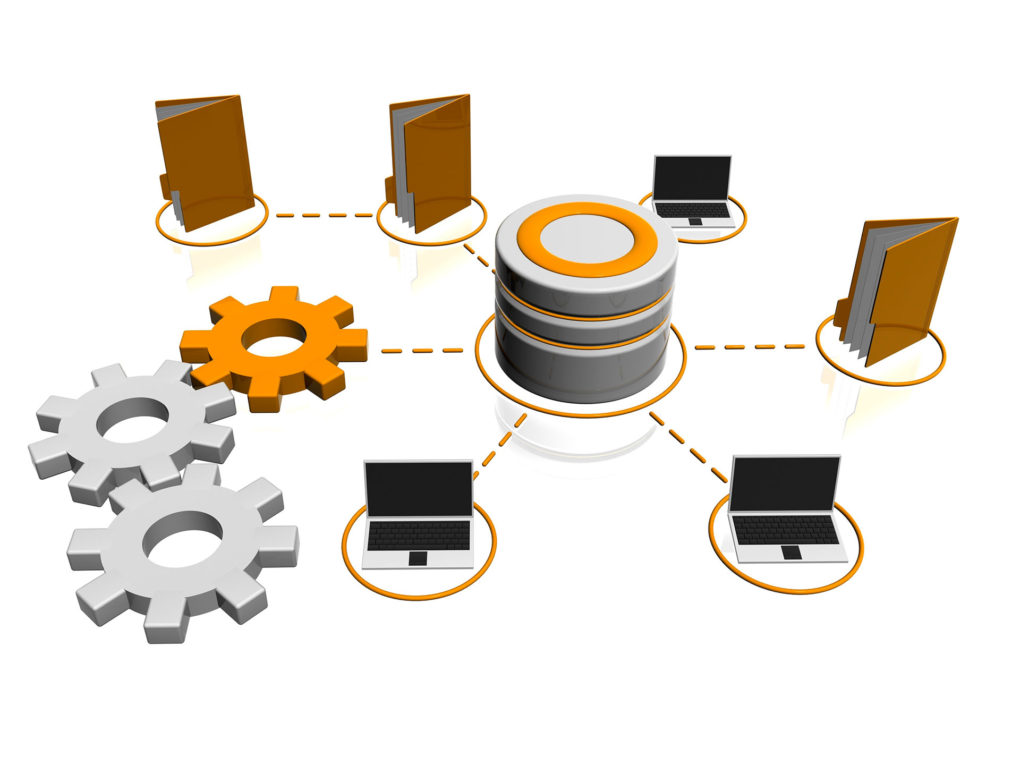


**ALMACÉN DE DATOS PARA EL ANÁLISIS DEL IMPACTO AMBIENTAL Y EL CONSUMO ENERGÉTICO DERIVADOS DE LA ACTIVIDAD ECONÓMICA**



**EDUARDO MORA GONZÁLEZ**

Contenido

[1. Introducción 3](#_Toc87197760)

[2. Análisis de los requerimientos 3](#_Toc87197761)

[3. Análisis de las fuentes de datos 5](#_Toc87197762)

[3.1. Estimación de volumetría 8](#_Toc87197763)

[4. Análisis funcional 9](#_Toc87197764)

[5. Diseño del modelo conceptual, lógico y físico del almacén de datos 11](#_Toc87197765)

[5.1. Diseño conceptual 11](#_Toc87197766)

[5.2. Diseño lógico 15](#_Toc87197767)

[5.3. Diseño físico 17](#_Toc87197768)

[5.3.1. Dimensiones 17](#_Toc87197769)

[5.3.2. Tablas de hechos 18](#_Toc87197770)

# Introducción

Actualmente, nuestra sociedad hace un uso irracional de la energía, por esta razón, se pretende diseñar un Almacén de datos para el análisis del impacto ambiental y consumo energético derivados de la actividad económica.

