UNIVERSIDAD POLITÉCNICA DE MADRID

Escuela Técnica Superior de Ingeniería de
SISTEMAS INFORMÁTICOS

MADRID

GRADO EN SISTEMAS DE INFORMACIÓN

Análisis de los datos de Calidad del Aire de la Comunidad de Madrid

Autores:
Katarina **AVAKIMYAN**
Antonio **MERINO RECIO**

Tutor:
Pedro Pablo **ALARCÓN**

JUNIO 2023

Resumen

En este Proyecto Fin de Grado vamos a realizar un análisis exhaustivo de los datos de la calidad del aire de la Comunidad de Madrid con el fin de realizar un cuadro de mando que aporte gran valor a la toma de decisiones, de manera ágil y con calidad.

En primer lugar, obtendremos nuestros datos de partida de la web de la Comunidad, revisaremos la información importada y realizaremos un pulido de información. Dicho pulido consta de adaptar el formato de los datos al que nos conviene para ser tratados en los procesos.

Posteriormente, detallaremos las herramientas a utilizar en el proyecto, Pentaho, Microsoft SQL Server y Power BI, además reflejaremos los motivos por los que han sido elegidas las herramientas mencionadas entre sus semejantes en el mercado.

Comenzaremos el proyecto definiendo la arquitectura de la solución a realizar, posteriormente los componentes que usaremos y los creados por nosotros.

Diseñaremos el Data Warehouse acorde a la solución planteada y los procesos de extracción transformación y carga de datos. Para finalmente configurar un cuadro de mando en Power BI.

Por último, realizaremos el análisis del cuadro de mando y obtendremos las conclusiones acordes a los datos reflejados.

Abstract

In this project we will perform an exhaustive analysis of the air quality data of the community of Madrid in order to create a dashboard that provides great value to decision making, in an agile and quality way.

First of all, we will obtain our starting data from the community's website, we will review the imported information and we will polish the information. This polishing consists of adapting the format of the data to the one that suits us to be treated in the continuous processes.

Subsequently, we will detail the tools to be used in the project, Pentaho, SQL Server and Power BI, in addition we will reflect the reasons why the mentioned tools have been chosen among their peers in the market.

We will start the project by defining the architecture of the solution to be implemented, then the components that we will use and those created by us.

We will design the Data Warehouse according to the proposed solution and the data extraction, transformation and loading processes. Finally, we will configure a dashboard in Power BI.

Finally, we will perform the analysis of the dashboard and we will obtain the conclusions according to the data reflected.

ÍNDICE DEL PROYECTO

1. Introducción **6**
1.1. Contexto del problema 7
1.2. Solución planteada 8
1.2.1. Objetivo principal 10
1.2.2. Objetivos secundarios 10
1.3. Contenido de la memoria 11
2. Datos de partida **11**
2.1. Conjunto de Datos Calidad del Aire 11
2.2. Conjunto de Datos Calidad del Aire Meteo 12
2.3. Conjunto de Datos de la calidad del aire de las estaciones 13
2.4. Conjunto de Datos de la calidad del aire por zonas 16
3. Herramientas Software **17**
3.1. Pentaho 17
3.1.1. Spoon 19
3.2. SQL Server 19
3.3. Power BI 21
4. Arquitectura de la solución **23**
4.1. Componentes de la solución 23
4.2. Componentes creados 24
4.2.1. Data Warehouse 24
4.2.2. Pentaho 25
4.2.3. Power BI 26
5. Data Warehouse **27**
5.1. Esquema Data Warehouse 28
5.2. Dimensión Calendario 29
5.3. Dimensión Estación 30
5.4. Dimensión Magnitudes 31
5.5. Dimensión Municipio 31
5.6. Tabla de hechos calidad del aire 32
5.7. Tabla de hechos meteorología 32
5.8. Script de creación 33
6. Procesos ETL **35**
6.1. Dimensión Municipio 35
6.2. Dimensión Estación 38
6.3. Dimensión Magnitudes 45
6.4. Dimensión Calendario 48
6.5. Tabla de hechos Calidad del aire 50
6.6. Tabla de hechos Metereología 53
6.7. Validación del Data Warehouse 54
6.8. Carga de las dimensiones 61
6.9. Carga de las tablas de hechos 63

6.10. Carga total 65
7. Cuadro de Mandos 66
8. Análisis de datos 74
8.1. Conclusión del análisis de datos 81
9. Conclusión 83
9.1. Resumen de aportaciones del trabajo 83
9.2. Vías de continuación 83
10. Impactos sociales y medioambientales del proyecto 84
Bibliografia 86

1.Introducción

Hemos centrado este proyecto en la Comunidad de Madrid, en donde uno de los principales problemas es la contaminación del aire.

Como líder de fuentes de contaminación nos encontramos al tráfico rodado, ya que este supera con creces las emisiones provenientes de aterrizaje y despegue del aeropuerto de Barajas, fuente que se encuentra en tercer lugar

En la ciudad de Madrid el tráfico rodado es con gran diferencia la principal fuente de emisiones de óxido de nitrógeno, esto es, que son los vehículos los que lideran con el 46,9 % este ranking negativo de causantes de la contaminación, según el 'Inventario de emisiones de contaminantes a la atmósfera en el municipio de Madrid 2016'.
El siguiente sector de esta lista de causas es el que agrupa los llamados "otros medios de transporte", que se atribuye a los ciclos de aterrizaje y despegue del aeropuerto de Barajas. Tal y como recoge el Inventario, el 25,5 % del total proceden de este origen, si bien sobre ellos no cabe actuar desde la administración local.
El tercer puesto de este podio le corresponde al sector residencial, comercial e institucional, que se refiere fundamentalmente a los sistemas de climatización no industrial, que representan un 18,1 % de las emisiones de óxidos de nitrógeno. Por último, el sector industrial ocupa el cuarto lugar sin llegar a alcanzar el 7 % de las emisiones.

Pero en este ranking de contaminantes, el impacto del tráfico es aún mayor si tenemos en cuenta la concentración a nivel del suelo o inmisión, ya que el porcentaje correspondiente al tráfico rodado se eleva hasta el 75 % en la estación de medición de calidad del aire donde más influye, según el 'Estudio para la cuantificación de la contribución de fuentes a los niveles de calidad del aire en el municipio de Madrid', realizado por la Escuela Técnica Superior de Ingenieros Industriales de la UPM. El tráfico es el factor que más afecta en todas las estaciones, siendo el 35% el porcentaje mínimo en una estación. Según este estudio, el sector residencial supone entre un 3 y un 16 % de la aportación anual en función de la estación.
Dentro del sector residencial, cabe destacar que en 2017 había 574 calderas de carbón, según los cálculos municipales. Aunque el número no es muy alto, es el sistema de calefacción menos sostenible. Por este motivo, el Plan A de Calidad del Aire y Cambio Climático recoge la eliminación de este combustible en la ciudad en 2020, si bien la competencia de renovación de calderas es de la Comunidad de Madrid.

A lo largo de este proyecto realizaremos una extracción, pulido y transformación de datos para posteriormente volcar dicha información en un cuadro de mandos, el cual facilitará a los analistas la toma de decisiones, en menor tiempo y de mayor calidad.

1.1. Contexto del problema

La calidad del aire en la Comunidad de Madrid ha sido un tema de preocupación en los últimos años debido a los altos niveles de contaminación en algunas áreas urbanas de la región. La Comunidad de Madrid es una de las regiones más pobladas de España, con más de 6 millones de habitantes y una superficie de 8.000 km². Esto ha llevado a un aumento en la demanda de transporte, energía y servicios, lo que a su vez ha generado altos niveles de emisiones de contaminantes atmosféricos.

El Plan de Calidad del Aire y Cambio Climático de la Comunidad de Madrid 2013-2020 establece una serie de medidas para mejorar la calidad del aire en la región. El plan incluye la reducción de emisiones de contaminantes atmosféricos, el fomento del transporte público y la movilidad sostenible, la promoción de la energía limpia y la protección de la salud de la población.

En la Comunidad de Madrid, los contaminantes atmosféricos más comunes son el dióxido de nitrógeno (NO2), el ozono troposférico (O3) y las partículas en suspensión (PM10 y PM2.5). Estos contaminantes pueden tener efectos negativos en la salud de la población, como enfermedades respiratorias y cardiovasculares.

La Agencia Europea de Medio Ambiente (EEA) publica informes anuales sobre la calidad del aire en las ciudades europeas. En su informe de 2020, la EEA señaló que la ciudad de Madrid todavía tiene niveles preocupantes de contaminación atmosférica. Los niveles de dióxido de nitrógeno en algunas áreas urbanas de la ciudad superan los límites establecidos por la UE, lo que puede tener un impacto negativo en la salud de la población.

Para combatir la contaminación del aire en la Comunidad de Madrid, se han implementado una serie de medidas como la restricción del tráfico en el centro de la ciudad, el fomento del uso de vehículos eléctricos y la promoción del transporte público y la movilidad sostenible. Además, se están llevando a cabo campañas de concienciación para informar a la población sobre los efectos negativos de la contaminación del aire y la importancia de reducir las emisiones de contaminantes.

En resumen, la calidad del aire en la Comunidad de Madrid es un tema de preocupación debido a los altos niveles de contaminación atmosférica en algunas áreas urbanas. Se están implementando medidas para reducir las emisiones de contaminantes y mejorar la calidad del aire en la región, con el objetivo de proteger la salud de la población y el medio ambiente.

1.2. Solución planteada

Se evalúa la creación de un cuadro de mandos para dar legibilidad a todos los datos de la calidad del aire publicados en la web de la Comunidad de Madrid.
Actualmente, en el apartado "Datos abiertos" encontramos un amplio catálogo de diferentes conjuntos de datos que podemos descargar en formato csv, json y/o pdf, como se muestra en la siguiente imagen.

1.2.1. Ejemplo de conjunto de datos - CCMM Datos Abiertos

Al seleccionar el conjunto de datos de ejemplo observamos cierta dificultad para realizar cualquier tipo de análisis sobre los datos. Puesto que se trata de datos brutos, véase en la siguiente imagen.

_id	provincia	municipio	estacion	magnitud	punto_...	ano	mes	dia	h01	v01
21	28	13	2	10	2801300...	2023	2	18	7	T
22	28	13	2	12	2801300...	2023	2	18	11	T
23	28	13	2	14	2801300...	2023	2	18	67	T
24	28	133	2	7	2813300...	2023	2	18	1	T
25	28	133	2	8	2813300...	2023	2	18	14	T
26	28	133	2	10	2813300...	2023	2	18	14	T
27	28	133	2	12	2813300...	2023	2	18	16	T
28	28	133	2	14	2813300...	2023	2	18	58	T
29	28	14	2	7	2801400...	2023	2	18	9	T
30	28	14	2	8	2801400...	2023	2	18	55	T
31	28	14	2	10	2801400...	2023	2	18	18	T
32	28	14	2	12	2801400...	2023	2	18	69	T
33	28	14	2	14	2801400...	2023	2	18	16	T
34	28	14	2	14	2801400...	2023	2	18		N
35	28	148	4	7	2814800...	2023	2	18	10	T
36	28	148	4	8	2814800...	2023	2	18	40	T
37	28	148	4	9	2814800...	2023	2	18	13	T
38	28	148	4	10	2814800...	2023	2	18	21	T
39	28	148	4	12	2814800...	2023	2	18	55	T
40	28	148	4	14	2814800...	2023	2	18	2	T
41	28	16	1	1	2801600...	2023	2	18	1	T
42	28	16	1	6	2801600...	2023	2	18	0.5	T

1.2.2. Ejemplo de datos obtenidos - CCMM Datos Abiertos

Con el fin de proporcionar una mayor legitimidad a los datos, lo cual se traduce directamente por un mejor entendimiento de datos, se plantea crear una solución que tome todos los datos referentes a la calidad del aire de la web de la Comunidad de Madrid y mediante un procesos de transformación de la información y posterior volcado de dicha información ofrezca al ciudadano una mejor experiencia visual, legibilidad de datos y entendimiento de los mismos.

El ciudadano podrá navegar en la solución desde su equipo local y tendrá acceso a numerosos filtros que le proporcionarán un mayor nivel de detalle de la información.

Entre las diferentes posibilidades de filtrado de datos se plantea poder analizar municipios de manera independiente y/o únicamente los que sean relevantes para el estudio, observar en base a las magnitudes meteorológicas los diferentes comportamientos a lo largo del tiempo de los agentes contaminantes descritos en el apartado anterior.

1.2.3. Ejemplo de cuadro de mandos PowerBI

1.2.1. Objetivo principal

El objetivo principal del proyecto es ofrecer al ciudadano de manera más clara el comportamiento de los agentes contaminantes dentro de nuestro entorno. Concienciar a la población de que se trata de un problema de todos, y al que a todos nos afecta.

1.2.2. Objetivos secundarios

El análisis de datos de calidad del aire tiene varios objetivos que se pueden clasificar en diferentes categorías. A continuación, se describen algunos de los objetivos más comunes del análisis de datos de calidad del aire:

- Monitorizar la contaminación del aire: Uno de los objetivos principales del análisis de datos de calidad del aire es monitorizar la calidad del aire en una determinada área geográfica. Los datos recopilados permiten determinar los niveles de contaminación del aire y establecer un histórico de la calidad del aire en el tiempo.

- Identificación de fuentes de contaminación: El análisis de datos de calidad del aire también se utiliza para identificar las fuentes de contaminación en una determinada área. Esto ayuda a los reguladores ambientales y a las autoridades locales a tomar medidas para reducir las emisiones contaminantes y mejorar la calidad del aire.

- Evaluación del impacto en la salud: El análisis de datos de calidad del aire también se utiliza para evaluar el impacto en la salud de la población expuesta a la contaminación del aire. Los datos recopilados permiten establecer la relación entre la exposición a contaminantes específicos y la aparición de enfermedades respiratorias y cardiovasculares.

- Evaluación del impacto en el medio ambiente: El análisis de datos de calidad del aire también se utiliza para evaluar el impacto en el medio ambiente de la contaminación del aire. Los datos recopilados permiten determinar los efectos de la contaminación del aire en los ecosistemas y los recursos naturales.

- Evaluación de las políticas ambientales: El análisis de datos de calidad del aire también se utiliza para evaluar la eficacia de las políticas ambientales y las regulaciones en la mejora de la calidad del aire. Los datos recopilados permiten determinar si las medidas implementadas están teniendo un impacto positivo en la calidad del aire y si se necesitan ajustes para mejorar la eficacia de las políticas.

- La creación de procesos ETL para el refinamiento de los datos proporcionados por la Comunidad de Madrid, así como un Data Warehouse para su almacenamiento y finalmente un cuadro de mando para poder dar una visión gráfica de esta calidad del aire a cualquier persona.

- En resumen, los objetivos del análisis de datos de calidad del aire incluyen el monitoreo de la contaminación del aire, la identificación de fuentes de contaminación, la evaluación del impacto en la salud y el medio ambiente, y la evaluación de las políticas ambientales. Estos objetivos tienen como objetivo principal mejorar la calidad del aire y proteger la salud y el bienestar de la población y el medio ambiente.

1.3. Contenido de la memoria

Comenzaremos el proyecto obteniendo la información de los datos abiertos de la Comunidad de Madrid, en concreto, el conjunto de datos referentes a la calidad del aire.
Posteriormente, analizaremos la información existente y utilizando procesos de extracción, transformación y carga de datos añadiremos la lógica correspondiente para la normalización de la información para finalmente cargar en la base de datos el contenido normalizado.
Finalmente, crearemos un cuadro de mandos para volcar toda la información mencionada de manera legible y clara ayudando así a la toma de decisiones.

2. Datos de partida

En este apartado describiremos qué datos y de dónde obtendremos los mismos para llevar a cabo la realización de este proyecto, así como los diferentes procesos de normalización de datos que hemos realizado.

En primer lugar, hemos obtenido los datos de la página web de la Comunidad de Madrid, del apartado “Datos Abiertos”, más concretamente en la sección de “Medio Ambiente”.

Obtenemos los ficheros descritos a continuación.

2.1. Conjunto de Datos Calidad del Aire

Conjunto de Datos Calidad del Aire: En donde obtenemos información referente a las magnitudes de medida de la calidad del aire de las diferentes provincias, municipios, estaciones agrupadas por fecha y hora. Cada hora tendrá su columna H0x y por cada hora tenemos una columna V0x que hará referencia a la validación de la hora.

