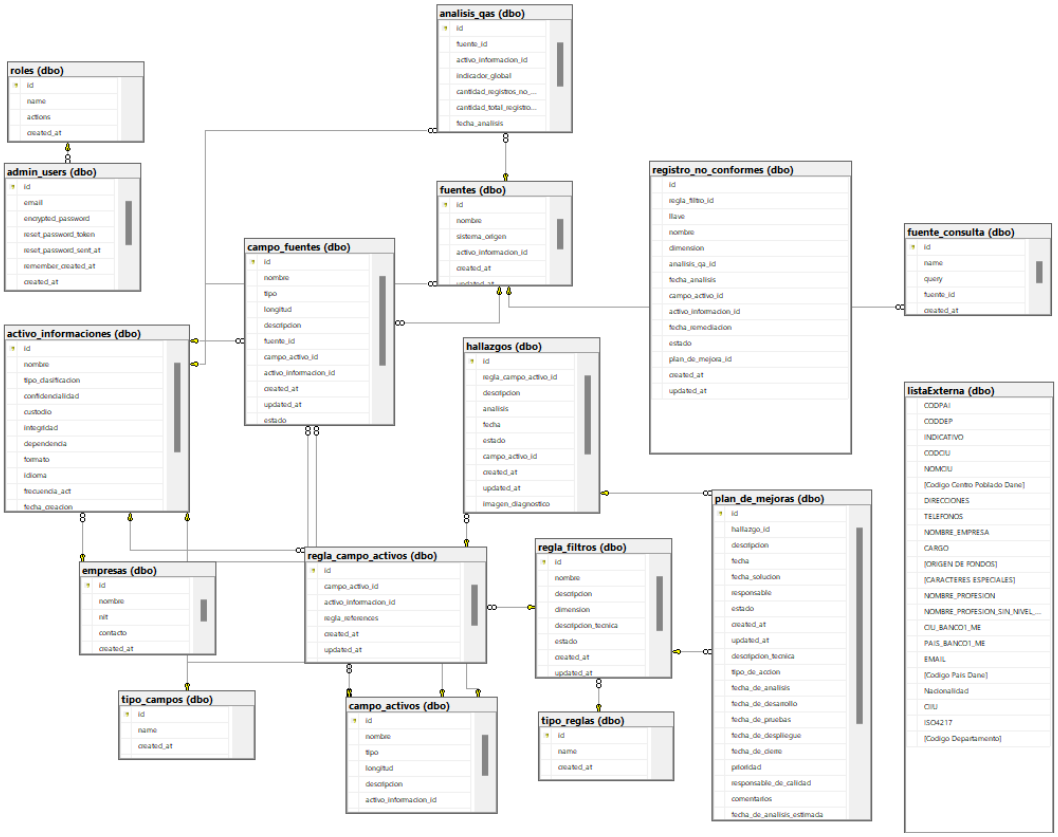

Figura 1. Modelo de datos relacional para la gestión y análisis de calidad de datos sobre el ambiente productivo

2.1 Descripción del modelo:

Este modelo de datos relacional está diseñado para gestionar y analizar la calidad de datos alojados en una base de datos SQL. El sistema permite la parametrización de diversas entidades (fuentes de datos, campos, activos de información, empresas, usuarios, etc.) a través de una aplicación. Las reglas de calidad y remediación se definen y almacenan en esta base de datos, y son procesadas por Azure Machine Learning. Los resultados del análisis, que incluyen informes de inconsistencias y registros no conformes, se almacenan nuevamente en la base de datos y se visualizan mediante Power BI.

2.2 Componentes del modelo:

Fuentes de Datos:

a. **fuentes**: Almacena información sobre las diferentes fuentes de datos.
b. **campo_fuentes**: Detalla los campos de las fuentes de datos, incluyendo sus características como tipo y longitud.

Activo de Información:

c. **activo_informaciones**: Describe los activos de datos incluyendo su tipo, confidencialidad, integridad, y otras propiedades.
d. **empresas**: Contiene información sobre las empresas que poseen o gestionan los activos de información.

Administración de Usuarios y Roles:

e. **admin_users**: Registra los usuarios administradores del sistema.
f. **roles**: Define los roles y permisos asignados a los usuarios.