# Análisis de los requerimientos

La necesidad principal de la organización encargada del análisis del impacto ambiental y del consumo energético es disponer de la información integrada para su análisis y su posterior difusión mediante las herramientas de inteligencia de negocio. Estas ayudarán a facilitar la toma de decisiones a todos los usuarios potenciales para garantizar el cumplimiento, entre otros, de los siguientes objetivos:

* Analizar el balance energético mundial de los países productores y consumidores de energía.
* Analizar el balance de producción de energía no contaminante y contaminante.
* Porcentajes de objetivos de Desarrollo Sostenibles conseguidos.
* Contrastar las inversiones realizadas (relativas al ambiente) con la consecución de objetivos de Desarrollo Sostenibles.

A continuación, se indica la información necesaria identificada para analizarlo desde diferentes perspectivas:

* por fecha,
* por país,
* por categoría de energía (producto),
* por población,
* por nivel de industrialización,
* por objetivo (Desarrollo Sostenibles) o
* por unidad de medida.

Si se tiene en cuenta toda esta información, el sistema podrá responder a múltiples preguntas y de esta manera conseguirá cubrir las necesidades de los usuarios potenciales.

A continuación, se indican de forma específica las preguntas que, como mínimo, el sistema debe ser capaz de responder:

* Análisis de los países que producen más energía y los que consumen más.
* Análisis del promedio de energía (contaminante o no) produce cada país.
* Determinar si las inversiones realizadas han causado una mejora en la creación de energía verde.
* Análisis de cómo se están cumpliendo los objetivos de Desarrollo Sostenible y cual tiene mayor y menor cumplimiento.
* Relación entre el crecimiento de las ciudades y el gasto de energía.
* Determinación de que años se ha producido más energía verde y cual más energía contaminante.
* Relacionar la producción de energía y el estado de los ecosistemas protegidos.
* Determinar los países menos y más ecológicos.

# Análisis de las fuentes de datos

En este apartado se pretende revisar las fuentes de datos proporcionadas, qué tipo de información contienen, cuál es su formato y qué datos deben ser cargados:

1. **02002.xlsx:** Encuesta del gasto de la industria en protección ambiental. Serie 2008-2018. Evolución de la inversión en protección ambiental por tipo de equipo e instalación, ámbito medioambiental y sector de actividad económica. A continuación, se analizan los campos del fichero .xlsx (Excel):

|  |  |  |
| --- | --- | --- |
| **Nombre campo** | **Tipo** | **Ejemplo** |
| Periodo | Numérico | 2018 |
| Sector de actividad económica | Texto | «C. industria manufacturera» |
| Tipo de equipo e instalación | Texto | «INVERSIÓN EN EQUIPOS E INSTALACIONES INDEPENDIENTES» |
| Ámbito medioambiental | Texto | «Protección del aire y el clima» |
| Comunidad autónoma | Texto | «Asturias, Principado de» |
| Inversión | Numérico | 6.280.443 |

Observaciones:

* Las columnas desde la E hasta la U corresponden a la inversión de cada comunidad autónoma.
* La columna V contiene el agregado de Total nacional.
* La unidad de la inversión son euros.

Total de registros: 154 registros.

1. **Countries.json:** Lista de todos los países. A continuación, se analizan los campos del fichero:

|  |  |  |
| --- | --- | --- |
| **Nombre campo** | **Tipo** | **Ejemplo** |
| Nombre | Texto | Cuba |
| Iso2 | Texto | CU |
| Iso3 | Texto | CUB |
| Código teléfono | Numérico | 53 |

Observaciones:

* Los 3 primeros campos contienen el nombre del país, pero en diferentes idiomas: español, inglés y francés.

Total de registros: 246 registros.

1. **DataGeneric.xml:** Este conjunto de datos presenta tendencias en las cantidades de residuos municipales (incluidos los residuos domésticos) y el método de tratamiento y eliminación utilizado por los distintos países.

A continuación, se analizan los campos del fichero:

|  |  |  |
| --- | --- | --- |
| **Nombre campo** | **Tipo** | **Ejemplo** |
| Country | Texto | COL |
| Var | Texto | TREATMENT |
| Time | Numérico | 2005 |
| Obsvalue | Numérico | 7583.09 |

Observaciones:

* Para ver las variables de forma correcta y entender la estructura se ha mirado la página de donde se han extraído los datos: <https://stats.oecd.org/viewhtml.aspx?datasetcode=MUNW&lang=en>
* Algunos datos que no tienen valores se han puesto a 0 en el xml

Total de registros: 48 por cada variable. Y tenemos el siguiente número de variables posibles:

* Country: 48.
* Variables: 36.
* Years: 30.