Nombre del fichero csv: *calidad_aire_datos_2021_2022*

provincia	municipio	estacion	magnitud	punto_muestreo	ano	mes	dia	h01	v01	h02	v02	h03	v03	h04	v04	h05	v05	h06	v06	h07	v07	h08	v08	h09	v09	h10	v10	h11	v11	h12	v12	h13	v13	h14	v14	h15	v15	h16	v16	h17	v17	h18	v18	h19	v19	h20	v20	h21	v21	h22	v22	h23	v23	h24	v24
28	102	1	82	28102001_82_89	2021	1	1	173	V	138	V	136	V	141	V	145	V	134	V	140	V	138	V	161	V	159	V	178	V	184	V	164	V	116	V	171	V	164	V	159	V	97	V	127	V	173	V	184	V	184	V	127	V	156	V
28	102	1	83	28102001_83_89	2021	1	1	7,5	V	7,1	V	6,9	V	7,1	V	7,2	V	6,9	V	6,8	V	7,1	V	7,6	V	9,1	V	11	V	14	V	15	V	14	V	14	V	14	V	13	V	13	V	12	V	11	V	11	V	11	V	11	V	10	V
28	102	1	86	28102001_86_89	2021	1	1	50	V	51	V	50	V	49	V	47	V	48	V	47	V	44	V	43	V	40	V	38	V	34	V	30	V	36	V	35	V	33	V	33	V	34	V	38	V	40	V	37	V	38	V	37	V	41	V
28	102	1	87	28102001_87_89	2021	1	1	932	V	932	V	932	V	931	V	931	V	931	V	931	V	931	V	931	V	931	V	931	V	930	V	930	V	929	V	928	V	928	V	928	V	928	V	928	V	928	V	928	V	928	V	928	V	929	V
28	102	1	88	28102001_88_89	2021	1	1	14	V	14	V	14	V	14	V	14	V	14	V	14	V	17	V	62	V	178	V	333	V	602	V	525	V	331	V	257	V	153	V	60	V	15	V	15	V	15	V	15	V	15	V	15	V	15	V
28	102	1	89	28102001_89_89	2021	1	1	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V
28	120	1	81	28120001_81_89	2021	1	1	1,8	V	2,3	V	2,6	V	2	V	1,7	V	1,1	V	2,1	V	2	V	2,9	V	2	V	2,6	V	4,1	V	4,8	V	4,7	V	3,7	V	3,8	V	3,1	V	3,9	V	4,1	V	3,7	V	4,1	V	3,7	V	4,1	V	3,6	V
28	120	1	82	28120001_82_89	2021	1	1	205	V	212	V	218	V	226	V	210	V	171	V	225	V	211	V	219	V	225	V	223	V	216	V	217	V	210	V	212	V	211	V	227	V	227	V	221	V	206	V	222	V	218	V	220	V	211	V
28	120	1	83	28120001_83_89	2021	1	1	5,7	V	5,6	V	5,9	V	5,6	V	5,5	V	4,9	V	5,2	V	4,9	V	4,6	V	4,4	V	5,4	V	6,1	V	7	V	7,2	V	6	V	5,9	V	5,2	V	5,2	V	5,2	V	4,9	V	4,9	V	4,3	V	3,2	V	3,3	V
28	120	1	86	28120001_86_89	2021	1	1	68	V	69	V	81	V	81	V	83	V	83	V	86	V	82	V	89	V	89	V	77	V	60	V	60	V	56	V	55	V	51	V	53	V	55	V	57	V	60	V	65	V	71	V	82	V	84	V
28	120	1	87	28120001_87_89	2021	1	1	818	V	818	V	818	V	817	V	817	V	817	V	817	V	817	V	817	V	818	V	817	V	816	V	815	V	815	V	814	V	814	V	814	V	814	V	814	V	814	V	814	V	814	V	814	V	814	V
28	120	1	88	28120001_88_89	2021	1	1	1	V	1	V	1	V	1	V	1	V	1	V	1	V	2	V	24	V	91	V	345	V	371	V	580	V	419	V	188	V	136	V	38	V	2	V	1	V	1	V	1	V	1	V	1	V	1	V
28	120	1	89	28120001_89_89	2021	1	1	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0,6	V	0,2	V	0,1	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0,1	V
28	123	2	81	28123002_81_89	2021	1	1	1,5	V	1,5	V	1,4	V	1,1	V	1,3	V	1	V	1	V	1,1	V	1,4	V	1,6	V	1,5	V	1,2	V	1,2	V	1,8	V	1,3	V	0,4	V	0,2	V	0,8	V	0,7	V	0,4	V	1,2	V	3,2	V	2,6	V	2,5	V
28	123	2	82	28123002_82_89	2021	1	1	91	V	68	V	66	V	57	V	9	V	4	V	31	V	24	V	89	V	67	V	58	V	84	V	101	V	154	V	165	V	319	V	353	V	45	V	27	V	56	V	184	V	164	V	164	V	214	V
28	123	2	83	28123002_83_89	2021	1	1	7,3	V	7	V	6,9	V	7	V	7,3	V	7,1	V	7,7	V	7,9	V	8,7	V	10	V	11	V	13	V	15	V	17	V	17	V	17	V	16	V	13	V	12	V	12	V	13	V	14	V	14	V	14	V
28	123	2	86	28123002_86_89	2021	1	1	72	V	73	V	72	V	71	V	68	V	68	V	65	V	67	V	65	V	60	V	56	V	49	V	41	V	37	V	38	V	39	V	45	V	56	V	58	V	54	V	51	V	38	V	38	V	44	V
28	123	2	87	28123002_87_89	2021	1	1	958	V	957	V	957	V	957	V	957	V	957	V	957	V	957	V	957	V	957	V	957	V	956	V	955	V	954	V	953	V	953	V	953	V	953	V	953	V	953	V	953	V	953	V	954	V	955	V
28	123	2	88	28123002_88_89	2021	1	1	8	V	8	V	8	V	8	V	8	V	8	V	8	V	8	V	63	V	131	V	228	V	356	V	377	V	416	V	190	V	107	V	44	V	8	V	8	V	8	V	8	V	8	V	8	V	8	V
28	123	2	89	28123002_89_89	2021	1	1	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V
28	13	2	81	28013002_81_89	2021	1	1	0,3	V	0,1	V	0,3	V	0,2	V	0,1	V	0,2	V	0,3	V	0,2	V	0,2	V	0,2	V	0,3	V	0,2	V	0,5	V	0,5	V	0,6	V	0,4	V	0,8	V	0,7	V	0,8	V	1	V	0,9	V	2,4	V	1,9	V	1,3	V
28	13	2	82	28013002_82_89	2021	1	1	250	V	43	V	47	V	43	V	42	V	48	V	313	V	28	V	45	V	204	V	68	V	115	V	199	V	167	V	137	V	7	V	172	V	165	V	153	V	169	V	163	V	182	V	175	V	198	V
28	13	2	83	28013002_83_89	2021	1	1	6,4	V	6	V	5,3	V	5,3	V	5,6	V	5,5	V	5,6	V	6	V	5,8	V	6,9	V	9,6	V	13	V	15	V	16	V	18	V	18	V	17	V	16	V	16	V	15	V	15	V	16	V	16	V	14	V
28	13	2	86	28013002_86_89	2021	1	1	74	V	74	V	76	V	75	V	73	V	75	V	75	V	72	V	74	V	68	V	57	V	47	V	39	V	37	V	34	V	37	V	35	V	36	V	33	V	33	V	31	V	29	V	29	V	38	V
28	13	2	87	28013002_87_89	2021	1	1	964	V	964	V	964	V	964	V	963	V	963	V	963	V	963	V	963	V	964	V	964	V	963	V	962	V	961	V	960	V	960	V	960	V	960	V	960	V	960	V	960	V	960	V	961	V	962	V
28	13	2	88	28013002_88_89	2021	1	1	0	V	0	V	0	V	0	V	0	V	0	V	0	V	1	V	35	V	92	V	175	V	250	V	210	V	152	V	202	V	75	V	22	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V
28	13	2	89	28013002_89_89	2021	1	1	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V
28	133	2	81	28133002_81_89	2021	1	1	0,5	V	0,5	V	0,4	V	0,9	V	0,5	V	0,7	V	0,6	V	0,8	V	1	V	1,3	V	0,9	V	0,8	V	1,5	V	1,9	V	1,9	V	0,8	V	0,5	V	1	V	1,1	V	1,9	V	1,5	V	1,6	V	3,1	V	2,5	V
28	133	2	82	28133002_82_89	2021	1	1	266	V	168	V	127	V	115	V	124	V	172	V	224	V	84	V	291	V	239	V	175	V	265	V	186	V	179	V	188	V	219	V	100	V	93	V	121	V	95	V	100	V	62	V	269	V	256	V
28	133	2	83	28133002_83_89	2021	1	1	9	V	9,6	V	10	V	10	V	10	V	9,5	V	9,7	V	9,5	V	9,4	V	9	V	9,8	V	11	V	14	V	15	V	15	V	14	V	13	V	12	V	12	V	11	V	13	V	13	V	11	V	10	V
28	133	2	86	28133002_86_89	2021	1	1	78	V	73	V	61	V	60	V	61	V	67	V	63	V	65	V	69	V	78	V	76	V	68	V	50	V	39	V	39	V	50	V	56	V	49	V	49	V	52	V	48	V	49	V	68	V	93	V
28	133	2	87	28133002_87_89	2021	1	1	943	V	942	V	942	V	942	V	941	V	941	V	941	V	942	V	942	V	942	V	942	V	941	V	940	V	939	V	938	V	938	V	938	V	938	V	937	V	937	V	937	V	938	V	939	V	939	V
28	133	2	88	28133002_88_89	2021	1	1	1	V	1	V	1	V	1	V	1	V	1	V	1	V	1	V	21	V	52	V	142	V	218	V	462	V	441	V	282	V	80	V	34	V	1	V	1	V	1	V	1	V	1	V	1	V	1	V

2.1. Cabeceras del fichero calidad_aire_datos_2021_2022

Columna	Tipo	Descripción
provincia	numeric	Identificador provincia
municipio	numeric	Identificador municipio
estacion	numeric	Identificador estacion
magnitud	numeric	Identificador magnitud
punto_muestreo	text	Compuesto por el codigo de la estación seguido de el identificador de la magnitud
ano	numeric	Año
mes	numeric	Mes
dia	numeric	Día
h01	text	Hora
v01	text	Validación de su respectiva H0x
h02	text	Hora
v02	text	Validación de su respectiva H0x
h03	text	Hora
v03	text	Validación de su respectiva H0x
h04	text	Hora
v04	text	Validación de su respectiva H0x

2.1.2. Tabla de cabeceras

2.2. Conjunto de Datos Calidad del Aire Meteo

Conjunto de Datos Calidad del Aire Meteo: En donde obtenemos información referente a las magnitudes de calidad del aire de meteorología de las diferentes provincias, municipios y estaciones agrupadas por fecha y hora. Cada hora tendrá su columna H0x y por cada hora tenemos una columna V0x que hará referencia a la validación de la hora.

Nombre del fichero csv: *calidad_aire_datos_meteo_2021_2022*

provincia	municipio	estacion	magnitud	punto_mues	ano	mes	dia	h01	v01	h02	v02	h03	v03	h04	v04	h05	v05	h06	v06	h07	v07	h08	v08	h09	v09
28	102	1	82	28102001_8	2021	1	1	173	V	138	V	136	V	141	V	145	V	134	V	140	V	138	V	161	V
28	102	1	83	28102001_8	2021	1	1	7,5	V	7,1	V	6,9	V	7,1	V	7,2	V	6,9	V	6,8	V	7,1	V	7,6	V
28	102	1	86	28102001_8	2021	1	1	50	V	51	V	50	V	49	V	47	V	48	V	47	V	44	V	43	V
28	102	1	87	28102001_8	2021	1	1	932	V	932	V	932	V	931	V	931	V	931	V	931	V	931	V	931	V
28	102	1	88	28102001_8	2021	1	1	14	V	14	V	14	V	14	V	14	V	14	V	14	V	17	V	62	V
28	102	1	89	28102001_8	2021	1	1	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V
28	120	1	81	28120001_8	2021	1	1	1,8	V	2,3	V	2,6	V	2	V	1,7	V	1,1	V	2,1	V	2	V	2,9	V
28	120	1	82	28120001_8	2021	1	1	205	V	212	V	218	V	226	V	210	V	171	V	225	V	211	V	219	V
28	120	1	83	28120001_8	2021	1	1	5,7	V	5,6	V	5,9	V	5,6	V	5,5	V	4,9	V	5,2	V	4,9	V	4,6	V
28	120	1	86	28120001_8	2021	1	1	68	V	69	V	81	V	81	V	83	V	83	V	86	V	82	V	89	V
28	120	1	87	28120001_8	2021	1	1	818	V	818	V	818	V	817	V	817	V	817	V	817	V	817	V	817	V
28	120	1	88	28120001_8	2021	1	1	1	V	1	V	1	V	1	V	1	V	1	V	1	V	2	V	24	V
28	120	1	89	28120001_8	2021	1	1	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0,6	V
28	123	2	81	28123002_8	2021	1	1	1,5	V	1,5	V	1,4	V	1,1	V	1,3	V	1	V	1	V	1,1	V	1,4	V
28	123	2	82	28123002_8	2021	1	1	91	V	68	V	66	V	57	V	9	V	4	V	31	V	24	V	89	V
28	123	2	83	28123002_8	2021	1	1	7,3	V	7	V	6,9	V	7	V	7,3	V	7,1	V	7,7	V	7,9	V	8,7	V
28	123	2	86	28123002_8	2021	1	1	72	V	73	V	72	V	71	V	68	V	68	V	65	V	67	V	65	V
28	123	2	87	28123002_8	2021	1	1	958	V	957	V	957	V	957	V	957	V	957	V	957	V	957	V	957	V
28	123	2	88	28123002_8	2021	1	1	8	V	8	V	8	V	8	V	8	V	8	V	8	V	8	V	63	V
28	123	2	89	28123002_8	2021	1	1	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V	0	V
28	13	2	81	28013002_8	2021	1	1	0,3	V	0,1	V	0,3	V	0,2	V	0,1	V	0,2	V	0,3	V	0,2	V	0,2	V
28	13	2	82	28013002_8	2021	1	1	250	V	43	V	47	V	43	V	42	V	48	V	313	V	28	V	45	V
28	13	2	83	28013002_8	2021	1	1	6,4	V	6	V	5,3	V	5,3	V	5,6	V	5,5	V	5,6	V	6	V	5,8	V

2.2. Cabeceras del fichero calidad_aire_datos_meteo_2021_2022

Columna	Tipo	Descripción
provincia	numeric	Identificador provincia
municipio	numeric	Identificador municipio
estacion	numeric	Identificador estacion
magnitud	numeric	Identificador magnitud
punto_muestreo	text	Compuesto por el codigo de la estación seguido de el identificador de la magnitud
ano	numeric	Año
mes	numeric	Mes
dia	numeric	Día
h01	text	Hora
v01	text	Validación de su respectiva H0x
h02	text	Hora
v02	text	Validación de su respectiva H0x
h03	text	Hora
v03	text	Validación de su respectiva H0x
h04	text	Hora
v04	text	Validación de su respectiva H0x

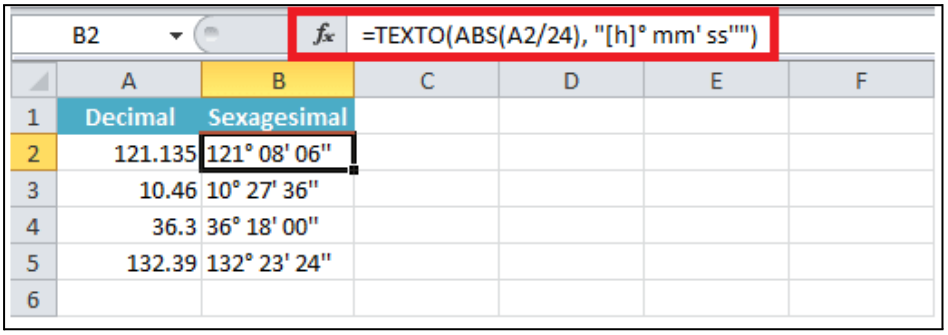
2.2.1. Tabla de cabeceras

2.3. Conjunto de Datos de la calidad del aire de las estaciones

Conjunto de Datos de la calidad del aire de las estaciones: En este archivo obtenemos información referente a las diferentes estaciones de medición de la calidad del aire, contamos con el identificador, el tipo de estación, las coordenadas de localización y los diferentes analizadores con los que cuenta cada estación.

En este fichero hemos realizado la normalización de los datos de las coordenadas tanto de latitud como de longitud, en donde hemos aplicado diferentes fórmulas para pasar los datos existentes en formato de coordenadas a numéricos para que fueran entendibles por Power BI.