Reglas de Calidad y Remediación:

g. **regla_filtros**: Define las reglas de calidad para los datos.
h. **regla_campo_activos**: Asocia las reglas de calidad con los campos específicos de los activos de información.
i. **tipo_reglas**: Clasifica los diferentes tipos de reglas de calidad.

Resultados del Análisis de Calidad:

j. **registro_no_conformes**: Almacena los registros que no cumplen con las reglas de calidad.
k. **hallazgos**: Registra los problemas de calidad identificados y su estado.
l. **plan_de_mejoras**: Contiene los planes de acción para corregir los problemas de calidad.

Análisis y Visualización:

m. **analisis_gas**: Guarda los resultados del análisis de calidad realizado por Azure Machine Learning.
n. **fuente_consulta**: Permite la generación de consultas para la visualización de datos en Power BI.
o. **lista_extena**: asegura la calidad de los datos mediante la comparación y validación de valores en el análisis de calidad.

2.3 Diagrama de Flujo

El diagrama de flujo ilustra el proceso integral de gestión de la calidad de datos en Bancoomeva utilizando los servicios de Microsoft Azure. A continuación, se describe cada uno de los pasos representados en el diagrama.

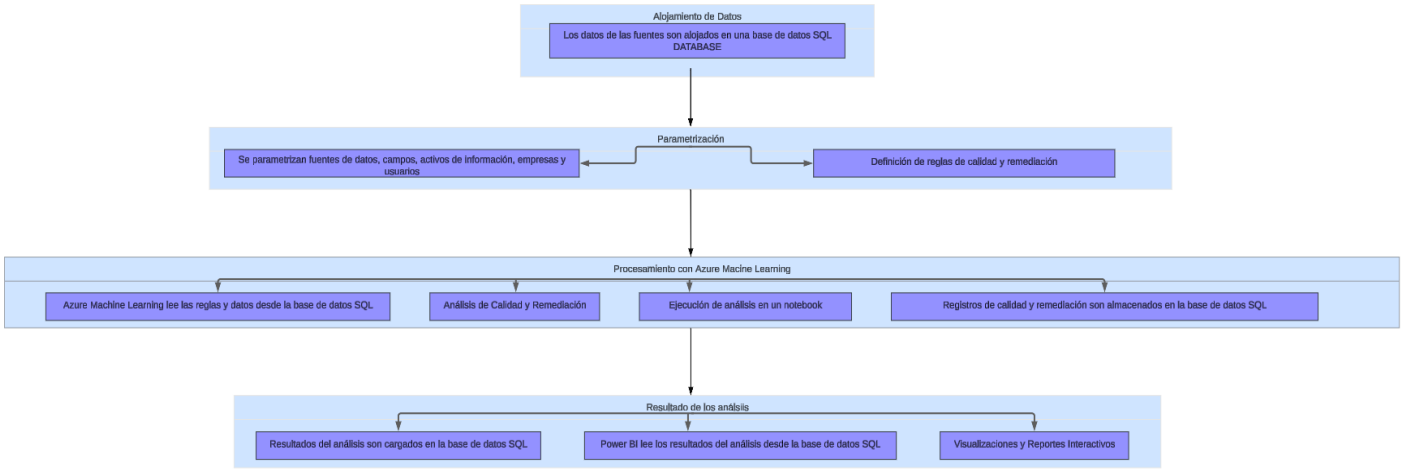

Figura 2. Proceso de diagnóstico y mejora continua

- **Alojamiento de Datos:**

  - **Los datos de las fuentes son alojados en una base de datos SQL**

  Este es el punto de inicio donde los datos provenientes de diversas fuentes son almacenados. Este almacenamiento inicial asegura que todos los datos estén centralizados y accesibles para su posterior procesamiento.

- **Parametrización:**

  - **Definición de reglas de calidad y remediación**

  En esta etapa, se definen las reglas de calidad y remediación que serán aplicadas a los datos. Estas reglas son configuradas a través de una aplicación específica.

- **Se parametrizan fuentes de datos, campos, activos de información, empresas y usuarios**

Los usuarios administradores pueden parametrizar y registrar las diversas fuentes de datos, campos específicos, activos de información, empresas y detalles de los usuarios en la base de datos SQL.

