Documentación Arquitectura de datos

Big Data II TerraWatt

Alma Gutierrez
Rafael Borge
Deniz Alcobendas
Íñigo Pérez

Índice

Introducción

La fundación de cualquier sistema de análisis predictivo sólido se basa en un diseño de base de datos sólido, que esté en capacidad de manejar de manera efectiva calidad, relevancia y accesibilidad de la información. Para este proyecto se concebía una base de datos enfocada en la unificación de información heterogénea obtenida de varias fuentes relevantes para las actividades energetizantes, con la misión de predecir el precio promedio por día de electricidad en España.

La estrategia implementada prioriza el ordenamiento del flujo de datos desde su origen hasta su aprovechamiento. A ese efecto, se delimitó un sistema apto para recopilar, transformar, almacenar y brindar datos procedentes de diversas entidades que afectan el comportamiento del mercado eléctrico. Entre las entidades analizadas se encuentran registros meteorológicos precisos de la Agencia Estatal de Meteorología (AEMET), precios pasados obtenidos mediante la API de Red Eléctrica Española (REE), y un calendario de días fiestas cubiertos tanto en ámbito nacional como autonómico, e se sumaron datos relacionados con infraestructura y demografía residencial.

Cada uno de los conjuntos de información se hizo pasar por procesos de normalización, limpieza y estructura, con el propósito de integrarlos en base de datos relacional centralizada. Esta plataforma tiene no sólo trazabilidad de información y adecuada alineación geográfica y temporal, sino que también ofrece posibilidades de escalabilidad del sistema ante fuentes nuevas o reformulación de las entradas en formatos.

Fuentes de datos

La arquitectura de datos desarrollada se apoya en una integración estructurada de múltiples fuentes de información que, combinadas, permiten reflejar la complejidad del mercado eléctrico español. Estas fuentes fueron cuidadosamente seleccionadas por su relevancia y complementariedad, e incorporadas al sistema mediante procesos ETL automatizados, garantizando la coherencia, trazabilidad y disponibilidad continua de los datos.

Una de las principales fuentes es la Agencia Estatal de Meteorología (AEMET), que proporciona registros diarios desde 1920 hasta 2024 a través de su API y archivos CSV. La información proviene de 947 estaciones distribuidas por todo el país e incluye variables como temperatura, precipitaciones, viento, presión atmosférica y horas de sol. Tras aplicar procesos de limpieza y normalización —como corregir errores y unificar nombres de

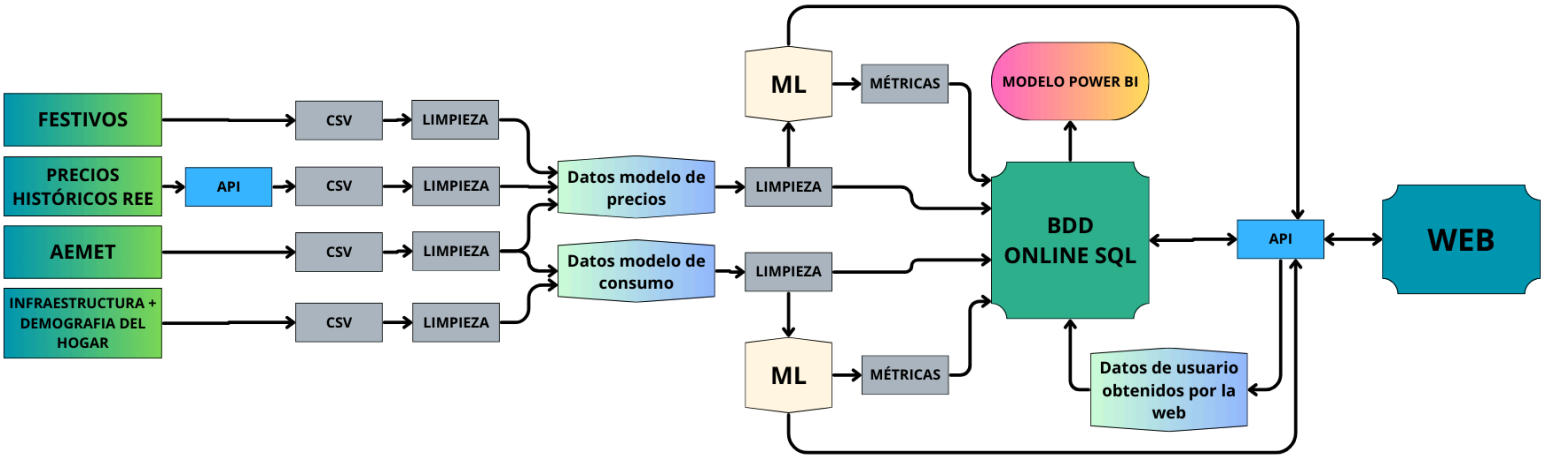
provincias— los datos se agrupan por región. Esta fuente es clave para entender cómo el clima influye en la demanda eléctrica, ya que permite anticipar aumentos en el consumo durante episodios de frío o calor extremos, y ajustar los modelos a nivel provincial.