1. **Env\_bio.tvs:** conjunto de áreas de conservación de la biodiversidad. A continuación, se analizan los campos del fichero:

|  |  |  |
| --- | --- | --- |
| **Nombre campo** | **Tipo** | **Ejemplo** |
| AreaProt | Texto | AREA\_KM2 |
| Geo | Texto | AT |
| Año | Numérico | 83944 |

Observaciones:

* Desde la tercera columna hasta el final (separando Areaprot y Geo como columnas independientes) se refieren al área que ocupa la zona en cada año.
* Existen 20 datos nulos.

Total de registros: 120

1. **ODS.xlsx:** Objetivos de Desarrollo Sostenible y su relación con los ámbitos medioambientales. Este fichero contiene dos pestañas
   1. **ODS:**  contiene los siguientes campos:

|  |  |  |
| --- | --- | --- |
| **Nombre campo** | **Tipo** | **Ejemplo** |
| Objetivo | Numérico | 10 |
| Nombre | Texto | Reducción de las desigualdades |
| Descripción | Texto | Reducir la desigualdad en y entre los países |

Total de registros: 17

* 1. **Ambito\_VAR\_Flow-ODS:** contiene los siguientes campos:

|  |  |  |
| --- | --- | --- |
| **Nombre campo** | **Tipo** | **Ejemplo** |
| Codigo | Texto | BULKY |
| Ambito/VAR/Flow | Texto | Bulky waste |
| ODS principal | Numérico | 9 |

Observaciones:

* El ultimo campo hace referencia al campo Objetivo de la pestaña ODS.

Total de registros: 50

1. **WorldEnergyBalancesHighlights\_final.xlsx:** sobre este fichero solo se analizará la pestaña "TimeSeries\_1971-2019" que contiene variables relevantes del balance energético mundial. A continuación, se analizan los campos del fichero:

|  |  |  |
| --- | --- | --- |
| **Nombre campo** | **Tipo** | **Ejemplo** |
| Country | Texto | Australia |
| Product | Texto | Nuclear |
| Flow | Texto | Imports (ktoe) |
| Año | Numérico | 0 |

Observaciones:

* Por cada país (Country) existen diferentes productos (Product), y por cada producto diferentes campos (Flow).
* Las columnas desde la G hasta la BC corresponden al número de Ktoe por cada año.
* Existen valores sin datos.

Total de registros: 6049

## Estimación de volumetría

Una posible estimación del volumen de datos del almacén para la carga inicial de los datos sería la siguiente:

|  |  |  |  |
| --- | --- | --- | --- |
| FICHERO | REGISTROS | VALORES | DATOS |
| 02002.xlsx | 154 | 22 | 3388 |
| Countries.json | 246 | 6 | 1476 |
| DataGeneric.xml \* | 48 | 31 | 53568 |
| Env\_bio.tvs | 120 | 10 | 1080 |
| ODS.xlsx 🡪 ODS | 17 | 3 | 51 |
| ODS.xlsx 🡪 Ambito\_VAR\_Flow-ODS | 50 | 3 | 150 |
| WorldEnergyBalancesHighlights\_final.xlsx | 6049 | 55 | 332640 |
| TOTAL | **6684** | **130** | **392353** |

\*Son 31 valores contemplando el país y cada uno de los años, pero existen 36 tipos diferentes para cada una de las combinaciones de los parámetros anteriores.