Los cambios han sido aplicados a la columna de coordenadas longitud y latitud usando la fórmula del siguiente ejemplo:



2.3.1. Fórmula de transformación de coordenadas

Nombre del fichero csv: *calidad_aire_estaciones*

estacion_codigo	zona_calidad_aire_descripcion	estacion_municipio	estacion_fecha_alta	estacion_tipo_area	estacion_tipo_estacion	estacion_subarea_rural	estacion_direccion_postal	estacion_coord_UTM_ETRS89_x	estacion_coord_UTM_ETRS89_y	est
28006001	Zona 2 Corredor del Henares	Alcalá de Henares	2000/01/16 00:00:00.000	Urbana	Tráfico	No aplica	C.E.I.P. Manuel Isaña, Avda. del Ejército, 5	462828	4481011	03
28006004	Zona 2 Corredor del Henares	Alcobendas	2000/01/19 00:00:00.000	Urbana	Tráfico	No aplica	Parque de Andalucía, C/ Pintor Murillo	445363	4487950	03
28007004	Zona 3 Urbana Sur	Alcorcón	2008/09/11 00:00:00.000	Urbana	Fondo	No aplica	Colegio Blas de Otero, Avda. del Oeste, 4	429190	4466041	03
28009001	Zona 2 Corredor del Henares	Algete	2007/03/31 00:00:00.000	Suburbana	Fondo	No aplica	Parque Europa, C/ Humilladero con C/ Francisco Quevedo	457418	4494449	03
28013002	Zona 3 Urbana Sur	Aranjuez	2001/10/06 00:00:00.000	Suburbana	Fondo	No aplica	Polideportivo Municipal Agustín Marañón, C/ Moreras (C/ Primero de Mayo, 4)	449541	4431618	03
28014002	Zona 2 Corredor del Henares	Arganda del Rey	2006/04/25 00:00:00.000	Urbana	Industrial	No aplica	Punto Limpio, C/ Río Tajuña, 5	461002	4461246	03
28016001	Zona 5 Sierra Norte	El Atazar	2006/08/22 00:00:00.000	Rural	Fondo	Remota	Instalaciones del Canal de Isabel II	460596	4528767	03
28045002	Zona 4 Urbana Noroeste	Colmenar Viejo	2001/10/06 00:00:00.000	Urbana	Tráfico	No aplica	Auditorio Municipal, C/ Molino de Viento, s/n	434586	4501818	03
28047002	Zona 4 Urbana Noroeste	Collado Villalba	2006/05/31 00:00:00.000	Urbana	Tráfico	No aplica	Colegio Público Rosa Chacel, C/ Santa Teresa de Jesús, 12	414233	4498525	04
28049003	Zona 2 Corredor del Henares	Coslada	2007/06/05 00:00:00.000	Urbana	Tráfico	No aplica	Polideportivo Municipal Valle Aguado, Avda. José Gárate, 8 (C/ Perú, 17)	454001	4475678	03
28058004	Zona 3 Urbana Sur	Fuenlabrada	2000/05/01 00:00:00.000	Urbana	Industrial	No aplica	C/ Grecia, 3 (trasera)	431913	4459313	03
28065014	Zona 3 Urbana Sur	Getafe	2006/10/31 00:00:00.000	Urbana	Tráfico	No aplica	C.E.I.P. Mariana Pineda, Avda. de las Ciudades, 33	439090	4462920	03
28067001	Zona 5 Sierra Norte	Guadalix de la Sierra	2006/10/31 00:00:00.000	Rural	Fondo	Regional	Camino de Chozas, s/n	440751	4514644	03
28074007	Zona 3 Urbana Sur	Leganés	2000/05/01 00:00:00.000	Urbana	Tráfico	No aplica	C.E.P.A. Rosalía de Castro, C/ Roncal (Avda. de la Mancha, 50)	435914	4465744	03
28080003	Zona 4 Urbana Noroeste	Majadahonda	2001/10/06 00:00:00.000	Suburbana	Fondo	No aplica	Campo de golf, C/ Isaac Albéniz	426307	4477645	03
28092005	Zona 3 Urbana Sur	Móstoles	2000/05/01 00:00:00.000	Suburbana	Fondo	No aplica	Parque Liana	425516	4464113	03
28102001	Zona 7 Cuenca del Tajuña	Orusco de Tajuña	2006/09/29 00:00:00.000	Rural	Fondo	Remota	Camino forestal de acceso al repetidor de telefonía	481205	4459700	03
28120001	Zona 5 Sierra Norte	Puerto de Cotos (Rascafría)	2019/01/01 00:00:00.000	Rural	Fondo	Remota	Centro de Visitantes Peñalara, Ctra. M-604, km 42 (Puerto de Cotos, Rascafría)	418948	4519788	03
28123002	Zona 2 Corredor del Henares	Rivas-Vaciamadrid	2006/08/02 00:00:00.000	Suburbana	Fondo	No aplica	Polideportivo Cerro del Telégrafo, C/ Juncal (Avda. Los Almendros, 1)	453906	4467829	03
28133002	Zona 6 Cuenca del Alberche	San Martín de Valdeiglesias	2006/08/24 00:00:00.000	Rural	Fondo	Regional	C/ Depósito s/n	381298	4469513	04
28148004	Zona 2 Corredor del Henares	Torrejón de Ardoz	2005/01/01 00:00:00.000	Suburbana	Fondo	No aplica	Parque del Ocio	459497	4477765	03
28161001	Zona 3 Urbana Sur	Valdemoro	2006/07/27 00:00:00.000	Suburbana	Fondo	No aplica	Colegio Público Pedro Antonio de Alarcón, C/ Artemisa, 5	442089	4448540	03
28171001	Zona 6 Cuenca del Alberche	Villa del Prado	2006/06/14 00:00:00.000	Rural	Fondo	Remota	Hospital Virgen de la Poveda, C/ Alejandro Peris Barrios, km. 5	391504	4456144	04
28180001	Zona 7 Cuenca del Tajuña	Villarejo de Salvanés	2006/04/25 00:00:00.000	Suburbana	Tráfico	No aplica	I.E.S. Nuestra Señora de la Victoria de Lepanto, C/ Luis de Requesens, 1-3	476442	4446354	03

2.3. Cabeceras del fichero calidad_aire_estaciones -I

estacion_coord_longitud	estacion_coord_latitud	estacion_altitud	estacion_analizador_NO	estacion_analizador_NO2	estacion_analizador_PM10	estacion_analizador_PM2_5	estacion_analizador_O3	estacion_analizador_TOL	estacion_analizador_BEN	estacion_analizador_XIL	estacion_analizador_CO	estacion_analizador_SO2
03°38'42,7"W	40°32'25,8"N	688	Si	Si	Si	No	Si	Si	Si	Si	No	No
03°50'01,4"W	40°20'30,9"N	693	Si	Si	No	Si	Si	No	No	No	No	No
03°30'11,8"W	40°35'59,1"N	800	Si	Si	No	Si	Si	No	No	No	No	No
03°35'29,1"W	40°01'59,8"N	501	Si	Si	Si	No	Si	No	No	No	No	No
03°27'32,0"W	40°18'02,9"N	586	Si	Si	Si	No	Si	No	No	No	No	No
03°28'04,4"W	40°54'32,6"N	995	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
03°46'26,0"W	40°39'52,7"N	905	Si	Si	Si	No	Si	No	No	No	Si	No
04°00'51,1"W	40°37'59,2"N	873	Si	Si	No	Si	Si	Si	Si	Si	No	Si
03°32'32,3"W	40°25'49,7"N	621	Si	Si	Si	Si	Si	No	No	No	No	No
03°48'03,4"W	40°16'53,5"N	699	Si	Si	Si	No	Si	Si	Si	Si	No	No
03°43'00,7"W	40°18'52,5"N	667	Si	Si	Si	Si	Si	No	No	No	No	No
03°42'07,8"W	40°46'50,3"N	853	Si	Si	Si	No	Si	No	No	No	No	No
03°45'16,3"W	40°20'23,2"N	676	Si	Si	Si	Si	Si	No	No	No	No	No
03°52'08,4"W	40°26'46,3"N	360	Si	Si	Si	No	Si	No	No	No	No	No
03°52'36,3"W	40°19'27,2"N	660	Si	Si	Si	No	Si	No	No	No	Si	Si
03°13'16,0"W	40°17'15,3"N	800	Si	Si	Si	No	Si	No	No	No	Si	Si
03°57'40,3"W	40°49'30,4"N	1870	Si	Si	Si	Si	Si	No	No	No	No	No
03°32'34,3"W	40°21'35,1"N	608	Si	Si	Si	No	Si	No	No	No	No	No
04°23'53,2"W	40°22'03,9"N	549	Si	Si	Si	No	Si	No	No	No	No	No
03°28'39,5"W	40°26'58,4"N	597	Si	Si	Si	Si	Si	No	No	No	No	No
03°40'49,0"W	40°11'06,9"N	615	Si	Si	No	Si	Si	No	No	No	No	No
04°16'32,4"W	40°14'55,4"N	510	Si	Si	Si	Si	Si	No	No	No	Si	Si
03°16'36,0"W	40°10'02,0"N	764	Si	Si	No	Si	Si	No	No	No	No	No

2.4. Cabeceras del fichero calidad_aire_estaciones -II

Columna	Tipo	Etiqueta	Descripción
estacion_codigo	numeric		Código numérico de la estación
zona_calidad_aire_descripcion	text		Zona en la que se ubica la estación, según zonificación de la Red de Calidad del Aire de la Comunidad de Madrid
estacion_municipio	text		Municipio donde se localiza la estación
estacion_fecha_alta	timestamp		Fecha en la que se dio de alta la estación
estacion_tipo_area	text		Tipo de área en la que se encuentra la estación (urbana, suburbana o rural)
estacion_tipo_estacion	text		Tipo de estación (fondo, tráfico o industrial)
estacion_subarea_rural	text		Tipo de subárea rural (remota o regional)

2.3.1. Tabla de cabeceras

estacion_direccion_postal	text		Dirección postal de la estación
estacion_coord_UTM_ETRS89_x	numeric		Coordenada X de la estación según sistema UTM ETRS89
estacion_coord_UTM_ETRS89_y	numeric		Coordenada Y de la estación según sistema UTM ETRS89
estacion_coord_longitud	text		Longitud de localización de la estación
estacion_coord_latitud	text		Latitud de localización de la estación
estacion_altitud	numeric		Altitud de la estación, en metros
estacion_analizador_NO	text		Analizador de óxido de nitrógeno (NO)
estacion_analizador_NO2	text		Analizador de dióxido de nitrógeno (NO2)
estacion_analizador_PM10	text		Analizador de partículas en suspensión de tamaño inferior a 10 micras (PM10)

2.3.2. Tabla de cabeceras

estacion_analizador_PM2_5	text		Analizador de partículas en suspensión de tamaño inferior a 2,5 micras (PM2.5)
estacion_analizador_O3	text		Analizador de ozono (O3)
estacion_analizador_TOL	text		Analizador de tolueno (TOL)
estacion_analizador_BEN	text		Analizador de benceno (BEN)
estacion_analizador_XIL	text		Analizador de xileno (XIL)
estacion_analizador_CO	text		Analizador de monóxido de carbono (CO)
estacion_analizador_SO2	text		Analizador de dióxido de azufre (SO2)
estacion_analizador_HCT	text		Analizador de hidrocarburos totales (HCT)
estacion_analizador_HNM	text		Analizador de hidrocarburos no metánicos (HNM)

2.3.3. Tabla de cabeceras

2.4. Conjunto de Datos de la calidad del aire por zonas

Conjunto de Datos de la calidad del aire por zonas: En este archivo obtenemos información referente a las diferentes zonas de la Comunidad de Madrid en donde se realizan mediciones de la calidad del aire.

Nombre del archivo csv: *calidad_aire_zonas*

zona_calidad_aire_codigo	zona_calidad_aire_descripcion	zona_calidad_aire_municipio
1	MADRID	MADRID
2	CORREDOR DEL HENARES	AJALVIR
2	CORREDOR DEL HENARES	ALCALÁ DE HENARES
2	CORREDOR DEL HENARES	ALCOBENDAS
2	CORREDOR DEL HENARES	ALGETE
2	CORREDOR DEL HENARES	ANCHUELO
2	CORREDOR DEL HENARES	ARGANDA DEL REY
2	CORREDOR DEL HENARES	CAMARMA DE ESTERUELAS
2	CORREDOR DEL HENARES	COBEÑA
2	CORREDOR DEL HENARES	COSLADA
2	CORREDOR DEL HENARES	DAGANZO DE ARRIBA
2	CORREDOR DEL HENARES	LOECHES

2.4. Cabeceras del fichero calidad_aire_zonas

Columna	Tipo	Etiqueta	Descripción
zona_calidad_aire_codigo	numeric		Código numérico de zona, según zonificación de la Red de Calidad del Aire de la Comunidad de Madrid
zona_calidad_aire_descripcion	text		Denominación de cada zona e la Red de Calidad del Aire
zona_calidad_aire_municipio	text		Municipios de la Comunidad de Madrid que integran cada zona

2.4.1. Tabla de cabeceras

3. Herramientas Software

En este apartado se van a detallar las características de las herramientas que han sido utilizadas para la realización del proyecto, evidenciando las ventajas de uso, así como la facilidad de relación entre ellas, los tipos de conexión posibles que se pueden realizar y la arquitectura de trabajo en cada una de ellas.

3.1. Pentaho

Se trata de una plataforma orientada a soluciones de Business Intelligence centrada en procesos que incluyen actividades para realizar la lógica deseada, de esta forma se implementan soluciones basadas en procesos, entre otros, ETL, generación de informes y minería de datos. Se trata de una de las herramientas pioneras en mejora de la capacidad de análisis y de toma de decisiones.

La plataforma Pentaho es propiedad de Hitachi Vantara. Fue creada en el año 2004 y actualmente lidera todas las soluciones de Business Intelligence Open Source, todas las soluciones de Pentaho están compuestas por código Java y el entorno de implementación también basado en Java, lo cual proporciona a Pentaho una gran flexibilidad y adaptabilidad para cualquier necesidad empresarial al usar este sistema tecnologías estándar.
Pentaho posee entornos básicos de desarrollo tanto en Pentaho Data Integration (PDI) como en BI server, ambos entornos tienen innumerables capacidades de desarrollo por medio de técnicas ETL cuyos datos posteriormente serán representados en forma de Cuadro de Mandos, los cuales muchas empresas utilizan como base para la elaboración de informes o la realización de seguimientos de cumplimientos de objetivos, acuerdos de nivel de servicio (SLAs),etc.

Esto se traduce en que todas aquellas empresas que cuenten con estos sistemas o herramientas van a conseguir grandes ventajas como, por ejemplo, poder mejorar su capacidad de análisis, tomar decisiones de una forma mucho más rápida, reducir el tiempo que se tarda en recopilar la información o mejorar el seguimiento y gestión de su estrategia.

Pentaho Data Integration, cuyo nombre en clave es Kettle, es una de las herramientas o componentes de Pentaho Suite que permite que se utilicen técnicas ETL, es decir, poder implementar procesos de extracción, transformación y carga de datos. Kettle, además, ofrece datos analíticos muy precisos, eliminando las complejidades involucradas en la codificación al proporcionar bibliotecas en profundidad para el mismo.

Proporciona a las empresas un ahorro de tiempo y de esfuerzo automatizando procesos manuales.
Entre las numerosas ventajas que tiene esta plataforma cabe destacar las presentadas a continuación:

- Software libre: Pentaho es un software libre sujeto a diferentes licencias, una de las más interesantes es la GPLv2, la cual implica el uso, copia distribución y modificación gratuitamente. Como conclusión a esta ventaja, tenemos la posibilidad de implantación de un sistema sin desembolso.

  Además de la versión libre, se ofrece una versión de pago por parte de los propietarios, dicha versión destaca entre otros por un servicio de soporte al usuario, en donde se podrán registrar las incidencias con el software. Al ser una herramienta multiplataforma funciona en Linux, Windows, etc.

- Velocidad: Cuenta con una rapidez de procesamiento y transformación de la información muy elevada, por lo tanto la toma de decisiones tras el procesamiento y transformación se realiza de manera más rápida y ágil. Una toma de decisiones más rápida tiene entre otras consecuencias una mayor fluidez entre las capas de la empresa, mejora el ROI (Retorno de la Inversión) puesto que el cálculo de rentabilidad de las inversiones, campañas o acciones será mucho más eficiente aumentando en última instancia la eficiencia de los empleados.

- Data mining y servidor OLAP: En cuanto al procesamiento de datos, facilita la interacción con los datos para identificar grupos o patrones y generar predicciones para futuros análisis respaldados por algoritmos de Machine Learning. Dicha facilidad en combinación con su poderoso servidor OLAP nos proporcionan una de las ventajas más importantes del aplicativo.

La suite Pentaho BI está compuesta por diferentes módulos que son:

- Pentaho Analysis Services: basado en Mondrian que es un servidor OLAP (procesamiento analítico en línea) escrito en Java. Es compatible con MDX (expresiones multidimensionales) y el lenguaje de consulta XML para el Análisis y especificaciones de la interfaz olap4j.

- Pentaho Reporting: Consiste en un motor de presentación, capaz de generar informes programáticos sobre la base de un archivo de definición XML.

- Pentaho Data Mining: Es una envoltura alrededor del proyecto Weka (Minería de datos). Es una suite de software que usa estrategias de aprendizaje de máquina, aprendizaje automático y minería de datos. Cuenta con series de clasificación, de

regresión, de reglas de asociación, y de algoritmos de clustering, para así apoyar las tareas de análisis predictivo.

- Pentaho Dashboard: Es una plataforma integrada para proporcionar información sobre sus datos, donde se pueden ver informes, gráficos interactivos y los cubos creados con las herramientas Pentaho Report Designer.

- Pentaho para Apache Hadoop: Es un conector de bajo nivel para facilitar el acceso a grandes volúmenes manejados en el proyecto Apache Hadoop.

Desventajas de Pentaho:

- Soporte limitado: Aunque hay una comunidad activa detrás de Pentaho, el soporte técnico puede ser limitado en comparación con otras plataformas comerciales.

- Integración con otros sistemas: Pentaho puede tener problemas de integración con otros sistemas, lo que puede requerir conocimientos técnicos avanzados para solucionarlos.

- Rendimiento: Pentaho puede ser lento en el procesamiento de grandes conjuntos de datos, lo que puede afectar el rendimiento.

- Personalización: Aunque Pentaho es altamente personalizable, puede requerir conocimientos técnicos avanzados para personalizarlo según las necesidades de una empresa.

- Funcionalidades limitadas: Pentaho puede tener funcionalidades limitadas en comparación con otras plataformas comerciales, lo que puede limitar su capacidad para manejar ciertos tipos de análisis de datos.

El componente Spoon es la interfaz gráfica de PDI, Gracias a Spoon vamos a poder realizar procesos de ETL de manera fácil y rápida. Permite hacer un Data Warehouse, pudiendo formar las tablas de Hechos y Dimensiones y sus relaciones entre ellas sin ninguna dificultad. Prácticamente nos va a permitir hacer de todo: crear conexiones a los datos, hacer todo tipo de transformaciones, insertar fórmulas, transformaciones directas ya implementadas gracias a su calculadora… Sencillamente en pocos pasos te permite hacer muchas cosas sin tener que programar directamente con código.

3.2. Microsoft SQL Server

Microsoft SQL Server es un sistema de gestión de bases de datos relacional desarrollado por Microsoft. Es utilizado por empresas y organizaciones para almacenar, organizar y recuperar grandes cantidades de datos. SQL Server utiliza el lenguaje SQL (Structured Query Language) para administrar y manipular los datos.

Entre las características principales de SQL Server se encuentran:

- Escalabilidad: SQL Server está diseñado para manejar grandes cantidades de datos y usuarios simultáneos, lo que lo hace adecuado para empresas y organizaciones de diferentes tamaños.