Por otro lado, los precios históricos de la electricidad ofrecidos por la Red Eléctrica Española (REE) también fueron incorporados mediante su API, consolidando la información en el archivo precios_energia.csv. Esta base de datos incluye precios horarios y diarios de los últimos diez años, desglosados en precio medio sin impuestos, impuestos aplicados e IVA, reflejando el coste total por megavatio hora. Luego de limpiar datos incompletos, rellenar valores faltantes y homogeneizar los formatos, esta fuente se convierte en una herramienta esencial para anticipar fluctuaciones de precios. Su uso permite prever los costes energéticos, identificar horarios punta y valle, y establecer vínculos entre precios, condiciones climáticas y días festivos, lo cual resulta útil para tomar decisiones financieras más informadas.

El calendario de días festivos, contenido en el archivo Festivos.csv y generado mediante ChatGPT, incluye una recopilación detallada de festivos nacionales, autonómicos y locales, especificando la fecha, la provincia y el nombre de cada celebración. Este conjunto de datos fue enriquecido con columnas adicionales como “Festivo (Sí/No)” y “Entre Semana (Sí/No)”, lo que permitió integrarlo fácilmente al conjunto final. Su valor principal radica en reflejar cómo varía el consumo eléctrico durante los días no laborables: aumenta en los hogares y disminuye en sectores industriales. Gracias a esta información, los modelos predictivos pueden ajustarse con mayor precisión en fechas especiales como la Navidad, y permiten aplicar segmentaciones regionales más exactas.

Por otro lado, los datos sobre infraestructura y demografía del hogar, que formarán parte del modelo de predicción de consumo, incluyen detalles como el tipo de vivienda (piso, adosado, dúplex o casa unifamiliar), la potencia contratada en kilovatios (kW), el número promedio de habitantes por hogar y el consumo energético por metro cuadrado (kWh/m²). Estos datos, al integrarse con la información meteorológica y ser normalizados para mantener la coherencia, permiten analizar cómo varía el consumo según el tipo de vivienda. Su incorporación en modelos como Random Forest o XGBoost hace posible personalizar las predicciones, recomendar cambios en la potencia contratada para reducir costes y promover un consumo más eficiente, adaptado a las características específicas de cada hogar.

Arquitectura de datos



Almacenamiento de Datos

En el proyecto TerraWatt, el almacenamiento de datos se realiza en una base de datos relacional SQL (como PostgreSQL o MySQL) tras un proceso ETL que transforma y combina múltiples fuentes de datos. A continuación, se describe cómo se organizan los repositorios de datos, el esquema de la base de datos y las consideraciones específicas para garantizar un almacenamiento y consulta eficiente.

Los datos procesados a través del flujo ETL se almacenan en tablas específicas dentro de una base de datos estructurada, diseñada para reflejar los conjuntos de datos clave del proyecto. Cada tabla se ha creado para contener un tipo específico de información transformada, proveniente de las distintas fuentes originales:

- **Tabla Meteorología**: Recoge los datos meteorológicos diarios, agregados por provincia, a partir de los registros proporcionados por AEMET.

- **Tabla Precios**: Almacena los precios diarios de la electricidad, calculados con base en los datos históricos obtenidos de la Red Eléctrica Española (REE).

- **Tabla Festivos**: Registra los días festivos por provincia, a partir del archivo Festivos.csv, incluyendo tanto festivos nacionales como autonómicos y locales.

- **Tabla Consumo Hogares**: Contiene información mensual sobre el consumo energético segmentado por tipo de vivienda, combinada con variables meteorológicas reales para mejorar la precisión del modelo.

Esquema de la Base de Datos

El diseño del esquema SQL está orientado a representar fielmente la estructura de los datos procesados. Cada tabla incluye columnas específicas que recogen las variables más relevantes para el análisis, asegurando una organización clara, eficiente y escalable.

Dimensiones

Estas tablas contienen atributos descriptivos que contextualizan los datos numéricos registrados en las tablas de hechos.

dim_fecha_dia

- **ID_Tiempo_Dia** *(PK)*
- Fecha completa: Fecha
- Componentes temporales: Año, Mes, Día, Nombre_Mes, Trimestre

Permite análisis diarios detallados por múltiples perspectivas temporales.

dim_fecha_mes_anio

- **ID_Tiempo_Mes_Anio** *(PK)*
- Año, Mes, Nombre_Mes, Trimestre, Año_Mes

Soporta agregaciones mensuales y trimestrales para tendencias de mediano plazo.

dim_provincia

- **ID_provincia** *(PK)*
- Nombre_provincia, Nombre_completo

Proporciona la dimensión geográfica del análisis.

dim_residentes

- **ID_residentes** *(PK)*
- Rango
- Residentes_min, Residentes_max

Clasifica los hogares según número de residentes, relevante para estimar consumo.

dim_potencia

- **ID_potencia** *(PK)*
- Rango
- Potencia_min, Potencia_m

Describe la capacidad contratada de suministro eléctrico.

dim_vivienda

- **ID_vivienda** *(PK)*
- tipo_de_vivienda (piso, casa, dúplex, etc.)