# Análisis funcional

A continuación, se describen algunos de los requerimientos funcionales para el diseño de una factoría de información para la organización, teniendo en cuenta las consideraciones del enunciado:

|  |  |  |  |
| --- | --- | --- | --- |
| # | REQUERIMIENTO | PRIORIDAD\* | EXIGIBLE/  DESEABLE |
| 1 | Se extraerá de forma adecuada la información de las fuentes de datos. | 1 | E |
| 2 | Se creará un almacén de datos. | 1 | E |
| 3 | Se cargará la información en el staging área | 1 | E |
| 4 | Se realizarán las transformaciones de los datos para crear el Data Mart | 2 | E |
| 5 | Se creará un modelo OLAP para consultas automáticas de los usuarios. | 2 | E |
| 6 | Se crearán los informes estáticos solicitados. | 2 | E |
| 7 | Cada usuario tendrá acceso a un conjunto de información relevante determinada | 2 | D |
| 8 | El sistema a través de los datos podrá predecir posibles resultados futuros o ciertos comportamientos | 2 | D |
| 9 | Análisis de viabilidad y análisis de riesgos | 2 | D |
| 10 | Se podrá realizar consultas a través de un dispositivo móvil. | 3 | D |
| 11 | Se redactará un manual de carga de datos  inicial e incremental. | 3 | D |
| 12 | Creación de procesos de cargas incrementales | 3 | D |

\* Se asigna una prioridad del 1 al 3, siendo 1 completamente prioritario para la actividad y 3 no prioritario.

En términos de la arquitectura funcional existen los siguientes elementos:

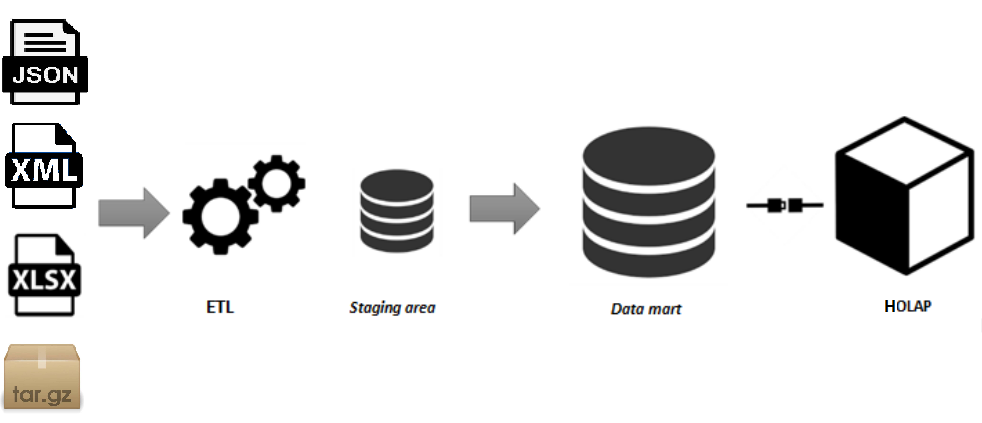
* Las fuentes de datos de las que se dispone son las siguientes:
  + 3 ficheros Excel (un fichero del tipo xls y dos ficheros xlsx).
  + 1 fichero en formato XML.
  + 1 fichero en formato json.
  + 1 fichero en formato tsv.gz.
* La arquitectura de la factoría de información estará formada por varios elementos alojados en la misma máquina:
  + **Staging Area**: para consolidar la información en una estructura de carga intermedia.
  + **Data mart** con el análisis del impacto ambiental y del consumo energético.
  + **HOLAP:** a partir de la información de la data mart se creará un MOLAP y un ROLAP.

Se ha considerado que la arquitectura HOLAP es la más adecuada por los siguientes motivos:

* Vamos a tener datos bastante estáticos que se puede tratar mediante la una arquitectura MOLAP, como por ejemplo los Objetivos de Desarrollo Sostenible y las áreas de conservación de la biodiversidad.
* Por otro lado, hay otros datos que anualmente se van a ir añadiendo y pueden modificarse su estructura, por eso un sistema ROLAP es lo mas adecuado, ya que su sensibilidad a los cambios es mucho más baja.

Una vez definido los elementos de la arquitectura se concluye que tenemos una arquitectura funcional que usa un área intermedia (staging area) y se crearía dentro de la misma base de datos, cuyos objetos se identificarán con un prefijo en los nombres.

El siguiente gráfico resume los elementos de la arquitectura necesarios para esta actividad:



# Diseño del modelo conceptual, lógico y físico del almacén de datos

## Diseño conceptual

Del análisis de las fuentes de datos y de los requerimientos iniciales, se puede determinar que uno de los hechos que se deben considerar es:

* **Mediciones ambientales.** Hace referencia a ciertas mediciones ambientales relevantes para un desarrollo sostenible.