- Seguridad: SQL Server ofrece opciones avanzadas de seguridad, como autenticación y autorización, encriptación de datos y auditoría, para garantizar que los datos estén protegidos contra accesos no autorizados.

- Integración: SQL Server se integra con otras herramientas y plataformas de Microsoft, como Visual Studio y Azure, lo que permite a los desarrolladores crear y desplegar aplicaciones y soluciones de datos en una amplia variedad de plataformas.

- Alta disponibilidad: SQL Server ofrece opciones avanzadas de alta disponibilidad, como replicación de datos y clustering, para garantizar que los datos estén disponibles en todo momento.
- Análisis de datos: SQL Server ofrece herramientas de análisis de datos, como Integration Services, Analysis Services y Reporting Services, para ayudar a los usuarios a analizar y visualizar los datos.

En resumen, SQL Server es un sistema de gestión de bases de datos relacional utilizado por empresas y organizaciones para almacenar, organizar y recuperar grandes cantidades de datos de manera segura y eficiente.

Ventajas de SQL Server:

- Escalabilidad: SQL Server es altamente escalable y puede manejar grandes volúmenes de datos.

- Alta disponibilidad: SQL Server ofrece opciones de alta disponibilidad, como la replicación y la agrupación de conmutación por error, para garantizar que los datos estén disponibles en todo momento.

- Integración con otras tecnologías de Microsoft: SQL Server se integra fácilmente con otras tecnologías de Microsoft, como .NET Framework y Microsoft Office.

- Seguridad: SQL Server ofrece características avanzadas de seguridad, como la autenticación de Windows y la encriptación de datos, para proteger los datos almacenados.

- Soporte: Microsoft ofrece un amplio soporte para SQL Server, incluyendo documentación, foros en línea y servicios de asistencia técnica.

Desventajas de SQL Server:

- Costo: SQL Server puede ser costoso, especialmente para empresas que necesitan licencias para múltiples usuarios.

- Requerimientos de hardware: Para ejecutar SQL Server correctamente, se necesita un hardware potente, lo que puede resultar costoso.

- Complejidad: SQL Server puede ser complejo y requiere conocimientos técnicos avanzados para su configuración y mantenimiento.

- Limitaciones de la edición Express: La versión gratuita de SQL Server, la edición Express, tiene algunas limitaciones en términos de tamaño de base de datos y número de usuarios simultáneos.

- Propietario: SQL Server es un producto propietario de Microsoft, lo que significa que las empresas que utilizan SQL Server están limitadas a las características y funcionalidades proporcionadas por Microsoft.

3.3. Microsoft Power BI

Power BI es una plataforma de análisis de datos que permite a los usuarios conectarse a diversas fuentes de datos, transformarlos y visualizarlos de una manera fácil y efectiva. Power BI es una herramienta desarrollada por Microsoft que permite a los usuarios crear informes y cuadros de mando interactivos, que se pueden utilizar para analizar y tomar decisiones basadas en los datos.

Power BI es una herramienta de análisis de datos que se utiliza para extraer datos de diversas fuentes, como bases de datos, hojas de cálculo y servicios en la nube, y luego transformarlos en modelos de datos que se pueden utilizar para crear informes y cuadros de mando. Power BI utiliza una interfaz de usuario intuitiva que permite a los usuarios crear visualizaciones de datos utilizando gráficos, tablas, mapas y otros elementos visuales.

Power BI también ofrece herramientas de análisis de datos avanzadas, como modelado de datos, análisis de tendencias y predicción, que permiten a los usuarios descubrir información oculta en los datos. Power BI también permite a los usuarios compartir informes y cuadros de mando con otros usuarios y trabajar juntos en tiempo real, lo que fomenta la colaboración.

Una de las ventajas de Power BI es su capacidad para integrarse con otras herramientas de Microsoft, como Excel, SharePoint y Azure. Esto significa que los usuarios pueden utilizar Power BI en combinación con estas herramientas para crear soluciones de análisis de datos más avanzadas.

Power BI también ofrece opciones avanzadas de seguridad y cumplimiento para garantizar que los datos estén protegidos y cumplan con las regulaciones. Esto incluye la capacidad de configurar el acceso a los datos a nivel de usuario, la encriptación de datos y la auditoría de datos.

Power BI se ofrece en una variedad de versiones, incluyendo una versión gratuita y una versión de pago. La versión gratuita tiene limitaciones en cuanto a la cantidad de datos que se

pueden procesar y la cantidad de informes y cuadros de mando que se pueden crear. La versión de pago ofrece más funcionalidades y capacidades avanzadas de análisis de datos.

Para utilizar Power BI, los usuarios deben tener conocimientos básicos de análisis de datos y visualización de datos. Power BI también requiere conocimientos técnicos para la configuración y el mantenimiento, lo que puede ser una limitación para los usuarios que no tienen experiencia en tecnología.

En resumen, Power BI es una plataforma de análisis de datos que permite a los usuarios conectar, transformar y visualizar datos de diversas fuentes. Power BI ofrece herramientas avanzadas de análisis de datos, opciones de seguridad y cumplimiento, y la capacidad de integrarse con otras herramientas de Microsoft. Power BI es una herramienta útil para empresas y organizaciones que necesitan analizar grandes cantidades de datos para tomar decisiones informadas.

A continuación se presentan algunas ventajas y desventajas de Power BI:

Ventajas de Power BI:

- Integración de datos: Power BI se integra con una amplia gama de fuentes de datos, lo que permite la extracción, transformación y carga de datos de múltiples fuentes.

- Visualización de datos: Power BI ofrece una amplia variedad de opciones para visualizar los datos, como gráficos, tablas y mapas, lo que permite a los usuarios comprender mejor la información.

- Análisis de datos avanzado: Power BI ofrece herramientas avanzadas de análisis de datos, como modelado de datos, análisis de tendencias y predicción, lo que permite a los usuarios descubrir información oculta en los datos.

- Colaboración y compartición: Power BI permite a los usuarios compartir informes y cuadros de mando con otros usuarios y trabajar juntos en tiempo real, lo que fomenta la colaboración.

- Seguridad y cumplimiento: Power BI ofrece opciones avanzadas de seguridad y cumplimiento para garantizar que los datos estén protegidos y cumplan con las regulaciones.

Desventajas de Power BI:

- Costo: Power BI es una herramienta de pago, lo que puede ser una limitación para las empresas que tienen un presupuesto limitado.

- Requiere conocimientos técnicos: Power BI requiere conocimientos técnicos para la configuración y el mantenimiento, lo que puede ser una limitación para los usuarios que no tienen experiencia en tecnología.

- Limitaciones en la versión gratuita: La versión gratuita de Power BI tiene limitaciones en cuanto a la cantidad de datos que se pueden procesar y la cantidad de informes y cuadros de mando que se pueden crear.

- Integración con otros sistemas: Power BI puede tener problemas de integración con otros sistemas, lo que puede requerir conocimientos técnicos avanzados para solucionarlos.

- Rendimiento: Power BI puede ser lento en el procesamiento de grandes conjuntos de datos, lo que puede afectar el rendimiento

4. Arquitectura de la solución

En este apartado se detallan los diferentes componentes y procesos definidos para la realización de este proyecto.

4.1. Componentes de la solución

En primer lugar obtendremos los datos a procesar de la página web oficial de la Comunidad de Madrid, teniendo varios formatos de descarga a elegir (json, csv,etc.). Hemos seleccionado el formato csv debido a la facilidad de manejo que nos proporciona, así como, la ventaja de poder consultar dicha información de manera más ágil.

Posteriormente, alimentaremos los datos añadiendo las magnitudes de medida necesarias para el correcto funcionamiento del proceso.
Además, realizaremos una normalización de los datos descargados para simplificar los procesos ETL.

Utilizaremos la herramienta spoon incluida en Pentaho, para la carga de la información en la base de datos y posteriormente mostraremos la información por medio de un cuadro de mandos creado en la herramienta Power BI.

Los componentes que forman la arquitectura de la solución nos proporcionan cierta facilidad para la integración de la información.

Pentaho nos proporciona la herramienta spoon que nos permite crear una conexión con Microsoft SQL Server utilizando la autenticación de Windows junto al controlador JDBC.

Esto nos proporciona conectividad con bases de datos a través de las interfaces de programación de aplicaciones estándar de JDBC disponibles en Java Platform, Enterprise Edition.

Power BI nos permite crear una conexión con la base de datos SQL Server seleccionando la opción de importar datos incluyendo las diferentes relaciones existentes en las tablas.

Por lo tanto, cabe destacar la agilidad con la que se pueden conectar los componentes y traspasar la información entre ellos.

4. Arquitectura de la solución

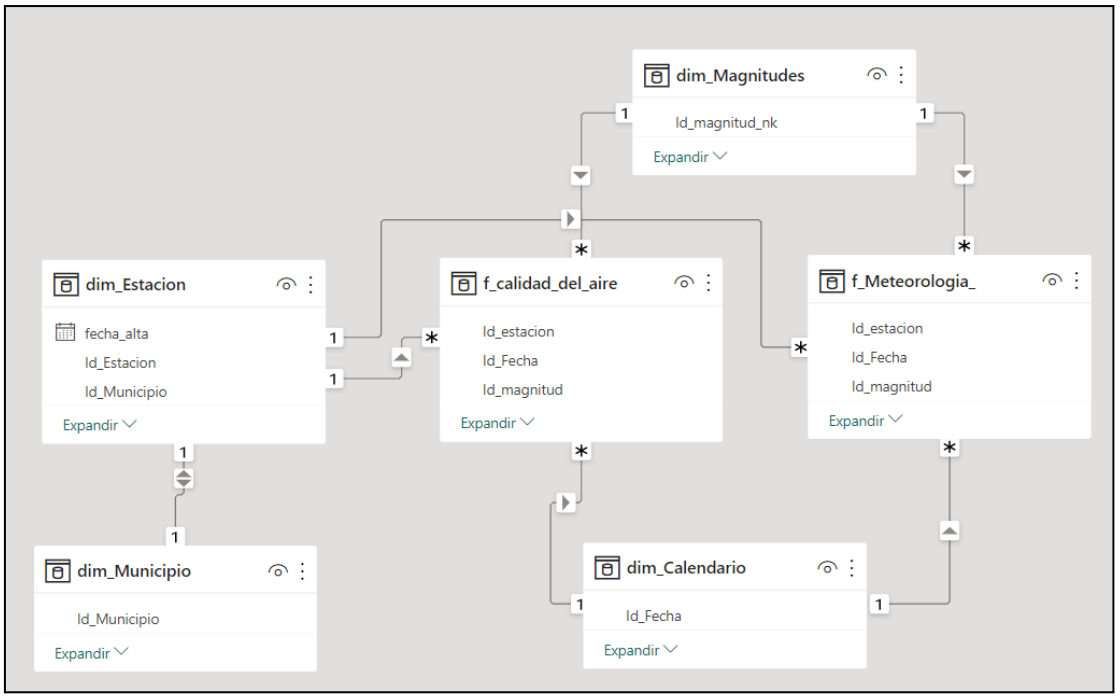
4.2. Componentes creados

En este apartado se detallarán los diferentes componentes que hemos creado para llevar a cabo la realización del proyecto.

4.2.1. Data Warehouse

Para la base de datos que va a conformar el data warehouse diseñaremos un script de creación para cada una de las tablas necesarias, 6 en total, siendo 4 tablas de dimensiones y 2 tablas de hechos, cada una de estas tablas tendrá una clave única para así poder mantener su integridad referencial.

Por lo tanto, obtenemos el esquema de tipo copo de nieve. Cuya definición realizaremos en el punto 5. Diseño del Data Warehouse:



4.2.1.1. Data Warehouse

4.2.2. Pentaho

Usaremos Pentaho para transformar los datos que obtenemos de la Comunidad de Madrid e incluirlos en el modelo del Data Warehouse que hemos creado.

La herramienta se divide en 2 principales elementos, transformaciones y jobs. Las transformaciones son el elemento más básico, con ellas podemos realizar todas las funciones de extracción, transformación e inserción de datos.

Los jobs son un conjunto de transformaciones agrupadas siguiendo una determinada lógica. También podemos encontrarnos agrupaciones de diferentes jobs en uno solo, lo que nos proporciona mayor agilidad en la ejecución del proceso.

Vamos a utilizar una transformación por cada tabla de dimensiones y tabla de hechos que tenemos en nuestro modelo, y finalmente utilizaremos los jobs para realizar una comprobación de la base de datos así como una carga conjunta para poder tener este proceso semi-automático. Los principales factores a tener en cuenta a la hora de crear estas transformaciones son: El formato de los datos en el origen y el tipo de base de datos que se utilice para el Data Warehouse.

En nuestro caso disponemos de los datos en el origen en forma de ficheros CSV, lo que hace un proceso más laborioso su obtención y refinamiento que si obtenemos los datos desde una BBDD.

Debido a la volumetría de información de nuestro proyecto crearemos las tablas del datawarehouse en SQL Server, puesto que contamos con un volumen de datos de aproximadamente 500 mil filas.

Si la volumetría de datos superará las 2 millones de filas convendrá migrar esta BBDD a otro gestor especializado en el manejo de volúmenes de datos mucho más grandes, es decir, un gestor de base de datos especializado en big data.

4.2.3. Power BI

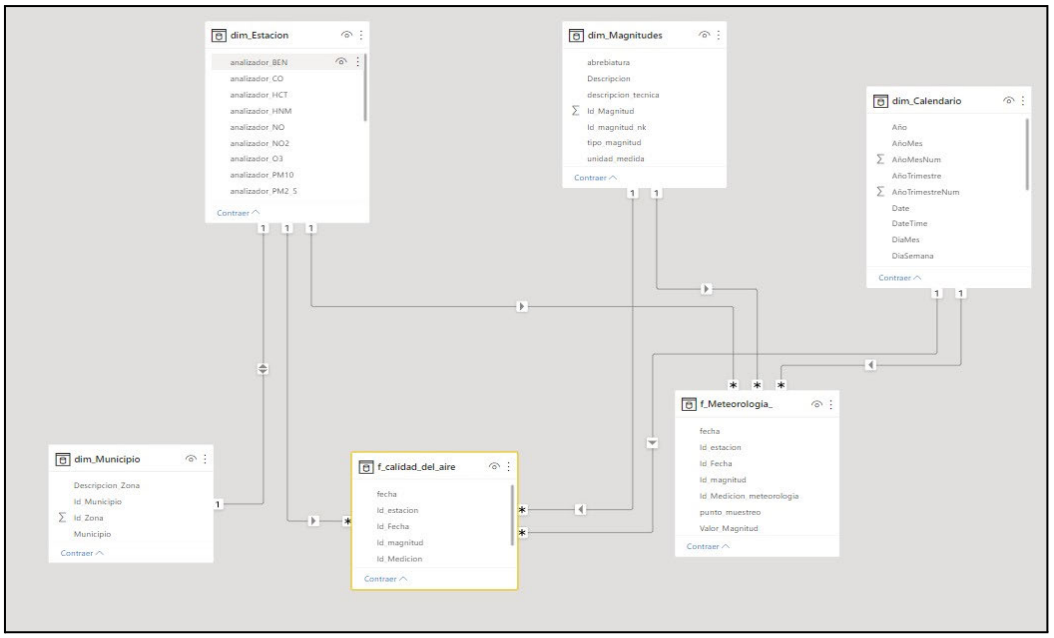
Se creará un informe de Power Bi para poder analizar los datos que previamente han sido refinados.

El objetivo de este informe será que mediante gráficos y tablas daremos un estudio de la calidad del aire pudiendo adaptar la forma de visualización de los datos según las necesidades que tengamos o el tipo de estudio que queramos realizar.

Desde un alto nivel o incluso llegando a centrarnos solo en casos muy concretos para así obtener una información muy completa para poder llegar a unas conclusiones sobre la calidad del aire que tengan una buena base.

5. Data Warehouse

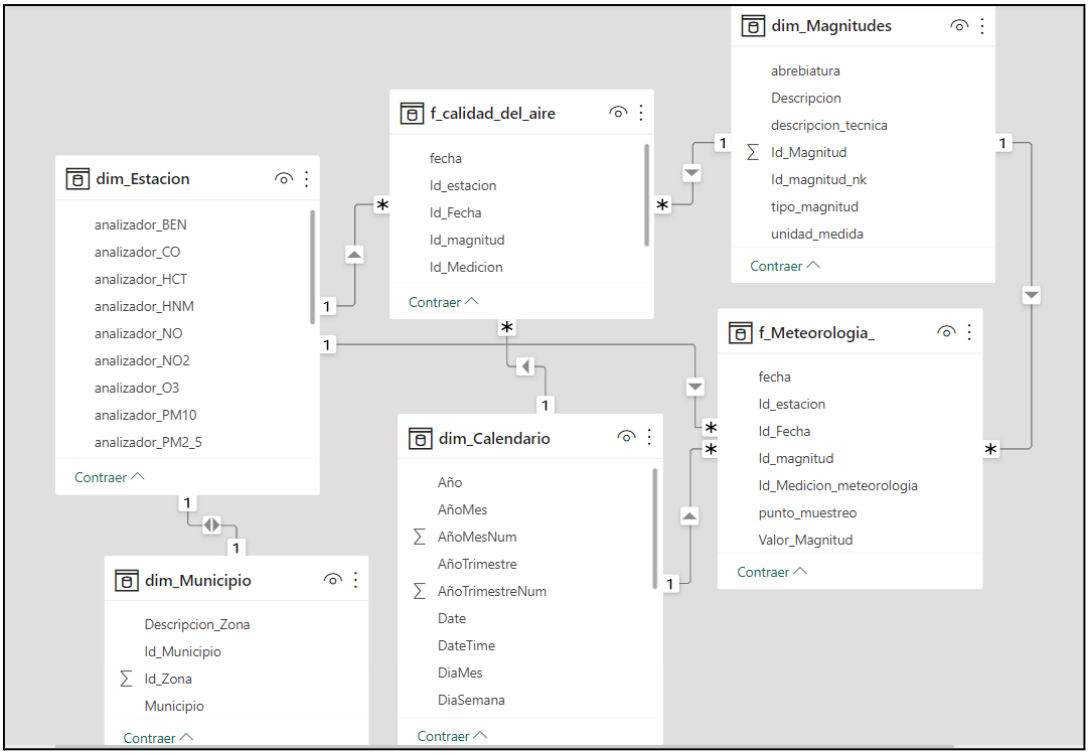
Idealmente, para el diseño de un Data Warehouse se debería intentar seguir un modelo en estrella, debido a los beneficios que este tiene a la hora de la utilización posterior de estos datos en Power Bi, el cual es uno de los objetivos principales para poder analizar los datos de manera correcta, esto es así debido a que Power Bi a la hora de representar los gráficos realiza consultas por debajo a la base de datos, por lo que cuantas más relaciones entre tablas tengamos estas consultas se realizarán de manera mucho menos eficiente.