Permite segmentar el consumo según tipo de vivienda.

dim_festivos

- **ID_festivos***(PK)*
- ID_fecha(en formato DDMMYYYY, que se conecta con la dimension **dim_fecha_dia**)
- Provincia
- Descripcion festividad

Permite segmentar el consumo según tipo de vivienda.

Tablas de Hechos

Almacenan los datos numéricos de consumo, precios y predicciones, junto con claves foráneas hacia las dimensiones.

Datos_precios_SQL_ID

- **Clave Primaria:** ID_Precios
- **Claves foráneas**:, ID_Tiempo_Dia, ID_provincia

- **Variables meteorológicas**: TMEDIA, TMIN, TMAX, VELMEDIA, PRESMAX, PRESMIN, SOL, RACHA, ALTITUD, DIR
- **Variables de precio**: Precio_KWH, Precio_total_con_impuestos
- **Atributos adicionales**: Festivo, Entre_semana

Analiza la variación diaria de precios eléctricos en relación con el clima y la geografía.

Datos_consumo_SQL_ID

- **Clave Primaria:** ID_Consumo
- **Claves foráneas**: ID_Tiempo_Mes_Anio, ID_provincia, ID_potencia, ID_residentes, ID_vivienda
- **Variable objetivo**: Consumo_energetico_kWh_m2
- **Variables meteorológicas**: TMEDIA, TMIN, TMAX, VELMEDIA, SOL, PRESMAX, PRESMIN

Evalúa el consumo energético mensual en función del hogar y el entorno climático.

Datos_predicciones_SQL_ID

- **Clave Primaria:** ID_Predicciones
- **Claves foráneas**: ID_Tiempo_Dia, ID_provincia, ID_potencia, ID_residentes, ID_vivienda
- **Predicciones del modelo**: PREDICCION_CONSUMO, PREDICCION_PRECIO
- **Estimaciones económicas**: COSTE_ESTIMADO, COSTE_POTENCIA

Contiene predicciones diarias personalizadas de consumo y precio, con impacto económico estimado.

Error_modelo_consumo

- **Clave primaria**: ID_modelo
- Fecha_generacion, Comunidad
- Métricas de evaluación: MAE, RMSE, R2

Evalúa el rendimiento del modelo de consumo por región y momento de generación.

Error_modelo_precios

- **Clave primaria**: ID_modelo
- Fecha_generacion
- Métricas: Accuracy, MAE, RMSE, R2, MAE_kWh