El análisis de las mediciones ambientales determina el diseño de la primera tabla de hechos, como se puede observar a continuación:

|  |  |
| --- | --- |
| TABLA DE HECHOS | DESCRIPCIÓN |
| FACT\_EnvironmentalMeasurements | Mediciones ambientales para un desarrollo sostenible |

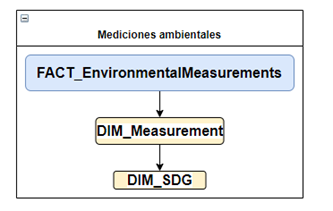
Las métricas de la tabla de hechos «FACT\_EnvironmentalMeasurements» son las siguientes:

|  |  |
| --- | --- |
| MÉTRICA | DESCRIPCIÓN |
| Value | Valor de la medición |

Las métricas de esta tabla de hechos podrán ser analizadas desde las diferentes perspectivas, a partir de las dimensiones:

|  |  |
| --- | --- |
| DIMENSIONES | DESCRIPCIÓN |
| SDG | Objetivos de desarrollo sostenible con los que se relaciona la medición |
| Measurement | Medición a tratar |

A nivel de diseño quedaría de la siguiente manera:



Otro de los hechos que se deben considerar es:

* **Inversiones en protección ambiental.** Hace referencia a ciertas inversiones realizadas por los países relevantes para la protección del medio ambiente.

El análisis de las inversiones en protección ambiental determina el diseño de la segunda tabla de hechos, como se puede observar a continuación:

|  |  |
| --- | --- |
| TABLA DE HECHOS | DESCRIPCIÓN |
| FACT\_ ProtectionEnvironmentalInvestments | Inversiones en la protección del medio ambiente |

Las métricas de la tabla de hechos «FACT\_ProtectionEnvironmentalInvestments» son las siguientes:

|  |  |
| --- | --- |
| MÉTRICA | DESCRIPCIÓN |
| Value | Valor de la inversión en un periodo de tiempo |

Las métricas de esta tabla de hechos podrán ser analizadas desde las diferentes perspectivas, a partir de las dimensiones:

|  |  |
| --- | --- |
| DIMENSIONES | DESCRIPCIÓN |
| Country | País que realiza la inversión |
| SDG | Objetivos de desarrollo sostenible con los que se relaciona la medición |
| Tiempo | Periodo de tiempo en que se ha realizado la inversión |
| Type of investment | Tipo de inversión a tratar |

A nivel de diseño quedaría de la siguiente manera:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Otro de los hechos que se deben considerar es:

* **Balance en producción de energía verde.** Hace referencia a la distribución que existe en las diversas formas de generar energía.

El análisis del balance determina el diseño de la tercera tabla de hechos, como se puede observar a continuación:

|  |  |
| --- | --- |
| TABLA DE HECHOS | DESCRIPCIÓN |
| FACT\_ GreenEnergyProductionBalance | Distribución entre las diferentes formas de generar energía verde |

Las métricas de la tabla de hechos «FACT\_ EnergyProductionBalance» son las siguientes:

|  |  |
| --- | --- |
| MÉTRICA | DESCRIPCIÓN |
| Percentage | Porcentaje de producción de energía verde |

Las métricas de esta tabla de hechos podrán ser analizadas desde las diferentes perspectivas, a partir de las dimensiones:

|  |  |
| --- | --- |
| DIMENSIONES | DESCRIPCIÓN |
| Country | País en que se realiza el balance |
| Tiempo | Periodo de tiempo en que se ha realizado el balance |
| Types | Tipos de energía generada |
| Reach | Alcance que tiene la producción (local, otros países…) |

A nivel de diseño quedaría de la siguiente manera:

Diagrama

Descripción generada automáticamente

Unido todos los diseños, quedaría de la siguiente manera:

Diagrama

Descripción generada automáticamente

## Diseño lógico

Una vez obtenido el modelo conceptual del almacén de datos para el análisis del impacto ambiental, pasamos a realizar el diseño lógico del mismo.