- **Procesamiento con Azure Machine Learning:**

  - **Azure Machine Learning lee las reglas y datos desde la base de datos SQL**

  Azure Machine Learning se conecta a la base de datos SQL para leer las reglas de calidad y los datos que necesitan ser analizados.

  - **Ejecución del Notebook**

  Los análisis de calidad y remediación son ejecutados en un entorno de notebook en Azure Machine Learning, utilizando los datos y reglas previamente definidos.

  - **Análisis de calidad y remediación**

  El sistema realiza el análisis de calidad, identificando inconsistencias y aplicando las reglas de remediación para mejorar la calidad de los datos.

- **Resultados del Análisis:**

  - **Resultados del análisis son cargados en la base de datos SQL**

  Una vez completado el análisis, los resultados, incluyendo informes de inconsistencias y registros no conformes, son almacenados nuevamente en la base de datos SQL.

  - **Power BI lee los resultados del análisis desde la base de datos SQL**

  Power BI se conecta a la base de datos SQL para leer los resultados del análisis de calidad.

  - **Visualizaciones y Reportes Interactivos:**

  Power BI genera visualizaciones y reportes interactivos basados en los resultados del análisis, permitiendo a los usuarios finales comprender mejor la calidad de los datos y las acciones de remediación realizadas.

3. DIMENSIONES

Tabla 1 Dimensiones del modelo multidimensional ROLAP

NUM.	TABLAS	Descripción	Campos principales
1	fuentes	Contiene los detalles de las diferentes fuentes de datos que alimentan el sistema, incluyendo el nombre de la fuente, su origen y el activo de información asociado.	id: Identificador único de la fuente, nombre: Nombre de la fuente de datos, origen: Origen de la fuente de datos, activo_informacion_id: Identificador del activo de información asociado
2	campo_fuentes	almacena información detallada sobre los campos individuales presentes en las diversas fuentes de datos. Cada registro en esta tabla representa un campo específico de una fuente de datos, incluyendo sus características, asociaciones y estado de análisis	id: Identificador único del campo, nombre: Nombre del campo, tal como se encuentra en la fuente original de datos, tipo: Tipo de dato del campo, nombre_amigable: Un nombre alternativo que pueda ser más fácil de interpretar en el contexto de reportes y análisis, estado: Este campo indica si el campo está sujeto a análisis o no.
3	campo_activo	Almacena la descripción y características de los campos dentro de las fuentes de datos, incluyendo tipo, longitud y otras propiedades	nombre: Nombre del campo, tipo: Tipo de dato del campo, longitud: Longitud del campo, descripcion: Descripción del campo, fuente_id: Identificador de la fuente de datos asociada, campo_activo_id: Identificador del campo activo asociado
4	activo_informaciones	Detalles de los activos de información, como nombre, tipo, confidencialidad, integridad, dependencia, y otros atributos relevantes.	nombre: Nombre del activo de información, tipo_identificacion: Tipo de identificación del activo, confidencialidad: Nivel de confidencialidad del activo, integridad: Nivel de integridad del activo, dependencia: Dependencia del activo con otros datos,
5	empresas	Información sobre las empresas, incluyendo nombre, NIT, y otros detalles identificativos.	nombre: Nombre de la empresa, nit: Número de identificación tributaria de la empresa, contactos: Información de contacto de la empresa
6	admin_users	Registra los usuarios administradores del sistema, incluyendo su correo electrónico, contraseña encriptada, y otros detalles de autenticación.	id: Identificador único del usuario, email: Correo electrónico del usuario, encrypted_password: Contraseña encriptada, reset_password_token: Token de restablecimiento de contraseña, reset_password_sent_at: Fecha y hora del envío del token de restablecimiento de contraseña, remember_created_at: Fecha y hora de creación del token de recordar sesión
7	lista_externa	Esta dimensión contiene una lista de valores utilizados para realizar comparaciones en el análisis de datos. Proporciona referencias necesarias para validar y categorizar la información procesada. Los campos que contiene son diversos y abarcan códigos geográficos, identificadores de entidades, y otras referencias que se utilizan en las reglas de calidad.	CCODPAI: Código del país, CODDEP: Código del departamento, INDICATIVO: Indicativo telefónico, CODCIU: Código de la ciudad, NOMCIU: Nombre de la ciudad, Codigo Centro Poblado Dane: Código del centro poblado según el DANE, DIRECCIONES: Dirección, TELEFONOS: Número de teléfono, NOMBRE_EMPRESA: Nombre de la empresa, CARGO: Cargo en la empresa, ORIGEN DE FONDOS: Origen de los fondos, CARACTERES ESPECIALES: Caracteres especiales utilizados, NOMBRE_PROFESION: Nombre de la profesión, NOMBRE_PROFESION_SIN_NIVEL_ACADEMICO: Nombre de la profesión sin nivel académico, CIU_BANCO1_ME: Código de identificación de la cuenta bancaria 1 en moneda extranjera, PAIS_BANCO1_ME: País de la cuenta bancaria 1