Mide la eficacia de los modelos de predicción de precios eléctricos

Relaciones

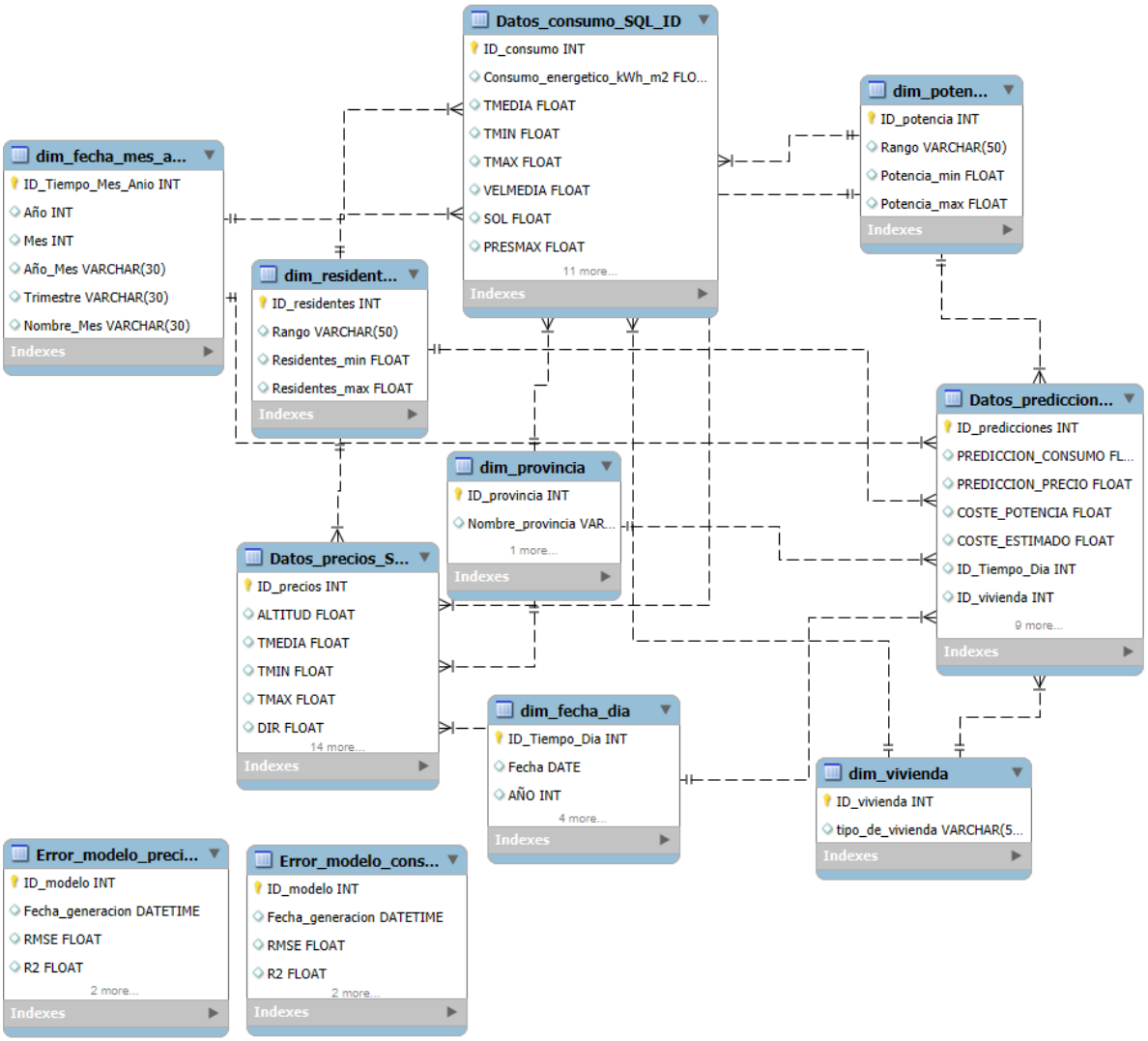
El modelo sigue una topología de estrella, donde **cada dimensión se relaciona con múltiples registros en las tablas de hechos**. Las claves foráneas permiten vincular los hechos con sus respectivos contextos:

- Datos_consumo_SQL_ID → se conecta con: dim_fecha_mes_anio, dim_provincia, dim_residentes, dim_potencia, dim_vivienda
- Datos_predicciones_SQL_ID → usa las mismas dimensiones, pero con dim_fecha_dia para granularidad diaria
- Datos_precios_SQL_ID → vinculado a dim_fecha_dia y dim_provincia

Diagrama en estrella de la base de datos

El diseño de modelos de datos en el proyecto TerraWatt sigue un enfoque relacional en estrella, estructurado para facilitar el análisis eficiente y detallado de datos energéticos. En este esquema, se integran tablas de hechos, que contienen métricas cuantitativas como consumo, precios y predicciones, con tablas de dimensiones que aportan contexto descriptivo, como fecha, ubicación geográfica, características de la vivienda, número de residentes o potencia contratada.