5.1. Data Warehouse

5.1. Esquema Data Warehouse

En la siguiente figura se muestra el esquema del Data Warehouse diseñado:



5.2. Data Warehouse I

Debido a la poca cantidad de los datos en las dimensiones se ha decidido hacer un esquema copo de nieve para desacoplar la dimensión municipio de la dimensión estación para así poder hacer un análisis por zona de manera mucho más clara y eficiente ya que en este caso queremos primar la capacidad de análisis de los datos frente a la rapidez de procesamiento.

Para entender el concepto de Data Warehouse primero necesitamos entender sus componentes:

Una tabla de hechos es un objeto que contiene toda la información de un acto o la información de negocio que es sometida a actividades de análisis.

Se utiliza para representar la ocurrencia de un evento en un momento temporal concreto. Por lo tanto, una medida de hechos es válida únicamente para un instante y hecho concreto, al contener un grado de detalle elevado, da lugar a múltiples tablas de dimensiones.

Llamamos dimensión de una base de datos al conjunto de atributos dentro de una tabla que nos proporcionan información referente a los datos de una tabla de hechos. Es decir, en estas tablas contendremos el detalle asociado a la tabla de hechos. Además, dichos atributos son organizables en jerarquías definidas por los usuarios.

Como se ha comentado antes se debería intentar seguir el modelo en estrella lo máximo posible debido a que a la hora de la creación de las ETL cuantas menos relaciones entre tablas tengamos, menos pasos Lookup en la BBDD tendremos que realizar.

Estos suelen tardar en torno a 1 hora por cada 30.000 filas, y en este caso que tenemos algunos de los csv que al transformarlos tienen más de 1.000.000 de filas por lo que este paso ralentiza en gran medida la carga.

Como lo que queremos primar es el análisis de los datos y actualmente no está contemplada una actualización diaria en el alcance del proyecto no hay un riesgo en tener un esquema en copo de nieve.

El sistema de gestión de Bases de Datos elegido para alojar el Data Warehouse ha sido Microsoft SQL Server ya que este tiene muchas ventajas y su única desventaja sería el precio en caso de querer llevar esta solución a un entorno de producción.

Por lo que se convierte en una opción ideal para lo que estábamos buscando ya que al nivel que vamos a implementar nuestra solución podemos usar las características gratuitas que nos proporciona la herramienta.

A continuación, vamos a realizar una descripción de cada una de las dimensiones y tablas de hechos del Data Warehouse.

5.2. Dimensión Calendario

En esta dimensión encontramos campos que nos ayudan a obtener un estudio de los datos a nivel temporal muy exhaustivo.

La clave de la tabla es el campo “Id_Fecha” de tipo Integer, el cual representa una fecha con la hora incluida en formato numérico. A diferencia del campo “FechaId” de esta tabla que solo representa la fecha sin incluir la hora.

Ejemplo: La fecha: 23/04/2022 13:00:00 quedaría con el siguiente identificador:

23042022130000.

Column Name	Data Type	Allow Nulls
Id_Fecha	int	☐
Date	varchar(100)	☑
Time	varchar(100)	☑
DateTime	varchar(100)	☑
Hour	int	☑
FechaId	int	☑
Año	float	☑
Trimestre	varchar(100)	☑
AñoTrimestreNum	float	☑
AñoTrimestre	varchar(100)	☑
MesNum	float	☑
AñoMesNum	float	☑
AñoMes	varchar(100)	☑
MesCorto	varchar(100)	☑
Mes	varchar(100)	☑
DiaMes	float	☑
DiaSemanaNum	float	☑
DiaSemana	varchar(100)	☑
DiaSemanaCorto	varchar(100)	☑

5.1. Dimensión Calendario

5.3. Dimensión Estación

La dimensión tiene como objetivo dar datos de cada una de las estaciones de medición para así complementar la información de la tabla de hechos.

La clave de esta tabla es el campo “Id_Estacion”, el cual tiene formato integer y se genera de manera autoincremental.

Cabe destacar que esta dimensión tiene la clave foránea “Id_Municipio” para poder tener relación con la tabla Dim_Municipio que hace referencia a la Dimensaión Municipio.

Column Name	Data Type	Allow Nulls
Id_Estacion	int	☐
Id_Municipio	int	☑
cod_estacion	bigint	☑
fecha_alta	datetime	☑
tipo_area	varchar(9)	☑
tipo_estacion	varchar(10)	☑
cord_longitud	float	☑
cord_latitud	float	☑
analizador_NO	varchar(3)	☑
analizador_NO2	varchar(3)	☑
analizador_PM10	varchar(3)	☑
analizador_PM2_5	varchar(2)	☑
analizador_O3	varchar(3)	☑
analizador_TOL	varchar(2)	☑
analizador_BEN	varchar(2)	☑
analizador_XIL	varchar(2)	☑
analizador_CO	varchar(2)	☑
analizador_SO2	varchar(2)	☑
analizador_HCT	varchar(2)	☑
analizador_HNM	varchar(2)	☑

5.2. Dimensión Estación

5.4. Dimensión Magnitudes

La dimensión tiene como objetivo dar información sobre los distintos contaminantes y valores a medir que influyen principalmente en la calidad del aire y en el ámbito meteorológico.

La clave de esta tabla es el campo “Id_Magnitud”, el cual tiene formato integer y se genera de manera autoincremental.

Vemos también el campo “Id_Magnitud_nk”, el cual es el identificador numérico que la comunidad de madrid le da a cada una de las magnitudes.

Column Name	Data Type	Allow Nulls
Id_Magnitud	int	☑
Id_magnitud_nk	float	☑
Descripcion	varchar(100)	☑
abrebiatura	varchar(100)	☑
unidad_medida	varchar(100)	☑
descripcion_tecnica	varchar(100)	☑
tipo_magnitud	varchar(100)	☑

5.3. Dimensión Magnitudes

5.5. Dimensión Municipio

Esta dimensión es la más escueta y su principal función es permitir el análisis de los hechos por municipios y zonas de la Comunidad de Madrid.

La clave de esta tabla es el campo “Id_Municipio”, el cual tiene formato integer y se genera de manera autoincremental.

Column Name	Data Type	Allow Nulls
Id_Municipio	int	☑
Id_Zona	bigint	☑
Descripcion_Zona	varchar(40)	☑
Municipio	varchar(44)	☑

5.4. Dimensión Municipio

5.6. Tabla de hechos calidad del aire

Esta tabla tiene las mediciones sobre la calidad del aire, a su vez tiene todas las claves foráneas de las dimensiones para así relacionarse y poder complementar la información con lo que se necesite en el momento.

Esta tabla de hechos tiene clave de esta tabla es el campo “Id_Medicion”, el cual tiene formato integer y se genera de manera autoincremental.

Column Name	Data Type	Allow Nulls
Id_Medicion	int	☐
Id_Fecha	varchar(100)	☑
Id_estacion	bigint	☑
Id_magnitud	bigint	☑
fecha	varchar(40)	☑
Valor_Magnitud	bigint	☑
punto_muestreo	varchar(15)	☑

5.5. Tabla de hechos de calidad del aire

5.7. Tabla de hechos meteorología

Esta tabla tiene las mediciones sobre los datos meteorológicos, a su vez tiene todas las claves foráneas de las dimensiones para así relacionarse y poder complementar la información con lo que se necesite en el momento.

Esta tabla de hechos tiene clave de esta tabla es el campo “Id_Medicion_meteorologia”, el cual tiene formato integer y se genera de manera autoincremental.

Column Name	Data Type	Allow Nulls
Id_Medicion_meteorologia	int	☐
Id_Fecha	varchar(100)	☑
Id_estacion	bigint	☑
Id_magnitud	bigint	☑
fecha	varchar(40)	☑
Valor_Magnitud	decimal(5, 1)	☑
punto_muestreo	varchar(14)	☑

5.8. Script de creación de las tablas del Data Warehouse

Script de generación de las tablas del Data Warehouse:

```
CREATE TABLE dim_Municipio
(
  Id_Municipio INT
, Id_Zona BIGINT
, Descripcion_Zona VARCHAR(40)
, Municipio VARCHAR(44)
)
;
ALTER TABLE dbo.dim_Municipio ADD CONSTRAINT D_MUNICIPIO_PK PRIMARY KEY
(Id_Municipio);

CREATE TABLE dim_Calendario
(
  Id_Fecha INT
, "Date" VARCHAR(100)
, "Time" VARCHAR(100)
, DateTime VARCHAR(100)
, "Hour" INT
, FechaId INT
, Año FLOAT(53)
, Trimestre VARCHAR(100)
, AñoTrimestreNum FLOAT(53)
, AñoTrimestre VARCHAR(100)
, MesNum FLOAT(53)
, AñoMesNum FLOAT(53)
, AñoMes VARCHAR(100)
, MesCorto VARCHAR(100)
, Mes VARCHAR(100)
, DiaMes FLOAT(53)
, DiaSemanaNum FLOAT(53)
, DiaSemana VARCHAR(100)
, DiaSemanaCorto VARCHAR(100)
)
;
ALTER TABLE dbo.dim_Calendario ADD CONSTRAINT D_Calendario_PK PRIMARY KEY
(Id_Fecha);

CREATE TABLE dim_Estacion
(
  Id_Estacion INT NOT NULL
, Id_Municipio INT
, cod_estacion BIGINT
, fecha_alta DATETIME
```

```
        , tipo_area VARCHAR(9)
, tipo_estacion VARCHAR(10)
, cord_longitud NUMERIC (10,18)
, cord_latitud NUMERIC(10,18)
, analizador_NO VARCHAR(3)
, analizador_NO2 VARCHAR(3)
, analizador_PM10 VARCHAR(3)
, analizador_PM2_5 VARCHAR(2)
, analizador_O3 VARCHAR(3)
, analizador_TOL VARCHAR(2)
, analizador_BEN VARCHAR(2)
, analizador_XIL VARCHAR(2)
, analizador_CO VARCHAR(2)
, analizador_SO2 VARCHAR(2)
, analizador_HCT VARCHAR(2)
, analizador_HNM VARCHAR(2)
)
;

ALTER TABLE dbo.dim_Estacion ADD CONSTRAINT D_EStacion_PK PRIMARY KEY
(Id_Estacion);
ALTER TABLE dbo.dim_Estacion
ADD CONSTRAINT FK_ID_MUNICIPIO FOREIGN KEY (Id_Municipio)
    REFERENCES dbo.dim_Municipio (Id_Municipio);

CREATE TABLE dim_Magnitudes
(
 Id_Magnitud INT
, Id_magnitud_nk FLOAT(53)
, Descripcion VARCHAR(100)
, abrebiatura VARCHAR(100)
, unidad_medida VARCHAR(100)
, descripcion_tecnica VARCHAR(100)
, tipo_magnitud VARCHAR(100)
)
;
ALTER TABLE dbo.dim_Magnitudes ADD CONSTRAINT D_Magnitud_PK PRIMARY KEY
(Id_Magnitud);

CREATE TABLE f_calidad_del_aire
(
 Id_Medicion INT not null identity(1,1)
, Id_Fecha VARCHAR(100)
, Id_estacion BIGINT
, Id_magnitud BIGINT
, fecha VARCHAR(40)
, Valor_Magnitud BIGINT
, punto_muestreo VARCHAR(15)
)
;
ALTER TABLE dbo.f_calidad_del_aire ADD CONSTRAINT D_Medicion_PK PRIMARY KEY
(Id_Medicion);

ALTER TABLE dbo.f_calidad_del_aire
ADD CONSTRAINT FK_ID_FECHA FOREIGN KEY (Id_Fecha)
    REFERENCES dbo.dim_Calendario (Id_Fecha);

ALTER TABLE dbo.f_calidad_del_aire
ADD CONSTRAINT FK_ID_ESTACION FOREIGN KEY (Id_estacion)
    REFERENCES dbo.dim_Estacion (Id_estacion);
```

```
ALTER TABLE dbo.f_calidad_del_aire
ADD CONSTRAINT FK_ID_MAGNITUD FOREIGN KEY (Id_magnitud)
    REFERENCES dbo.dim_magnitud (Id_magnitud);
```

```
CREATE TABLE f_meteorologia
(
  Id_Medicion_meteorologia INT not null identity(1,1)
, Id_Fecha VARCHAR(100)
, Id_estacion BIGINT
, Id_magnitud BIGINT
, fecha VARCHAR(40)
, Valor_Magnitud DECIMAL(5,1)
, punto_muestreo VARCHAR(14)
)
;
ALTER TABLE dbo.f_meteorologia ADD CONSTRAINT D_Medicion_meteorologia_PK
PRIMARY KEY (Id_Medicion_meteorologia);

ALTER TABLE dbo.f_calidad_del_aire
ADD CONSTRAINT FK_ID_FECHA FOREIGN KEY (Id_Fecha)
    REFERENCES dbo.dim_Calendario (Id_Fecha);

ALTER TABLE dbo.f_calidad_del_aire
ADD CONSTRAINT FK_ID_ESTACION FOREIGN KEY (Id_estacion)
    REFERENCES dbo.dim_Estacion (Id_estacion);

ALTER TABLE dbo.f_calidad_del_aire
ADD CONSTRAINT FK_ID_MAGNITUD FOREIGN KEY (Id_magnitud)
    REFERENCES dbo.dim_magnitud (Id_magnitud);
```

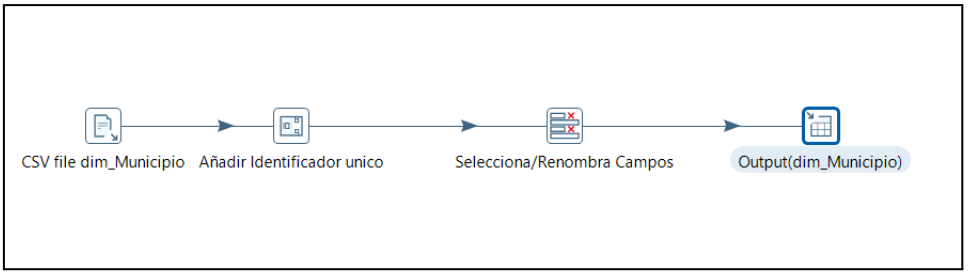
6. Procesos ETL

En este punto vamos a realizar una descripción de cada uno de los procesos ETL diseñados en el proyecto, así como una descripción más detallada de los mismos, las actividades usadas, y cómo se han configurado para llevar a cabo la transformación adecuada de los datos de partida.

Estos procesos han sido desarrollados mediante pentaho, haciendo uso de la herramienta spoon, ya que nos permite hacerlos de manera gráfica sin la necesidad de implementación de código.

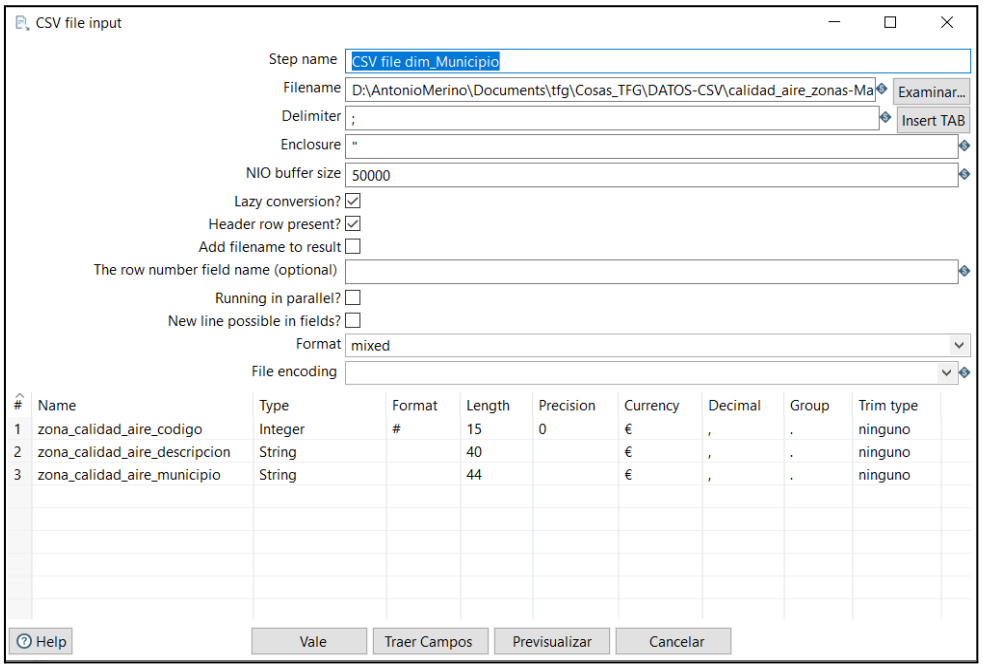
6.1. Dimensión Municipio

La dimensión Municipio será cargada con los datos del fichero csv *calidad_aire_zonas* y para ello usaremos la actividad CSV file, en donde añadiremos la ruta del archivo a leer y las cabeceras del fichero.