NUM.	TABLAS	Descripción	Campos principales
			en moneda extranjera, EMAIL: Correo electrónico, Codigo Pais Dane: Código del país según el DANE, Nacionalidad: Nacionalidad, CIIU: Código de la Clasificación Industrial Internacional Uniforme, ISO4217: Código de la moneda según el estándar ISO 4217, Codigo Departamento: Código del departamento

4. FACT

En el proceso se crean las siguientes Fact:

Tabla 2. FACT dentro del modelo multidimensional ROLAP

NUM.	TABLAS	Descripción	Campos principales
1	registro_no_conformes	Registra los datos que no cumplen con las reglas de calidad definidas. Cada registro contiene información sobre el filtro fallido, la fuente de datos, la dimensión afectada, y el detalle del error identificado.	id: Identificador único del registro, regla_filtro_id: Identificador de la regla de filtro aplicada, fuente_id: Identificador de la fuente de datos, dimension_id: Identificador de la dimensión afectada, campo_fuente_id: Identificador del campo de la fuente afectado, activo_informacion_id: Identificador del activo de información afectado, detalle_error: Descripción detallada del error encontrado, creado_at: Fecha y hora de creación del registro, actualizado_at: Fecha y hora de la última actualización del registro
2	hallazgos	Almacena los problemas de calidad de datos identificados, incluyendo una descripción del problema, el estado actual, y el ID del activo de datos afectado.	id: Identificador único del hallazgo, campo_activo_id: Identificador del campo activo afectado, descripcion: Descripción del problema de calidad identificado, estado: Estado actual del hallazgo (e.g., Abierto, En Proceso, Cerrado), creado_at: Fecha y hora de creación del hallazgo, actualizado_at: Fecha y hora de la última actualización del hallazgo, imagen_diagnostico: Referencia a una imagen o archivo relacionado con el diagnóstico del hallazgo
3	plan_de_mejoras	Contiene los planes de acción para corregir los problemas de calidad de datos identificados en los hallazgos. Cada registro incluye detalles de la solución propuesta y su estado.	id: Identificador único del plan de mejora, hallazgo_id: Identificador del hallazgo relacionado, descripcion: Descripción de la solución propuesta, fecha_solucion: Fecha en que se implementó la solución, estado: Estado actual del plan de mejora (e.g., Planificado, En Proceso, Completado), creado_at: Fecha y hora de creación del plan de mejora, actualizado_at: Fecha y hora de la última actualización del plan de mejora
4	analisis_gas	Registra los resultados del análisis de calidad de datos, incluyendo índices de calidad y detalles sobre los datos analizados	id: Identificador único del análisis, fuente_id: Identificador de la fuente de datos analizada, activo_informacion_id: Identificador del activo de información analizado, indice_calidad_global: Índice de calidad global calculado, cantidad_registros_incorrectos: Número de registros incorrectos encontrados, cantidad_registros_totales: Número total de

NUM.	TABLAS	Descripción	Campos principales
			registros analizados, fecha_analisis: Fecha en que se realizó el análisis