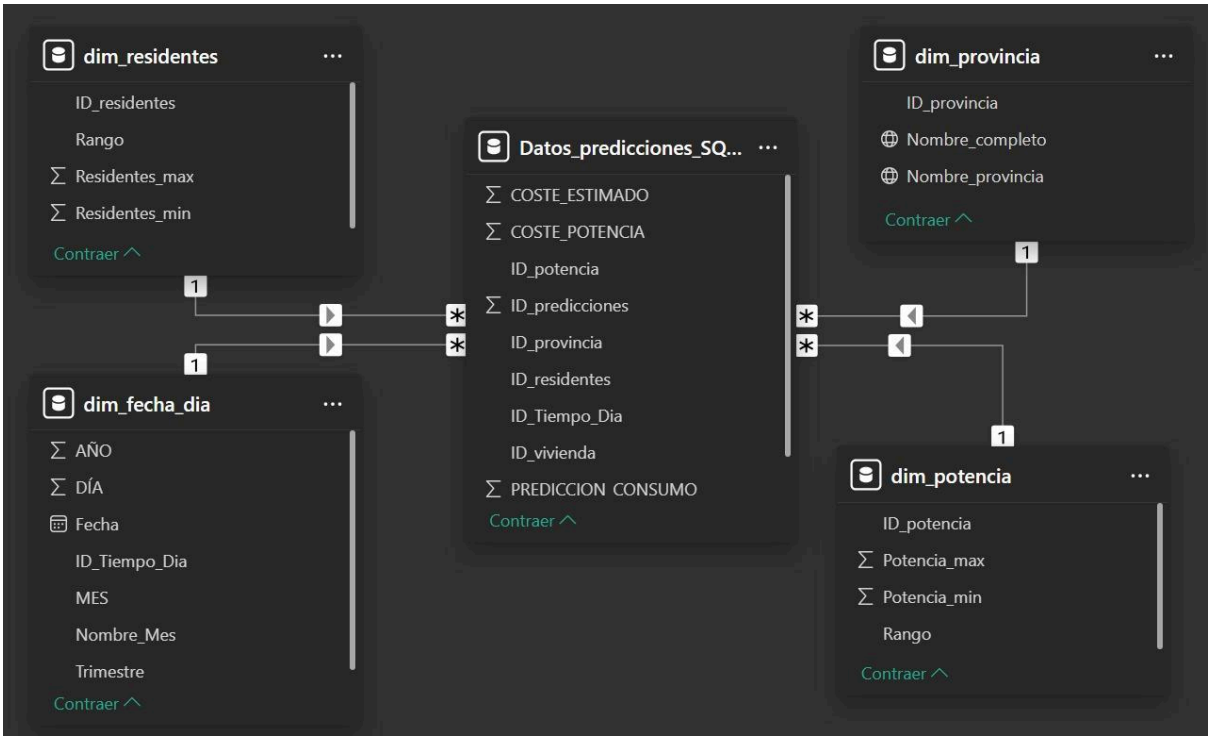
Este enfoque permite realizar consultas complejas, segmentaciones precisas y análisis multidimensionales que son fundamentales para la toma de decisiones en el sector energético.



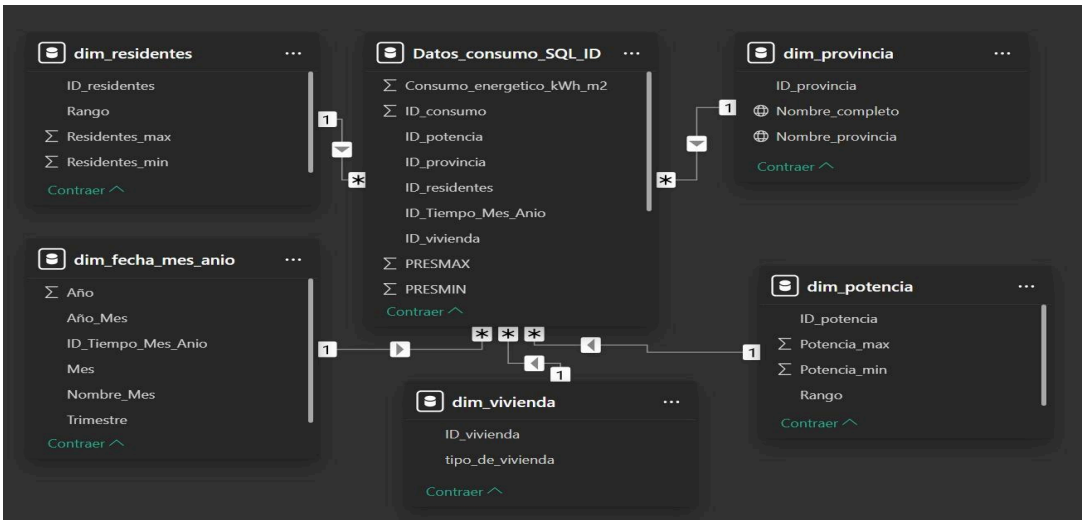
Características del Modelo en Estrella

- **Tablas de hechos múltiples**: Se incluyen Datos_consumo, Datos_predicciones, y Datos_precios_SQL, todas ellas conectadas a dimensiones comunes que permiten cruzar información entre consumo, predicciones y precios.
- **Dimensiones compartidas**: Tablas como dim_fecha_mes_anio, dim_fecha_dia, dim_provincia, dim_residente, dim_potencia y dim_vivienda se reutilizan en los distintos hechos, reduciendo la redundancia y promoviendo consistencia.
- **Escalabilidad estructural**: El modelo permite añadir nuevas tablas de hechos sin alterar las dimensiones existentes, lo cual es esencial para incorporar futuros conjuntos de datos dentro del marco del proyecto.