6.1.1. Dimensión Municipio

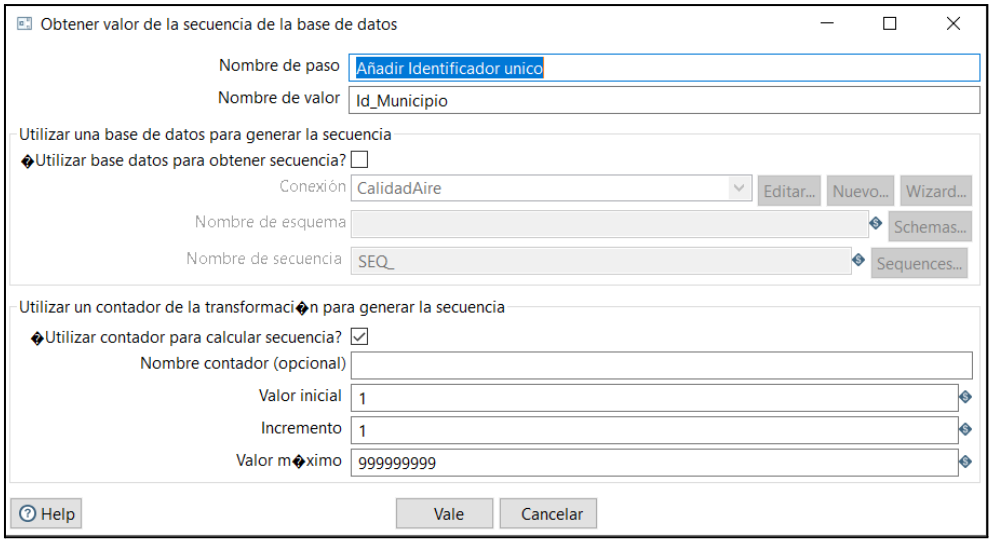
La opción “traer campos” nos ayudará a que automáticamente sean leídas las cabeceras y se les asigne el tamaño del campo en función al máximo de los tamaños existentes como datos del campo



6.1.2. Configuración de la actividad csv file

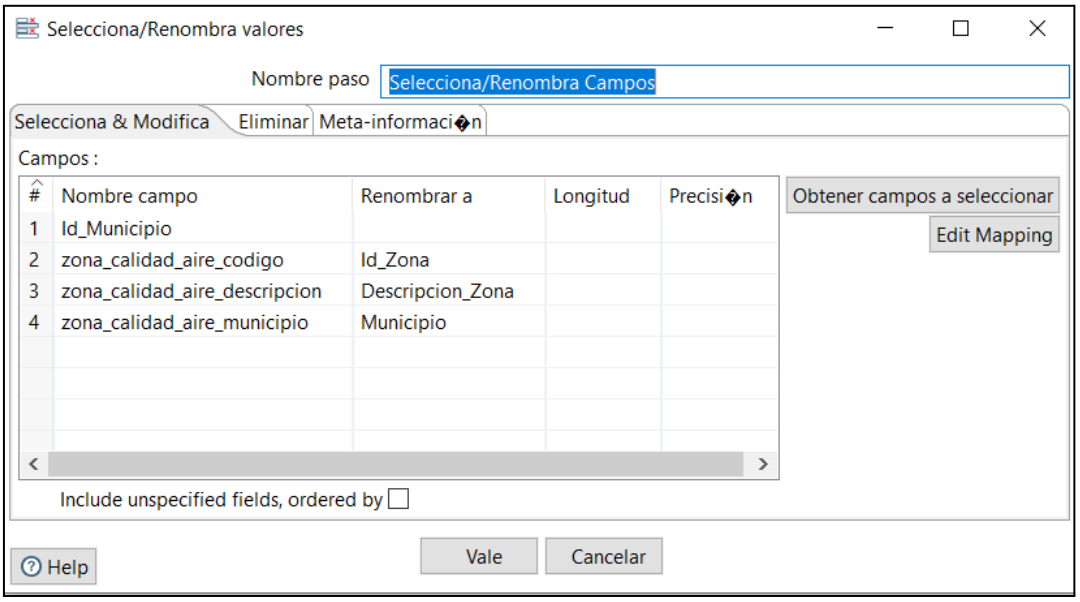
Posteriormente, crearemos un identificador para cada fila del archivo, pues esto nos ayudará a realizar futuros mapeos para obtener los valores de cada identificador.

El nombre del identificador será “*Id_Municipio”,* detallaremos en la configuración que se trata de un valor único, numérico y de tipo auto incremental.



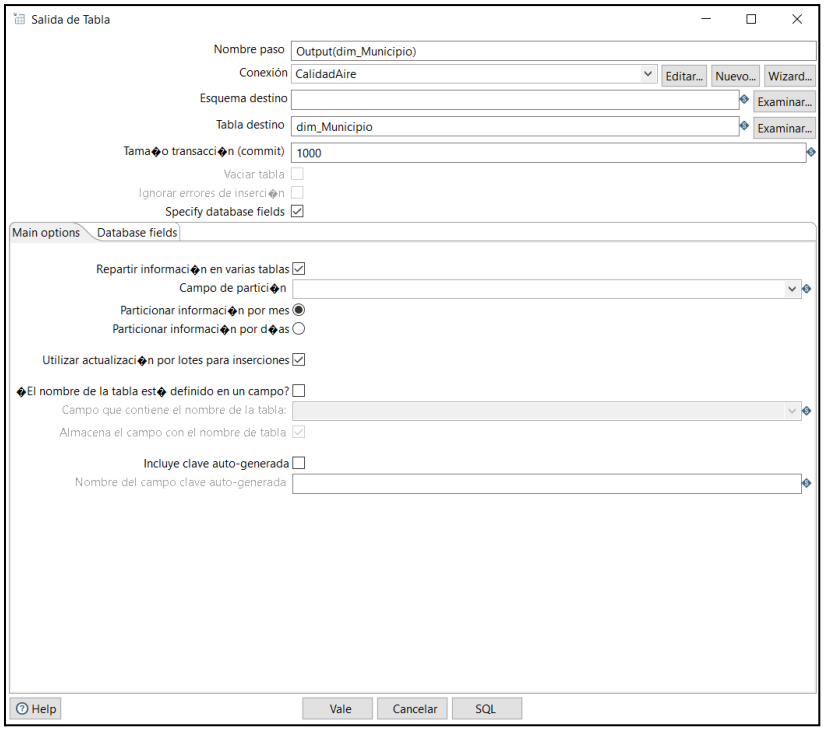
6.1.3. Configuración del identificador de la dimensión

Una vez que tenemos el identificador creado el siguiente paso es seleccionar los campos que son relevantes para el Data Warehouse y que van a aportar información al análisis posterior, en este paso también se nos ofrece la posibilidad de cambiar los campos de nombre a algo más representativo.



6.1.4. Seleccionar y renombrar datos

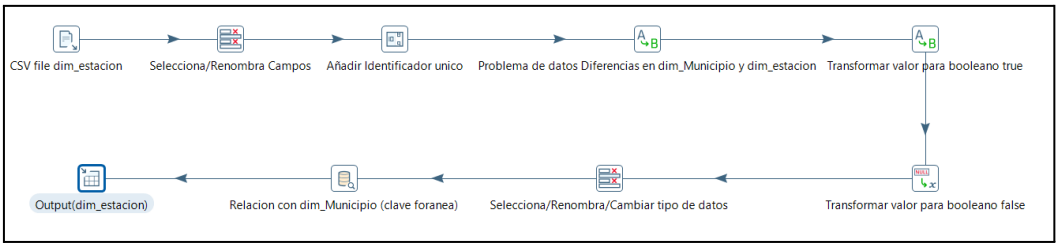
Por último, usaremos la actividad “Output” que nos ayudará a insertar la información previamente seleccionada a nuestra base de datos.



6.1.5. Configuración de la actividad output

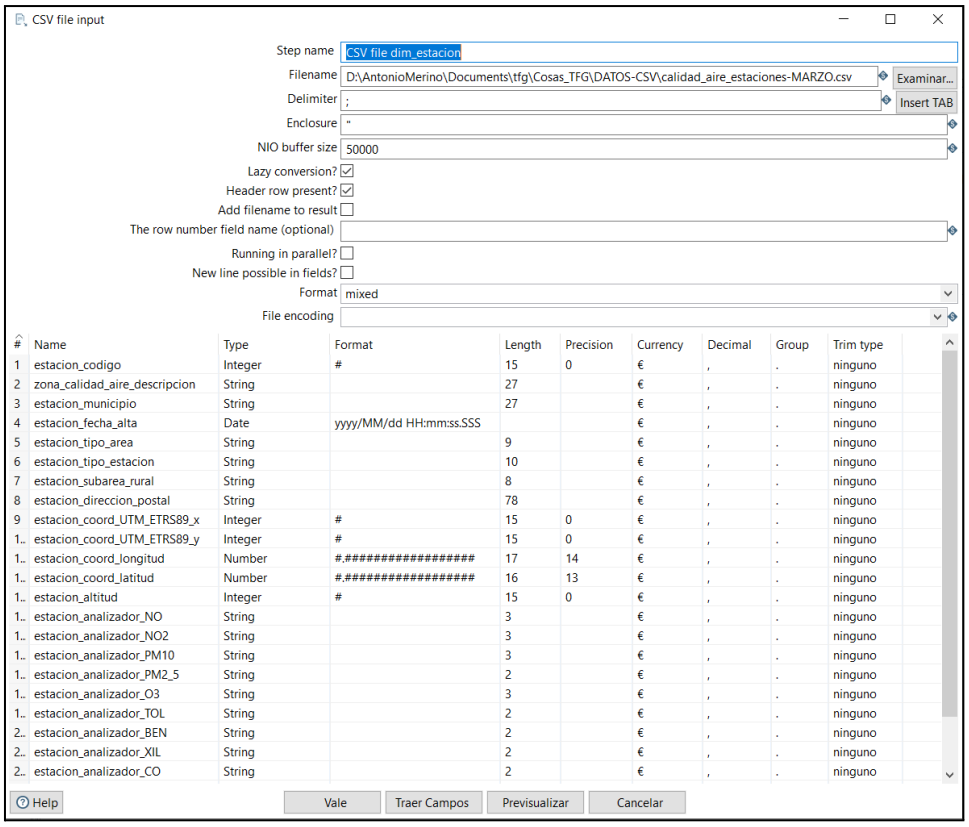
6.2. Dimensión Estación

La dimensión Estación será cargada con los datos del fichero csv *calidad_aire_estaciones* y para ello usaremos la actividad CSV file, en donde añadiremos la ruta del archivo a leer y las cabeceras del fichero.



6.2.1 Dimensión Estación

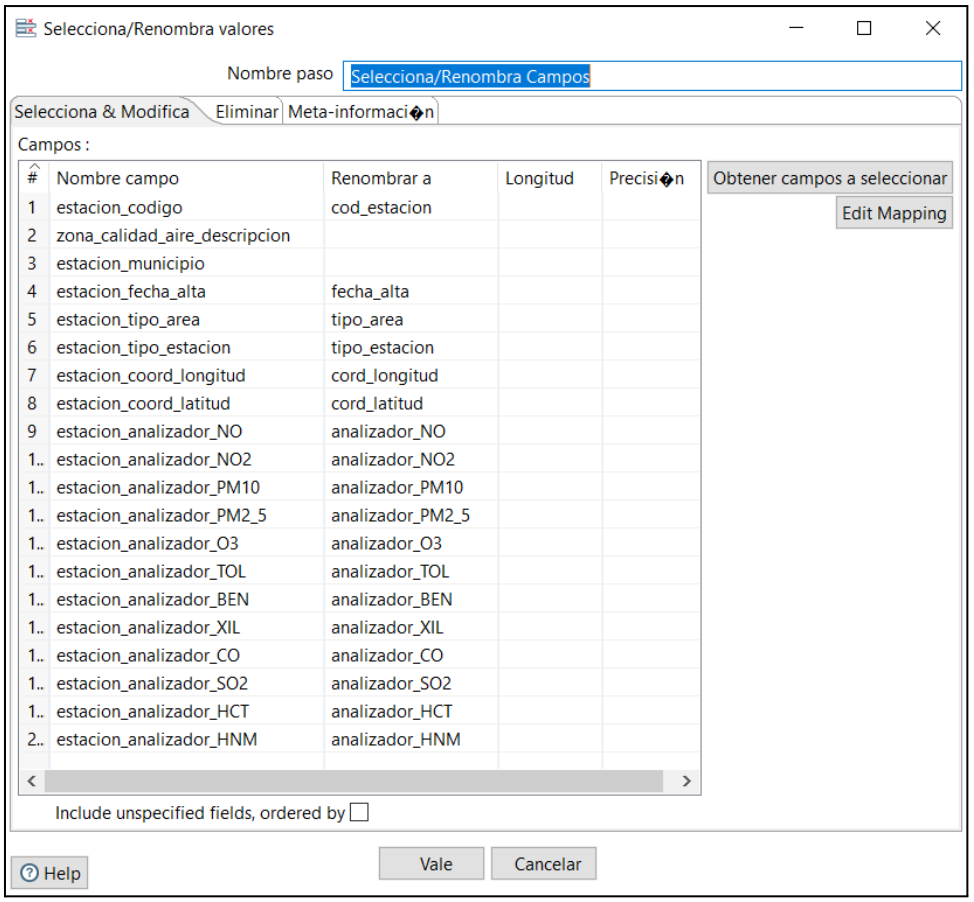
La opción “traer campos” nos ayudará a que automáticamente sean leídas las cabeceras y se les asigne el tamaño del campo en función al máximo de los tamaños existentes como datos del campo.



#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	estacion_codigo	Integer	#	15	0	€	,	.	ninguno
2	zona_calidad_aire_descripcion	String		27		€	,	.	ninguno
3	estacion_municipio	String		27		€	,	.	ninguno
4	estacion_fecha_alta	Date	yyyy/MM/dd HH:mm:ss.SSS			€	,	.	ninguno
5	estacion_tipo_area	String		9		€	,	.	ninguno
6	estacion_tipo_estacion	String		10		€	,	.	ninguno
7	estacion_subarea_rural	String		8		€	,	.	ninguno
8	estacion_direccion_postal	String		78		€	,	.	ninguno
9	estacion_coord_UTM_ETRS89_x	Integer	#	15	0	€	,	.	ninguno
1..	estacion_coord_UTM_ETRS89_y	Integer	#	15	0	€	,	.	ninguno
1..	estacion_coord_longitud	Number	#.##################	17	14	€	,	.	ninguno
1..	estacion_coord_latitud	Number	#.##################	16	13	€	,	.	ninguno
1..	estacion_altitud	Integer	#	15	0	€	,	.	ninguno
1..	estacion_analizador_NO	String		3		€	,	.	ninguno
1..	estacion_analizador_NO2	String		3		€	,	.	ninguno
1..	estacion_analizador_PM10	String		3		€	,	.	ninguno
1..	estacion_analizador_PM2_5	String		2		€	,	.	ninguno
1..	estacion_analizador_O3	String		3		€	,	.	ninguno
1..	estacion_analizador_TOL	String		2		€	,	.	ninguno
2..	estacion_analizador_BEN	String		2		€	,	.	ninguno
2..	estacion_analizador_XIL	String		2		€	,	.	ninguno
2..	estacion_analizador_CO	String		2		€	,	.	ninguno

6.2.2. Configuración de la actividad csv file

Una vez que tenemos el identificador creado el siguiente paso es seleccionar los campos que son relevantes para el Data Warehouse y que van a aportar información al análisis posterior, en este paso también se nos ofrece la posibilidad de cambiar los campos de nombre para que sean más representativos.