5. Generación de Vistas

Tabla 3. Generación de vista normalizada que integra todas las fuentes

NUM.	TABLAS	Descripción	Campos principales
1	FuenteConcatenada	La vista se encarga de unificar y homogenizar las diferentes fuentes de datos en una estructura común para facilitar el análisis en el proceso de diagnóstico y calidad de datos. Esta vista integra datos de múltiples fuentes, alineando los campos comunes y asegurando una representación consistente de la información. Al normalizar los datos de diversas fuentes, FuenteConcatenada permite una evaluación más eficiente y precisa de la calidad de los datos, identificando y remediando problemas de manera efectiva.	id: Identificador único del registro, regla_filtro_id: Identificador de la regla de filtro aplicada, fuente_id: Identificador de la fuente de datos, dimension_id: Identificador de la dimensión afectada, campo_fuente_id: Identificador del campo de la fuente afectado, activo_informacion_id: Identificador del activo de información afectado, detalle_error: Descripción detallada del error encontrado, creado_at: Fecha y hora de creación del registro, actualizado_at: Fecha y hora de la última actualización del registro

6. Gestión de Entornos: Productivo vs Desarrollo

En el contexto del proyecto de Diagnóstico y Mejora Continua de la Calidad de Datos en Azure para Bancoomeva, se utilizan dos entornos diferentes para asegurar la integridad y precisión de los datos: el entorno productivo y el entorno de desarrollo. Ambos entornos contienen las mismas tablas y relaciones, pero se diferencian principalmente en el esquema de la base de datos utilizado. A continuación, se describen las características y propósitos de cada entorno:

Entorno Productivo (.dbo)

El entorno productivo es el entorno operativo principal donde se ejecutan las operaciones mensuales para el diagnóstico y se procesan los datos en tiempo real. Este entorno utiliza el esquema .dbo para sus tablas y relaciones. (Ver Figura 1)

- **Propósito:** Gestionar y procesar datos reales de la organización, asegurando la integridad, seguridad y disponibilidad de los datos.

Características:

- **Esquema:** .dbo
- **Usuarios:** Utilizado por los usuarios finales y sistemas en producción.

- **Datos:** Contiene datos reales y actualizados.
- **Seguridad:** Implementa medidas de seguridad estrictas para proteger los datos sensibles.
- **Monitoreo:** Constantemente monitoreado para garantizar el rendimiento y la disponibilidad.
- **Copias de Seguridad:** Se realizan copias de seguridad regulares para evitar pérdida de datos.

Entorno de Desarrollo (.TestSchema)

El entorno de desarrollo es un entorno aislado utilizado para pruebas y desarrollo. Este entorno utiliza el esquema .TestSchema para sus tablas y relaciones.

- **Propósito:** Permitir a los desarrolladores y equipos de pruebas experimentar y validar cambios sin afectar el entorno productivo.
- **Características:**
  - **Esquema:** .TestSchema
  - **Usuarios:** Utilizado por desarrolladores, testers y otros miembros del equipo de desarrollo.
  - **Datos:** Contiene datos de prueba que pueden ser ficticios o una copia anonimizada de los datos reales.
  - **Flexibilidad:** Mayor flexibilidad para realizar cambios y pruebas sin riesgos para el entorno productivo.
  - **Integración Continua:** Se integra con sistemas de integración continua para automatizar pruebas y despliegues.
  - **Seguridad:** Aunque menos estricta que en producción, aún se implementan medidas de seguridad para proteger datos sensibles de prueba.