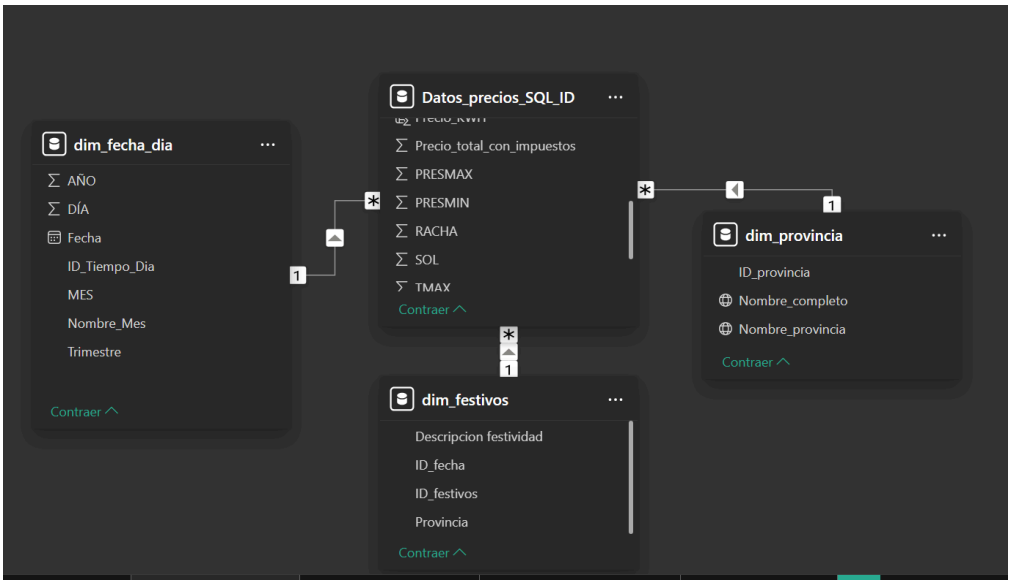
- **Optimización para análisis multidimensional**: La arquitectura admite segmentaciones por variables clave como tiempo, ubicación, tipo de vivienda o características demográficas, lo que mejora el poder predictivo de los modelos.



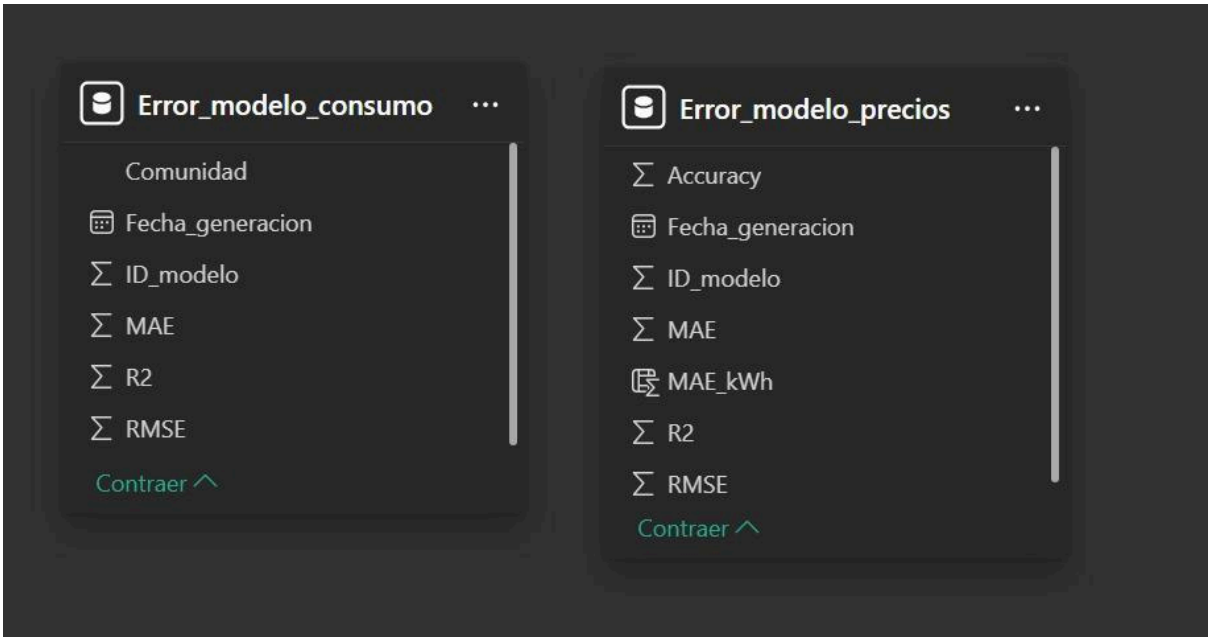
Datos_predicciones_SQL: Esta tabla integra predicciones de consumo y coste, y se vincula con dimensiones como dim_residente, dim_fecha_dia, dim_provincia y dim_potencia. Esta configuración permite generar estimaciones precisas por día, provincia y potencia contratada, fundamentales para anticipar demandas energéticas y planificar la infraestructura.



