|  |
| --- |
|  |
| **INTEGRACION DE DATOS Data Warehouse** |
| *Versión <1.0>* |
| ***Dirección General de Tecnologías de la Información y las Comunicaciones***  ***Datos, BI y Analítica*** |
|  |
| aconcheso[@energia.gob.ar](mailto:sramirez@minem.gob.ar)  Paseo Colón 189| Piso 7 | C1063ACC |Capital Federal |

# Tabla de Contenido

[Tabla de Contenido 2](#_Toc56097214)

[1. Propósito del Documento 3](#_Toc56097215)

[2. Confidencialidad de la Información 3](#_Toc56097216)

[3. Glosario 3](#_Toc56097217)

[4. Referencias 3](#_Toc56097218)

[5. Descripción General 3](#_Toc56097219)

[5.1. Definiciones Generales 3](#_Toc56097220)

[5.1.1. Actores del sistema 3](#_Toc56097221)

[6. Arquitectura 3](#_Toc56097222)

[6.1. definiciones 3](#_Toc56097223)

[6.1.1. Area de Ensayo o Staging 4](#_Toc56097224)

[6.1.2. Calidad de datos 4](#_Toc56097225)

[6.1.3. Manejo de datos maestros y almacén de datos normalizados (NDS) 4](#_Toc56097226)

[6.1.4. Almacén de datos dimensionales (DDS) 4](#_Toc56097227)

[6.1.5. Base de datos multidimensional (MDB) 4](#_Toc56097228)

[6.1.6. Herramientas de usuarios y base de datos multidimensional (MDB) 4](#_Toc56097229)

[7. Control del Documento 5](#_Toc56097230)

[7.1. Ubicación 5](#_Toc56097231)

[7.2. Versiones 5](#_Toc56097232)

# Propósito del Documento

El objetivo de este documento es proveer una visión del contexto general de la información existente y una línea base de validación del contenido de las estructuras de datos actuales, para poder realizar la unificación de estos en un data warehouse ministerial y tener una sola versión de la verdad.

En este documento se detalla un marco de calidad completo que cubra todas fases del ciclo de vida de los procesos de recolección, limpieza, procesamiento y análisis de información existente en los distintos sistemas del sector, para la unificación y centralización de información, así como el modelo de datos relacional, cubos multidimensionales, diseño de tableros de control, rutinas de validación de reglas de negocio y de detección de valores atípicos (outliers) en la información suministrada, para poder tener la calidad de datos requerida.

# Confidencialidad de la Información

La información contenida en este documento es considerada confidencial. La Dirección de Informática mantendrá en reserva dicha información, debiendo contar con la expresa autorización del Cliente para su divulgación en forma parcial o total.

# Glosario

|  |  |
| --- | --- |
| **Término /Abreviatura** | **Definición** |
|  |  |
|  |  |
|  |  |

# Referencias

|  |  |
| --- | --- |
| **Nombre del Documento** | **Ubicación** |
|  |  |
|  |  |

# Descripción General

## Definiciones Generales

### Actores del sistema

* Secretarías, subsecretarías y direcciones del ministerio

# Arquitectura

## definiciones

Se definirá un almacén de datos (varios Data Mart o Data Warehouse departamental) con la finalidad de contener toda la información requerida, siendo éste el único lugar al que se conectarán los tableros de control.

Estará basado en el modelo dimensional que permite visualizar los datos en un espacio usualmente llamado cubo o hipercubo y se define mediante dimensiones y hechos usando el concepto definido por Ralph Kimball ya que al requerir un enfoque de abajo a arriba o bottom-up, permite un sistema fácilmente explotable y entendible por el usuario y de rápido desarrollo.

El o los Data Mart será alimentado de un almacén de datos normalizado (NDS) que tendrá información validada y corregida, proveniente del procesamiento de las distintas fuetes de datos que existen en la actualidad, con los controles de calidad de información correspondientes.

Lo antedicho se realizará de la siguiente manera:

### Area de Ensayo o Staging

Se realizará un proceso de ETL (Extracción, transformación y carga de datos) que dejará una copia de los datos del sistema de origen en un almacenamiento temporal (Staging) de manera de no interferir con el sistema productivo y sin realizar ninguna transformación a priori, como se muestra en la Figura 1 - Modelo de Data Warehouse resaltado en color celeste.

### Calidad de datos

Luego de completado el paso anterior se procederá a realizar el control de calidad de los datos almacenados en el área de staging, según las reglas de negocios definidas por los usuarios del sistema, junto con un sistema de manejo de sinónimos y reglas de lógica difusa que permitirá corregir de manera automática ciertos errores comunes como ser abreviaturas o errores de tipeo, y se generará un log de auditoría con los procesos realizados, además de informar los datos que no puedan ser validados, para que los usuarios corrijan en el sistema de origen. Todo este proceso será manejado por el sistema de control de calidad de datos y un ETL que dejará los datos validados en el almacenamiento normalizado o NDS que no podrán ser modificados por otros procesos distintos a los aquí definidos, según se ve en colores verde y anaranjado en la Figura 1 - Modelo de Data Warehouse.

### Manejo de datos maestros y almacén de datos normalizados (NDS)

Dada la existencia de tablas que son genéricas para todos los sistemas, como ser, geografía (provincias, partidos/departamentos, localidades y municipios), domicilios, empresas, etc., se crearán tablas maestras para contar con un solo lugar de consulta de dicha información compartida por todos los sistemas. Estas tablas formarán parte del NDS y serán las que alimenten el data wharehouse entre otras, en color violeta en la Figura 1 - Modelo de Data Warehouse.

### Almacén de datos dimensionales (DDS)

El DDS es otra base de datos que almacena información de forma diferente al NDS de los sistemas transaccionales o OLTP. La razón de pasar los datos del NDS al DDS y luego consultar el DDS en lugar de consultar el NDS directamente es que en un DDS los datos se organizan en un formato dimensional que es más adecuado y rápido para el análisis, dado el volumen de información y la menor cantidad de relaciones necesarias entre tablas. Esto se encuentra representado en la Figura 1 - Modelo de Data Warehouse en color morado.

### Base de datos multidimensional (MDB)

Dada la cantidad de información que pueden contener algunas tablas se requiere de una base de datos multidimensional, que cargará los datos necesarios del DDS en dicha base (MDB), conocida también como cubo multidimensional, para mejorar la capacidad de respuesta de los sistemas que así lo requieran, manteniendo agregados precalculados y encontrándose representados en color gris en la Figura 1 - Modelo de Data Warehouse.

### Herramientas de usuarios y base de datos multidimensional (MDB)

Los usuarios utilizan varias herramientas, como generadores de informes, herramientas de consulta SQL y análisis de información, para recuperar y analizar los datos en el DDS y de los cubos multidimensionales, y poder realizar visualizaciones dinámicas, minería de datos y ciencia de datos, según las necesidades de cada sector.



Figura 1 - Modelo de Data Warehouse

# Control del Documento

## Ubicación

|  |  |
| --- | --- |
| **Ubicación** | https://github.com/BI-MINPROD/DataWarehouse |
| **Archivo** | MINPROD-DA-BI-AN |

## Versiones

|  |  |  |  |
| --- | --- | --- | --- |
| **Fecha** | **Autor** | **Versión** | **Descripción** |
| 12/11/2020 | Andrés G. Concheso | 1.0 | Versión inicial |