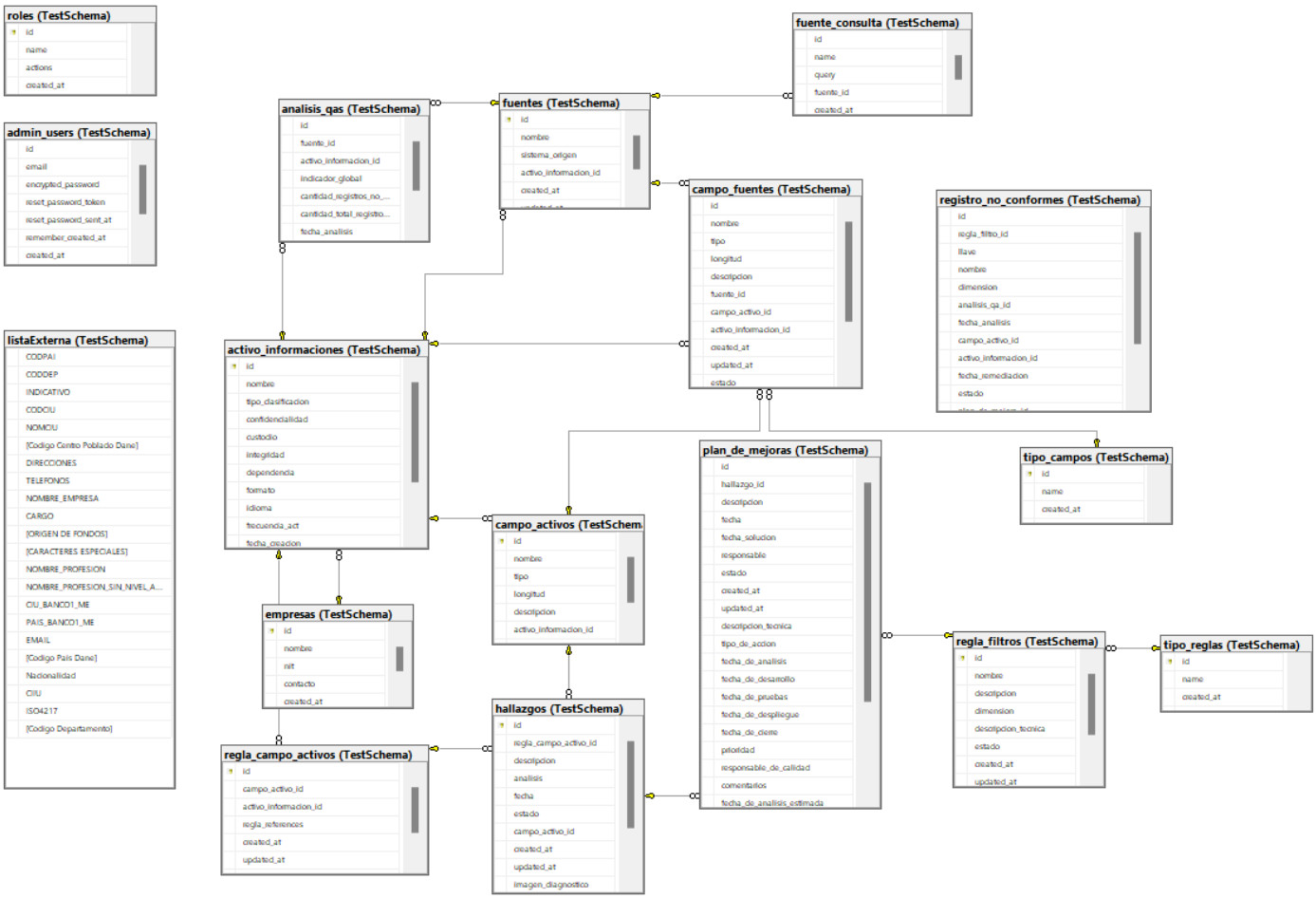

Figura 3. Modelo de datos relacional para la gestión y análisis de calidad de datos sobre el ambiente de desarrollo

Comparación de Tablas y Relaciones

Ambos entornos contienen las mismas tablas y relaciones, asegurando que cualquier cambio realizado y probado en el entorno de desarrollo sea consistente con el entorno productivo. Aquí están las tablas y relaciones presentes en ambos esquemas:

- **Tablas:**
  - roles
  - admin_users
  - activo_informaciones
  - empresas
  - campo_fuentes
  - fuentes
  - analisis_gas
  - registro_no_conformes
  - fuente_consulta
  - plan_de_mejoras
  - regla_filtros

- tipo_reglas
- tipo_campos
- campo_activos
- hallazgos
- listaExterna

Relaciones:

- Las tablas están interconectadas por claves foráneas que aseguran la integridad referencial entre los diferentes componentes del sistema.