Datos_consumo_SQL_ID: Diseñada para capturar el consumo energético por metro cuadrado, se relaciona con dimensiones como dim_residente, dim_provincia, dim_fecha_mes_anio, dim_vivienda y dim_potencia. Esto permite segmentar el consumo mensual según características sociodemográficas y técnicas, clave para análisis de eficiencia energética y diseño de políticas sostenibles.



Datos_precios_SQL_ID: Contiene precios energéticos con granularidad diaria y está conectada a las dimensiones dim_fecha_dia y dim_provincia. El objetivo es facilitar el análisis de tendencias temporales y regionales en los precios de la electricidad, lo cual resulta esencial para ajustar tarifas y prever costes.



El modelo también incluye tablas de hechos centradas en **métricas de error** como MAE, RMSE y R^2, que permiten evaluar el rendimiento de los modelos de predicción tanto para consumo como para precios. Estas tablas utilizan dimensiones como:

- **Fecha_generación:** para hacer seguimiento temporal del rendimiento,
- **ID_modelo:** para identificar el algoritmo evaluado,
- **Comunidad:** usada en el caso del consumo, para incluir una dimensión regional adicional.

Este diseño proporciona una base sólida para validar la calidad de las predicciones y ajustar los algoritmos de manera continua, asegurando que las decisiones basadas en datos sean precisas y fiables en el contexto energético del proyecto TerraWatt.

Procesamiento de datos

A continuación, se detalla cada etapa del flujo de trabajo en un párrafo por documento, explicando las acciones realizadas, su propósito y cómo contribuyen al objetivo final de generar un dataset unificado y limpio para análisis o modelado.

Extraccion_precios_energia.py

Este script automatiza la extracción de datos históricos de precios energéticos desde la API de ESIOS (Red Eléctrica de España) para el periodo comprendido entre el 1 de abril de 2014 y el 1 de abril de 2024. Utiliza autenticación mediante una clave API y encabezados

personalizados. El script itera diariamente dentro del rango de fechas, solicitando datos en formato JSON que contienen precios horarios bajo claves como "PVPC" o "PCB", según la fecha (diferenciadas por cambios regulatorios ocurridos en junio de 2021).

Para cada día, calcula el precio medio diario sin impuestos, aplica el impuesto eléctrico (5,11%) y el IVA (21%), y genera el precio total con impuestos. Estos resultados se almacenan en un DataFrame y se exportan como CSV (precios_energia.csv) en la carpeta Datos_brutos. Además, incorpora manejo de errores para prevenir fallos ante JSON malformados o problemas de conexión, garantizando un proceso robusto y fiable. Esta etapa es crítica para construir una base sólida de precios energéticos que luego será integrada con otras fuentes para análisis predictivos.

Datos_limpieza_meteorologicos.py

Este script se encarga de limpiar y estandarizar los datos meteorológicos almacenados inicialmente en la carpeta Datos_brutos_meteorologicos, generando archivos depurados en Datos_limpios_meteorologicos. El proceso elimina columnas no necesarias (como INDICATIVO y NOMBRE), normaliza los nombres de provincias (por ejemplo, convierte “ARABA/ALAVA” en “ALAVA”), y descarta registros anteriores a abril de 2014 para alinearlos temporalmente con los datos de precios.

Luego, agrupa los datos por provincia y fecha, calculando medias diarias de variables meteorológicas (temperatura, viento, etc.) para consolidar múltiples estaciones. También se excluyen los datos de Ceuta y Melilla debido a la ausencia de registros de consumo energético. Como resultado, se obtiene un conjunto uniforme y coherente de datos listos para su integración con otras fuentes.

Generacion_csv_modelos.py

Este script unifica los datos meteorológicos limpios, los precios energéticos y el calendario de festivos en un solo archivo (Modelo_Precios_Met_Fest.csv), que será la base para el modelado predictivo. Para asegurar compatibilidad, los archivos meteorológicos se convierten a UTF-8. Luego, se cargan los precios (precios_energia.csv) y los festivos (Festivos.csv) y se estandarizan tanto las fechas como los nombres de provincias.

Cada archivo meteorológico se combina con los precios mediante un inner join por fecha, y se enriquecen los registros con columnas como Festivo (Sí/No) y Entre semana (Sí/No). El script concatena todos los resultados, selecciona columnas clave (como FECHA, Provincia, TMEDIA, Precio total con impuestos), y gestiona valores nulos rellenándolos con la media provincial cuando el porcentaje de datos faltantes es bajo (por ejemplo: SOL 0.21%, PRESMAX 0.04%). Finalmente, calcula estadísticas preliminares y guarda el dataset consolidado.

Ingesta de datos

Flujo de Ingesta de Datos a la Base de Datos SQL

El proceso de carga de datos a la base de datos SQL comienza con la recopilación y transformación de conjuntos históricos relacionados con precios energéticos, consumo eléctrico y variables meteorológicas. Una vez preparados, estos datos se insertan en una base de datos MySQL remota mediante una conexión gestionada con SQLAlchemy.