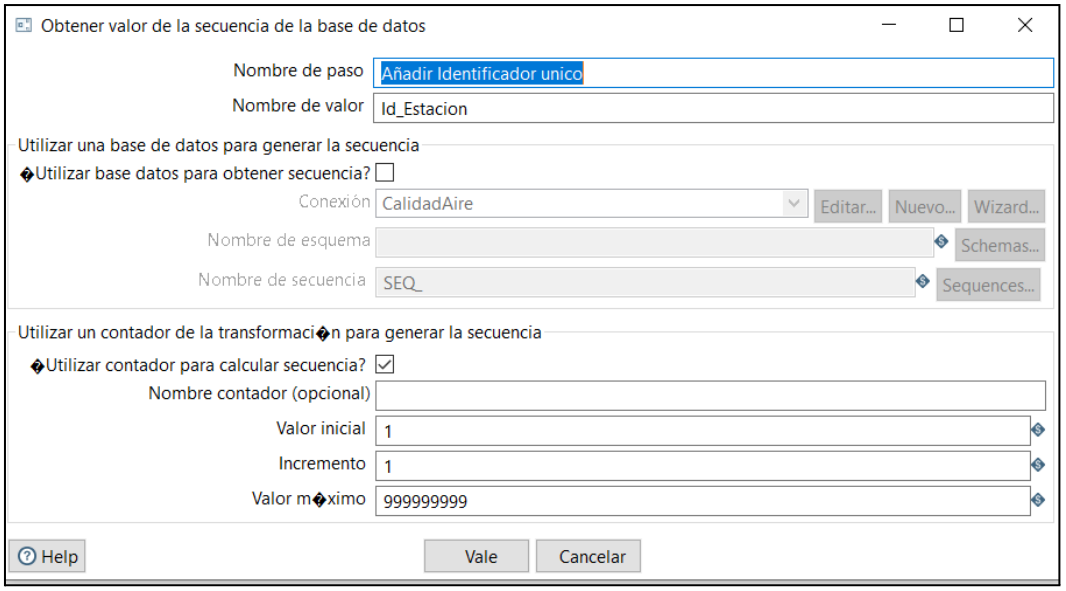


#	Nombre campo	Renombrar a	Longitud	Precisi►n
1	estacion_codigo	cod_estacion		
2	zona_calidad_aire_descripcion			
3	estacion_municipio			
4	estacion_fecha_alta	fecha_alta		
5	estacion_tipo_area	tipo_area		
6	estacion_tipo_estacion	tipo_estacion		
7	estacion_coord_longitud	cord_longitud		
8	estacion_coord_latitud	cord_latitud		
9	estacion_analizador_NO	analizador_NO		
1..	estacion_analizador_NO2	analizador_NO2		
1..	estacion_analizador_PM10	analizador_PM10		
1..	estacion_analizador_PM2_5	analizador_PM2_5		
1..	estacion_analizador_O3	analizador_O3		
1..	estacion_analizador_TOL	analizador_TOL		
1..	estacion_analizador_BEN	analizador_BEN		
1..	estacion_analizador_XIL	analizador_XIL		
1..	estacion_analizador_CO	analizador_CO		
1..	estacion_analizador_SO2	analizador_SO2		
1..	estacion_analizador_HCT	analizador_HCT		
2..	estacion_analizador_HNM	analizador_HNM		

6.2.3. Seleccionar y renombrar datos

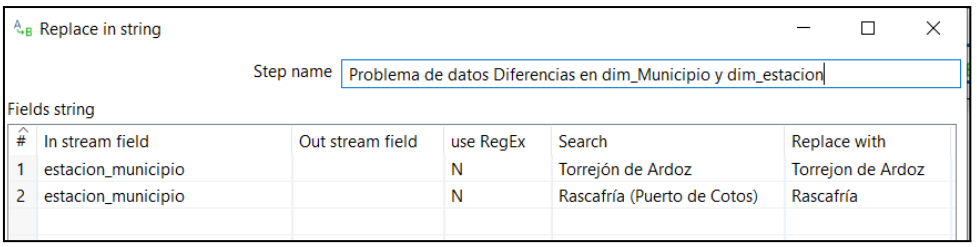
Crearemos un identificador para cada fila del archivo, pues esto nos ayudará a realizar futuros mapeos para obtener los valores de cada identificador.

El nombre del identificador será *Id_Estación,* detallaremos en la configuración que se trata de un valor único, numérico y de tipo auto incremental.



6.2.5. Configuración del identificador de la dimensión

La dimensión estación está relacionada con la dimensión anterior municipio, es por ello que ambas dimensiones han de llevar los mismos valores de relación. Para solventar el problema que nos hemos encontrado con la nomenclatura de los dos ficheros csv, hemos utilizado la actividad “Replace in string” para normalizar los datos de ambos ficheros.



6.2.6. Configuración para normalizar los datos de entrada

En el fichero csv correspondiente a esta dimensión encontramos columnas cuyos datos son asociados a la disponibilidad de analizadores por estación, los analizadores se encargan de detallar si una partícula es contaminante o no.
Por ello, hemos reemplazado los valores procedentes del csv por booleanos, si el dato del fichero contiene un “SI” será sustituido por el booleano “true” y si contiene “null” por “false”.

La búsqueda de información usando booleanos es mucho más efectiva que su respectiva búsqueda por datos de tipo string.

Replace in string

Step name: Transformar valor para booleano true

Fields string

#	In stream field	Out stream field	use RegEx	Search	Replace with	Set empty string?	Replace with field	Whole Word	Case se
1	analizador_BEN		N	Si	1	N		N	N
2	analizador_CO		N	Si	1	N		N	N
3	analizador_HCT		N	Si	1	N		N	N
4	analizador_HNM		N	Si	1	N		N	N
5	analizador_NO		N	Si	1	N		N	N
6	analizador_NO2		N	Si	1	N		N	N
7	analizador_O3		N	Si	1	N		N	N
8	analizador_PM10		N	Si	1	N		N	N
9	analizador_PM2_5		N	Si	1	N		N	N
1..	analizador_SO2		N	Si	1	N		N	N
1..	analizador_TOL		N	Si	1	N		N	N
1..	analizador_XIL		N	Si	1	N		N	N

Help | Vale | Get fields | Cancelar

6.2.7. Configuración para normalizar los datos de entrada - I

If field value is null

Step name: Transformar valor para booleano false

Replace Null for all fields

Replace by value

Set empty string?

Mask (Date)

Select fields ☑

Select value type ☐

Value types

#	Type	Replace by value	Conversion mask (Date)	Set empty string?

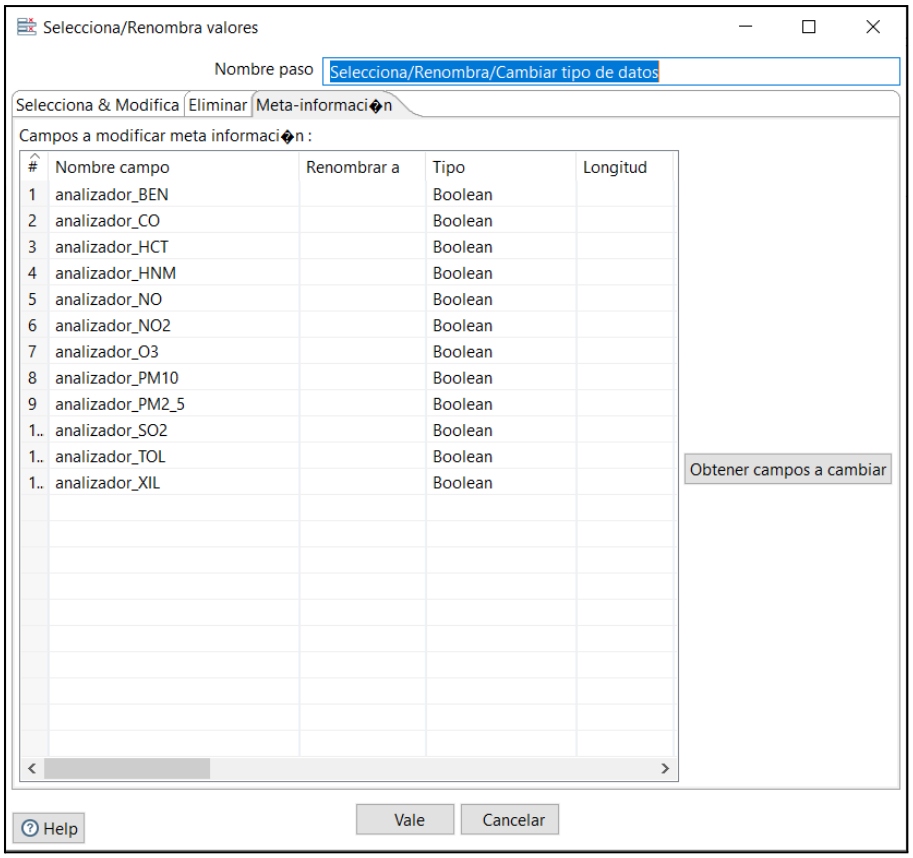
Fields

#	Field	Replace by value	Conversion mask (Date)	Set empty string?
1	analizador_PM10	0		N
2	analizador_BEN	0		N
3	analizador_CO	0		N
4	analizador_HCT	0		N
5	analizador_HNM	0		N
6	analizador_NO	0		N
7	analizador_NO2	0		N
8	analizador_O3	0		N
9	analizador_PM2_5	0		N
1..	analizador_SO2	0		N
1..	analizador_TOL	0		N
1..	analizador_XIL	0		N

Help | Vale | Traer Campos | Cancelar

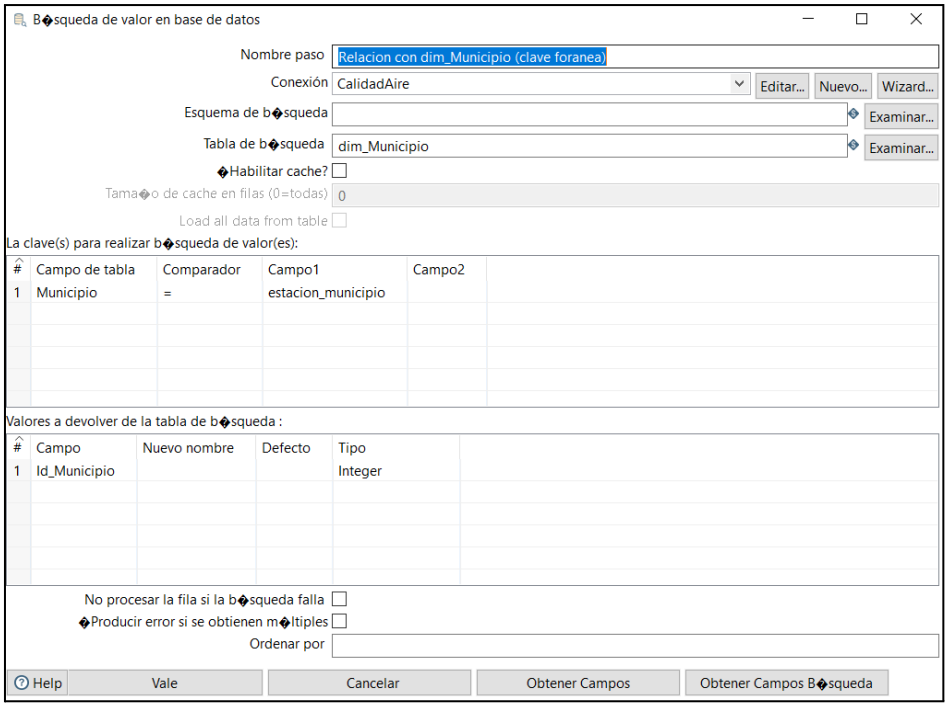
6.2.8. Configuración para normalizar los datos de entrada - II

Una vez que hemos realizado esta transformación ya estamos listos para transformar todos los campos de analizadores a formato booleano.



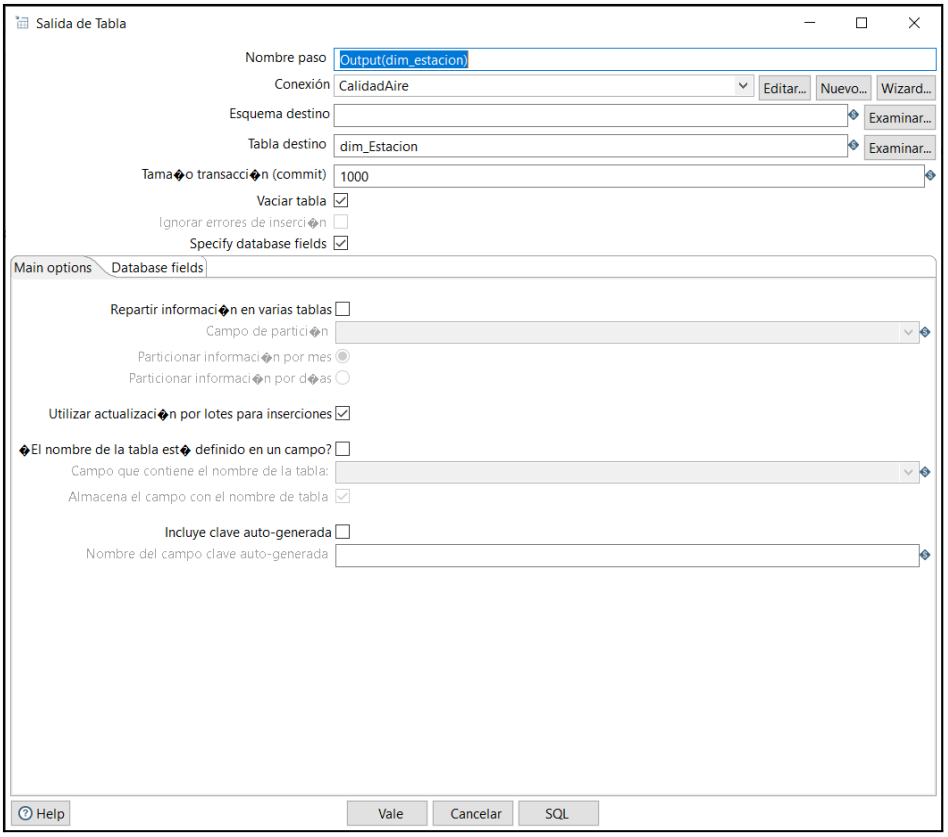
6.2.9. Seleccionar y renombrar datos

Una vez que ya contamos con los campos, nombres adecuados y sus formatos de manera correcta solo nos falta crear la relación con la dimensión municipio como muestra el modelo. Para ello usaremos la actividad “lookup” la cual nos va a permitir mediante el campo municipio en ambas dimensiones obtener el identificador único de la dimensión municipio a esta dimensión para así crear la relación entre los identificadores.



6.2.10. Configuración de la relación entre los identificadores

Por último, usaremos la actividad "Output" que nos ayudará a insertar la información previamente seleccionada a nuestra base de datos.



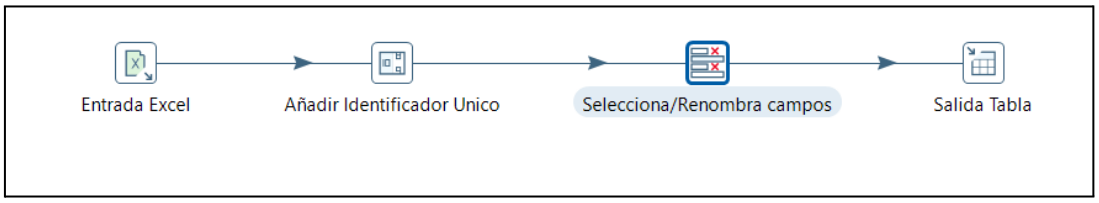
6.2.11. Configuración de la actividad output

6.3. Dimensión Magnitudes

La creación de esta dimensión ha sido la más costosa, puesto a que la Comunidad de Madrid no nos ha proporcionado detalles acerca de las magnitudes, únicamente han sido expuestas por medio de una imagen. Hemos trasladado la información de la imagen a un archivo excel de manera manual, para así finalmente incluirlos en esta nueva dimensión.

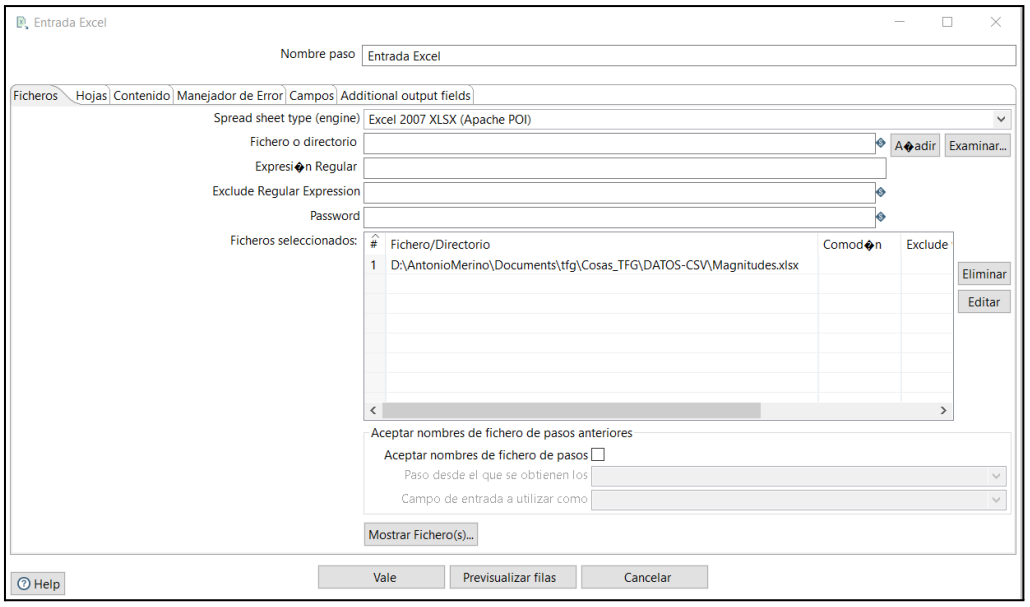
Id_magnitud_nk	descripcion_magnitud	abrebiatura_magnitud	unidad_medida_magnitud	descripcion_tecnica	tipo_magnitud
1	Dióxido de azufre	SO2	µg/m3	Fluorescencia ultravioleta	calidad_aire
6	Monóxido de carbono	CO	mg/m3	Absorción infrarroja	calidad_aire
7	Monóxido de nitrógeno	NO	µg/m3	Quimioluminiscencia	calidad_aire
8	Dióxido de nitrógeno	NO2	µg/m3	Id.	calidad_aire
9	Particulas < 2.5 um	PM2.5	µg/m3	Microbalanza	calidad_aire
10	Particulas < 10 um	PM10	µg/m3	Id.	calidad_aire
12	Óxidos de nitrogeno	Nox	µg/m3	Quimioluminiscencia	calidad_aire
14	Ozono	O3	µg/m3	Absorcion ultravioleta	calidad_aire
20	Tolueno	TOL	µg/m3	Cromatografía de gases	calidad_aire
30	Benzeno	BEN	µg/m3	Id.	calidad_aire
35	Etilbenceno	EBE	µg/m3	Id.	calidad_aire
37	Metaxileno	MXY	µg/m3	Id.	calidad_aire
38	Parasileno	PXY	µg/m3	Id.	calidad_aire
39	Ortoxileno	OXY	µg/m3	Id.	calidad_aire
42	Hidrocarburos totales (hexano)	TCH	mg/m3	Ionización de llama	calidad_aire
43	Metano	CH4	mg/m3	Id.	calidad_aire
44	Hidrocarburos no metanicos (hexano)	NMHC	mg/m3	Id.	calidad_aire
80	Radiacion ultravioleta		Mw/m2		meteorologia
81	Velocidad viento		m/s		meteorologia
82	Direccion viento		-		meteorologia
83	Temperatura		ºC		meteorologia
86	Humedad relativa		%		meteorologia
87	Presión barométrica		mb		meteorologia
88	Radiacion solar		Kw/m2		meteorologia
89	Precipitación		l/m2		meteorologia
-1	No consta				No consta

Este archivo Excel se compone por todos los contaminantes que influyen en la calidad del aire y componentes de medición meteorológicos,como campos principales de este fichero encontramos el nombre, abreviatura y en qué unidad de medida se expresan.