7. Diseño ETL:

El diseño del proceso ETL (Extract, Transform, Load) para el proyecto de calidad y mejoramiento continuo de los datos en Bancoomeva es fundamental para el proceso de generación de inconformidades y remediación sobre las inconsistencias detectadas mediante la definición de planes de mejora. Por esta razón, a continuación, se ilustra el diseño de estos procesos:

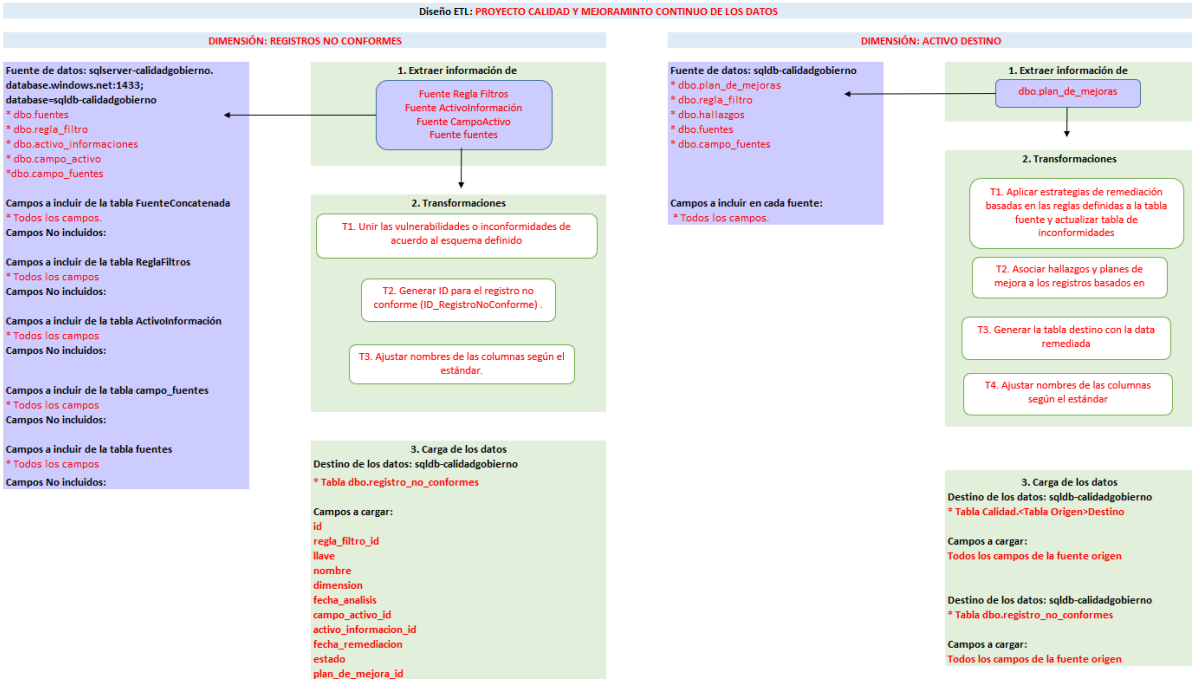

Figura 4. Diseño ETL para la dimensión de inconformidades y salida de remediación

Conclusiones

Visibilidad y Control: El uso de herramientas como Azure Machine Learning y Power BI proporciona visibilidad completa sobre la calidad de los datos. Los informes interactivos y las visualizaciones detalladas permiten a los usuarios monitorear el estado de los datos, identificar rápidamente las áreas problemáticas y tomar medidas correctivas oportunas.

Flexibilidad y Adaptabilidad: El diseño modular y parametrizable del proceso ETL facilita la adaptación a nuevas fuentes de datos y cambios en las reglas de calidad. Esta flexibilidad es crucial para mantener la relevancia y efectividad del sistema de calidad de datos en un entorno empresarial dinámico y en constante evolución.

Integridad y Precisión de los Datos: El manual técnico detalla un proceso ETL que asegura la integridad y precisión de los datos a través de un enfoque sistemático. Este proceso es esencial para mantener altos estándares de calidad de datos en Bancoomeva, permitiendo una toma de decisiones informada y confiable.

Mejora Continua: El enfoque en la mejora continua de la calidad de datos, respaldado por la definición y aplicación de planes de mejora, garantiza que Bancoomeva no solo resuelva las inconsistencias actuales, sino que también fortalezca sus procesos para prevenir problemas futuros. Esto contribuye a un ciclo de mejora continua que eleva la calidad y confiabilidad de los datos a largo plazo.