Los datos se distribuyen en tablas específicas como Datos_precios y Datos_consumo, con una configuración que permite agregar nueva información sin sobrescribir los registros existentes. Además, se establece una tabla de dimensión para provincias, asignando identificadores únicos a cada una, lo que facilita las relaciones entre tablas.

Flujo de Ingesta de Datos del Usuario

La ingesta de datos desde el usuario se realiza a través de una **interfaz web**, donde un formulario permite introducir manualmente información relevante como:

- Potencia contratada
- Número de residentes
- Tipo de vivienda
- Provincia
- Mes de consumo

Al enviar el formulario, los datos son recogidos mediante JavaScript utilizando la función **enviarDatos()**, que convierte los campos al formato adecuado (por ejemplo, potencia y residentes como valores numéricos, y el resto como cadenas de texto). Esta información se transmite al backend a través de una **petición HTTP POST** al endpoint de una API construida con **FastAPI**, en formato **JSON**.

En el backend, estos datos son procesados y, si es necesario, combinados con información histórica y sintética previamente almacenada en la base de datos SQL. Esto permite generar predicciones personalizadas en tiempo real. El sistema así diseñado permite una integración ágil entre la entrada del usuario y el motor analítico, haciendo posible análisis dinámicos y altamente personalizados.

Conclusiones

TerraWatt se consolida como una solución integral para la predicción del precio medio diario de la electricidad en España. Su fortaleza radica en una arquitectura de datos robusta, escalable y relacional, estructurada bajo un esquema en estrella. La integración de fuentes heterogéneas —como datos meteorológicos, históricos de precios, festivos y características del hogar— ha dado lugar a un sistema analítico completo que mejora significativamente la precisión de las predicciones y respalda decisiones tanto técnicas como estratégicas.

Gracias a un riguroso tratamiento de datos mediante técnicas de limpieza, normalización y unión, se garantiza la consistencia y fiabilidad del modelo. Además, el diseño modular del sistema facilita la incorporación futura de nuevas fuentes o componentes analíticos. Una de las funcionalidades más destacadas es la capacidad de ofrecer predicciones personalizadas a usuarios individuales, lo que incrementa el valor práctico y la adaptabilidad del sistema.

Retos y Limitaciones

1. **Heterogeneidad de fuentes de datos**
   La integración de fuentes con distintas estructuras, formatos y niveles de granularidad (como AEMET vs. REE) supuso el desarrollo de procesos ETL complejos y altamente personalizados.

2. **Calidad de los datos históricos**
   Aunque en general robustos, algunos conjuntos meteorológicos incluían valores nulos o inconsistentes. Se aplicaron técnicas de imputación que, si bien controladas, podrían introducir cierto sesgo.

3. **Cobertura geográfica incompleta**
   Ceuta y Melilla fueron excluidas por falta de datos de consumo. Además de las Islas canarias, que fueron excluidas debido a que no obtienen electricidad de la red eléctrica española

4. **Limitaciones temporales**
   El modelo trabaja con datos históricos hasta abril de 2024. Sin una actualización periódica, su capacidad predictiva puede disminuir con el tiempo.

5. **Carga computacional**
   El procesamiento y modelado diario de grandes volúmenes de datos representa un desafío técnico, que puede intensificarse al ampliar la cobertura geográfica o incluir nuevas variables.

Mejoras Propuestas y Roadmap Futuro

Una de las principales mejoras planteadas es la automatización del pipeline de datos. Para ello, se propone establecer tareas programadas (como cron jobs) que permitan la ingesta diaria automática de información proveniente de las APIs de AEMET y Red Eléctrica Española (REE). Además, se plantea automatizar los procesos de limpieza y validación de los datos en tiempo real, garantizando así una calidad continua en todas las etapas del flujo de información.

Otra línea de avance consiste en la incorporación de nuevas variables relevantes para enriquecer el análisis. En este sentido, se busca integrar datos socioeconómicos, como niveles de ingresos o patrones de consumo energético por hogar, lo cual permitiría construir modelos más precisos y personalizados. Asimismo, se propone añadir información relativa a tarifas dinámicas y ofertas regionales disponibles en distintas comercializadoras eléctricas, ampliando así el alcance de las simulaciones y predicciones.

En cuanto a la ampliación geográfica, se contempla la integración de las ciudades autónomas de Ceuta y Melilla mediante el uso de estimaciones basadas en fuentes complementarias o nuevas conexiones con servicios de datos. A largo plazo, también se abre la posibilidad de extender el proyecto a contextos internacionales o a regiones que cuenten con estructuras energéticas comparables, lo cual permitiría evaluar la adaptabilidad del sistema a distintas realidades regulatorias y climáticas.