A continuación, se muestra la tabla con las métricas identificadas en el diseño conceptual de la tabla de hechos «FACT\_EnvironmentalMeasurements»:

|  |  |
| --- | --- |
| TABLA DE HECHOS | MÉTRICAS |
| FACT\_EnvironmentalMeasurements | Value |

Detalle de los atributos descriptores de las dimensiones de cada hecho:

|  |  |
| --- | --- |
| DIMENSIONES | ATRIBUTOS |
| SDG | Código, nombre, descripción |
| Measurement | Código, nombre de la medición y unidad de medida |

A continuación, se muestra la tabla con las métricas identificadas en el diseño conceptual de la tabla de hechos «FACT\_ProtectionEnvironmentalInvestments»:

|  |  |
| --- | --- |
| TABLA DE HECHOS | MÉTRICAS |
| FACT\_ ProtectionEnvironmentalInvestments | Value |

Detalle de los atributos descriptores de las dimensiones de cada hecho:

|  |  |
| --- | --- |
| DIMENSIONES | ATRIBUTOS |
| Country | Nombre, iso2, iso3 |
| SDG | Código, nombre, descripción |
| Tiempo | Año, mes y día |
| Type of investment | Código, nombre de la inversión, cantidad, plazo, sector |

A continuación, se muestra la tabla con las métricas identificadas en el diseño conceptual de la tabla de hechos «FACT\_EnergyProductionBalance»:

|  |  |
| --- | --- |
| TABLA DE HECHOS | MÉTRICAS |
| FACT\_EnergyProductionBalance | Percentage |

Detalle de los atributos descriptores de las dimensiones de cada hecho:

|  |  |
| --- | --- |
| DIMENSIONES | ATRIBUTOS |
| Country | Nombre, iso2, iso3 |
| Tiempo | Año, mes y día |
| Types | Código, tipo de energía generada, porcentaje de producción, cantidad |
| Reach | Consumo Energía Verde, Producción Energía, Periodo |

## Diseño físico

Una vez que se han determinado las tablas de hechos, las dimensiones, las métricas y los atributos que existen en el modelo lógico, podemos pasar a realizar el diseño físico, lo cual significa obtener una implementación del modelo lógico en términos del sistema gestor de bases de datos escogido.

### Dimensiones

El modelo físico de las dimensiones identificadas es el siguiente:

● **DIM\_SDG**: contiene los datos de los objetivos de desarrollo sostenible.

|  |  |  |  |
| --- | --- | --- | --- |
| NOMBRE CAMPO | TIPO | TAMAÑO | EJEMPLO |
| Pk\_sdg (PK) | Numérico | 4 | 13 |
| Sdg\_name | Texto | 50 | «Acción por el clima» |
| Sdg\_description | Texto | 500 | «El cambio climático es un reto global que no respeta las fronteras nacionales» |

● **DIM\_Measurement**: contiene los datos de la medición.

|  |  |  |  |
| --- | --- | --- | --- |
| NOMBRE CAMPO | TIPO | TAMAÑO | EJEMPLO |
| pk\_measurement (PK) | Numérico | 4 | 1 |
| measurement\_code | Texto | 100 | «Protección del aire y el clima» |
| measurement\_name | Texto | 200 | «Inversión en protección ambiental» |
| Unit | Texto | 25 | «euros» |
| fk\_sdg | Numérico | 4 | 13 |

* **DIM\_Country**: contiene los datos del país donde se hacen las mediciones o los cálculos.

|  |  |  |  |
| --- | --- | --- | --- |
| NOMBRE CAMPO | TIPO | TAMAÑO | EJEMPLO |
| pk\_country (PK) | Numérico | 4 | 15 |
| Country\_name | Texto | 100 | «CUBA» |
| Country\_ISO2 | Texto | 200 | «CU» |
| Country\_ISO3 | Texto | 25 | «CUB» |

* **DIM\_Tiempo**: corresponde a la dimensión temporal de las mediciones o inversiones realizadas.

|  |  |  |  |
| --- | --- | --- | --- |
| NOMBRE CAMPO | TIPO | TAMAÑO | EJEMPLO |
| pk\_Tiempo (PK) | Numérico | 4 | 25 |
| Year | Numérico | 4 | 2020 |
| Month | Numérico | 2 | 6 |
| Day | Numérico | 2 | 20 |
| Date | Fecha | 10 | 20/06/2020 |

* **DIM\_Type\_of\_investment**: contiene los datos de la inversión

|  |  |  |  |
| --- | --- | --- | --- |
| NOMBRE CAMPO | TIPO | TAMAÑO | EJEMPLO |
| pk\_TypeOfInvestement (PK) | Numérico | 4 | 5 |
| Investment\_Name | Texto | 25 | «Cambio Industria del Carbón» |
| amount | Numérico | 10 | 150000€ |
| Time\_Frame | Fecha | 10 | 10/12/2021 |
| sector | Texto | 25 | «Producción Energía» |

* **DIM\_Type**: contiene los datos de la energía generada.

|  |  |  |  |
| --- | --- | --- | --- |
| NOMBRE CAMPO | TIPO | TAMAÑO | EJEMPLO |
| pk\_Type (PK) | Numérico | 4 | 2 |
| Type\_Name | Texto | 25 | «Energía Solar» |
| Percentage | Numérico | 4 | 60% |
| amount | Numérico | 10 | 52000 |

* **DIM\_Reach**: contiene los datos del consumo y alcance de la energía.

|  |  |  |  |
| --- | --- | --- | --- |
| NOMBRE CAMPO | TIPO | TAMAÑO | EJEMPLO |
| pk\_Reach (PK) | Numérico | 4 | 3 |
| Name | Texto | 25 | «Energía Solar» |
| GreenEnergyConsumption | Numérico | 4 | 12% |
| EnergyProduction | Numérico | 4 | 50% |
| Period | Numérico | Fecha | 10/12/2021 |

### Tablas de hechos

El modelo físico de las tablas de hechos del almacén de datos para el impacto ambiental y el consumo energético está compuesto, entre otras, de las siguientes tablas:

* **FACT\_EnvironmentalMeasurements**: es la tabla física que contendrá la información que permitirá realizar el análisis de los datos de las mediciones ambientales para un desarrollo sostenible. Tendrá los siguientes campos:

|  |  |  |  |
| --- | --- | --- | --- |
| NOMBRE CAMPO | TIPO | TAMAÑO | EJEMPLO |
| pk\_id (PK) | Numérico | 4 | 17 |
| fk\_date (FK) | Numérico | 4 | 25 |
| fk\_region (FK) | Numérico | 4 | 10 |
| fk\_activitysector (FK) | Numérico | 4 | 1 |
| fk\_typeequipinstall (FK) | Numérico | 4 | 1 |
| fk\_measurement (FK) | Numérico | 4 | 1 |
| value | Numérico | 8 | 6.390.920 |

* **FACT\_ProtectionEnvironmentalInvestments**: es la tabla física que contendrá la información que permitirá realizar el análisis de los datos de las inversiones para la protección del medio ambiente. Tendrá los siguientes campos:

|  |  |  |  |
| --- | --- | --- | --- |
| NOMBRE CAMPO | TIPO | TAMAÑO | EJEMPLO |
| pk\_id (PK) | Numérico | 4 | 50 |
| fk\_date (FK) | Numérico | 4 | 25 |
| fk\_region (FK) | Numérico | 4 | 10 |
| fk\_SDG (FK) | Numérico | 4 | 1 |
| fk\_typeOfInvestement (FK) | Numérico | 4 | 1 |
| value | Numérico | 8 | 6.390.920 |

* **FACT\_EnergyProductionBalance**: es la tabla física que contendrá la información que permitirá realizar el análisis de los datos de la producción de energía. Tendrá los siguientes campos:

|  |  |  |  |
| --- | --- | --- | --- |
| NOMBRE CAMPO | TIPO | TAMAÑO | EJEMPLO |
| pk\_id (PK) | Numérico | 4 | 23 |
| fk\_date (FK) | Numérico | 4 | 25 |
| fk\_region (FK) | Numérico | 4 | 10 |
| fk\_activitysector (FK) | Numérico | 4 | 1 |
| fk\_type (FK) | Numérico | 4 | 1 |
| Percentage | Numérico | 8 | 0.85 |