6.3.1 Dimensión Magnitud

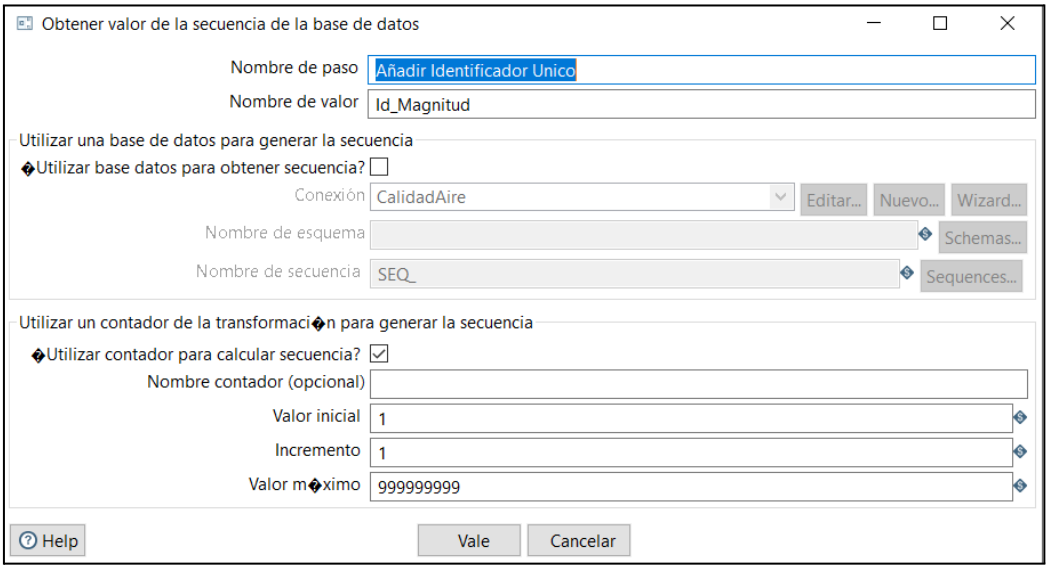
Hemos usado la actividad “entrada excel” que por medio de la ruta de nuestro archivo en local obtiene la información del mismo.



6.3.2. Configuración de la actividad entrada excel

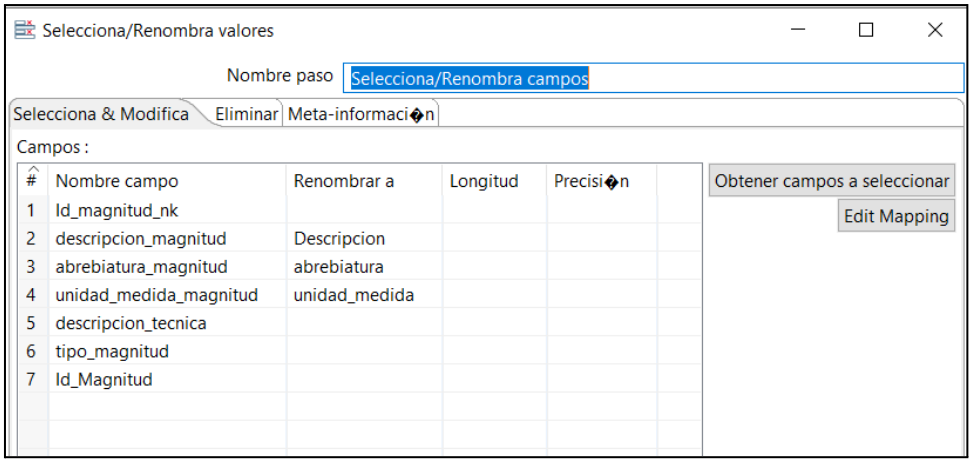
Crearemos un identificador para cada fila del archivo, pues esto nos ayudará a realizar futuros mapeos para obtener los valores de cada identificador.

El nombre del identificador será *Id_Magnitud,* detallaremos en la configuración que se trata de un valor único, numérico y de tipo auto incremental.



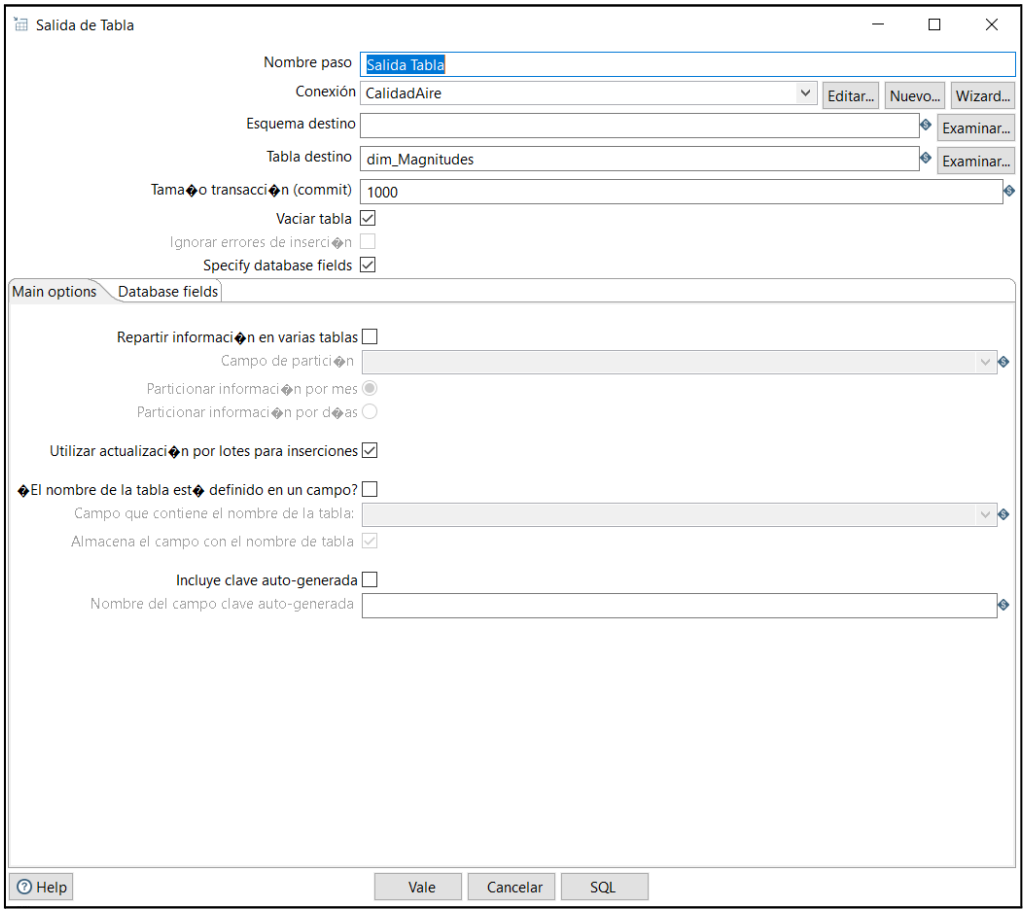
6.3.3. Configuración del identificador de la dimensión

Debido a que el archivo excel lo hemos creado de manera manual nos hemos asegurado de que la calidad de los datos sea correcta por lo tanto solo añadimos un paso más en el que renombramos los campos utilizando un nombre mucho más descriptivo.



6.3.4. Seleccionar y renombrar datos

Por último, usaremos la actividad "Output" que nos ayudará a insertar la información previamente seleccionada a nuestra base de datos.



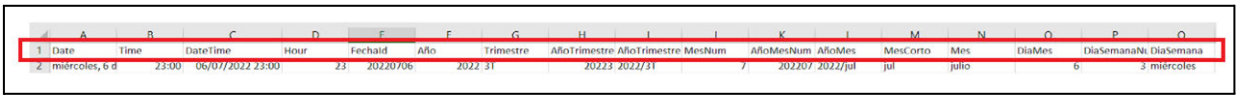
6.3.5. Configuración de la actividad output

6.4. Dimensión Calendario

Esta dimensión tiene la misma casuística que la anterior, ya que, estos datos no vienen por parte de la Comunidad de Madrid si no que los hemos generado nosotros de manera particular para así dar sentido al Data Warehouse.

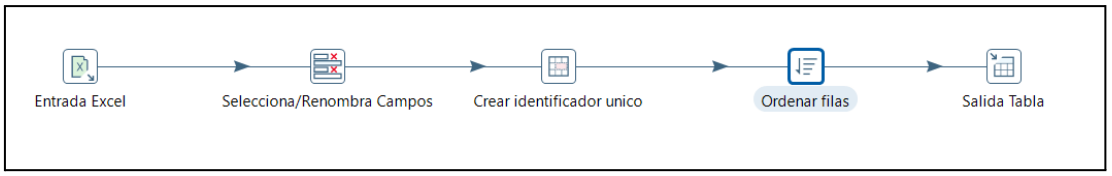
Por lo general las dimensiones de tiempo suelen tener la granularidad de los datos hasta el campo diario, pero en este caso necesitamos tener todas las horas durante el día para poder realizar un análisis más exhaustivo de los datos.

Debido a esto hemos decidido reutilizar los datos de un proyecto propio anterior en el que teníamos generada en la herramienta de Power BI una dimensión de tiempo con las características buscadas, se ha procedido a exportar los datos a un fichero excel para su posterior tratamiento en la ETL.



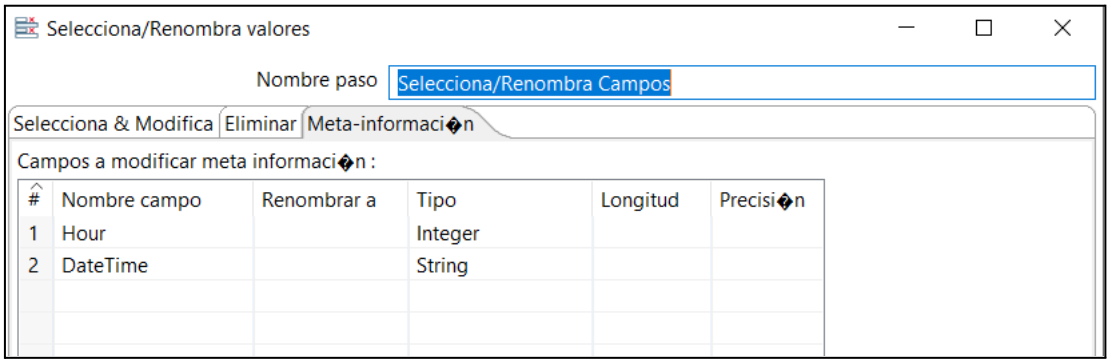
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Date	Time	DateTime	Hour	FechaId	Año	Trimestre	AñoTrimestre	AñoTrimestreNu	MesNum	AñoMesNum	AñoMes	MesCorto	Mes	DiaMes	DiaSemanaNu	DiaSemana
2	miércoles, 6 d	23:00	06/07/2022 23:00	23	20220706	2022	3	2022	3	2022/3	7	202207	2022/jul	jul	julio	6	3 miércoles

6.4.1. Cabecera del fichero excel creado para la dimensión



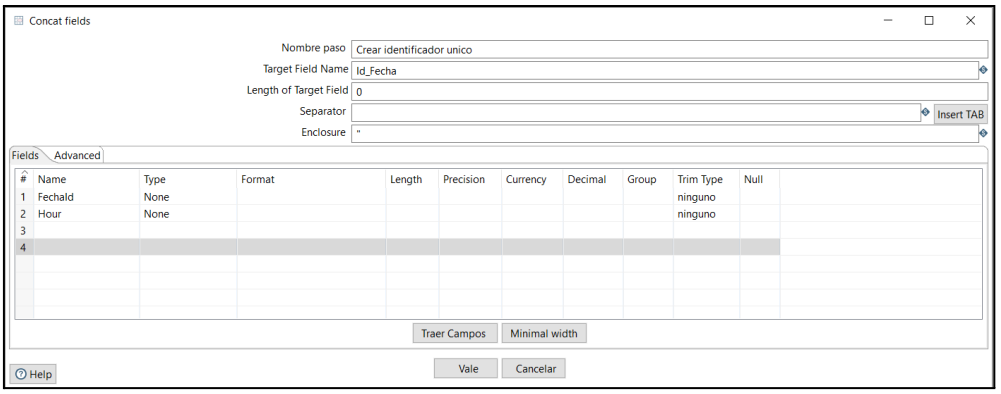
6.4.2. Dimensión Calendario

Trás la importación de datos, usaremos la actividad de renombrado de valores, pues es necesario cambiar el tipo de valor a los campos hora y fecha, gracias a formatear estos dos valores nuestro proyecto contará con un filtrado agrupado por fechas y horas.



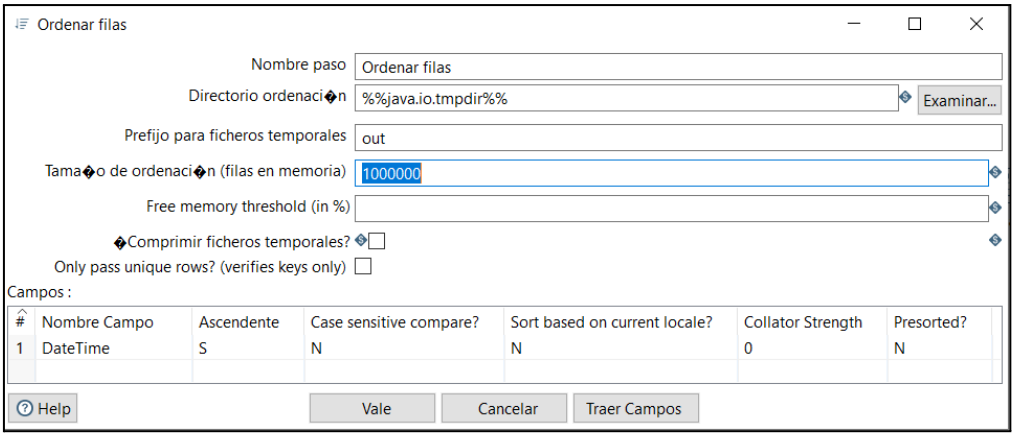
6.4.3. Formateo de la fecha y la hora

Una vez cambiados los formatos, tenemos que crear, como en las anteriores, dimensiones un identificador único, pero en este caso no vamos a optar por un valor entero incremental, sino, optamos por un identificador numérico y entero que represente el valor de la fecha más la hora, por lo que vamos a utilizar un paso para concatenar cadenas de caracteres, el campo *Fechald* junto al campo *Hour* para así obtener este identificador que posteriormente usaremos para relacionar con las tablas de hechos y poder tener un correcto seguimiento en el tiempo de los hechos.



6.4.5. Creación del identificador de la dimensión

Posteriormente, utilizamos la actividad de ordenación para reorganizar la información de esta dimensión.

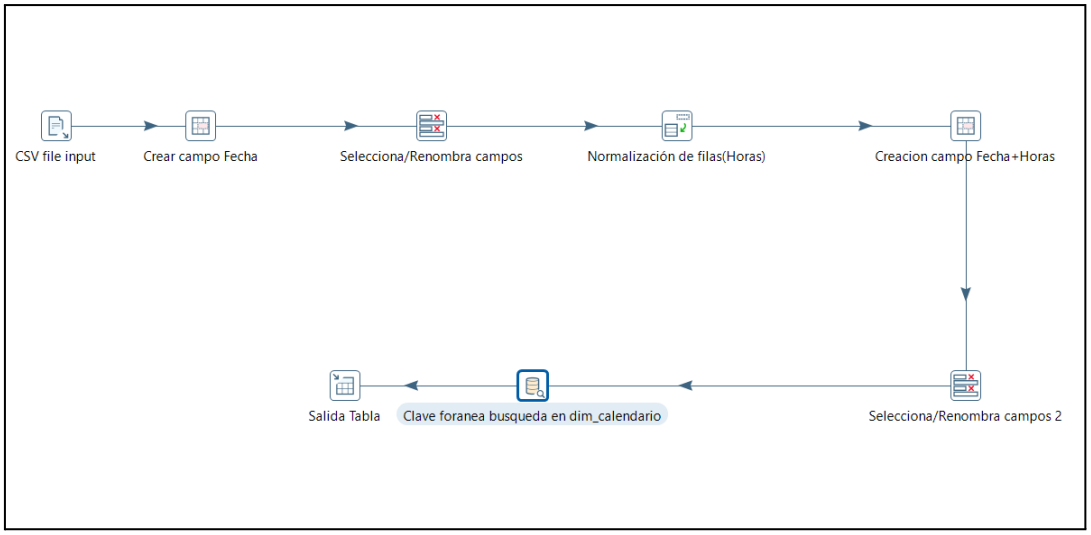


6.4.6. Reorganización de la dimensión

Por último, usaremos la actividad “Output” que nos ayudará a insertar la información previamente seleccionada a nuestra base de datos.

6.5. Tabla de hechos Calidad del aire

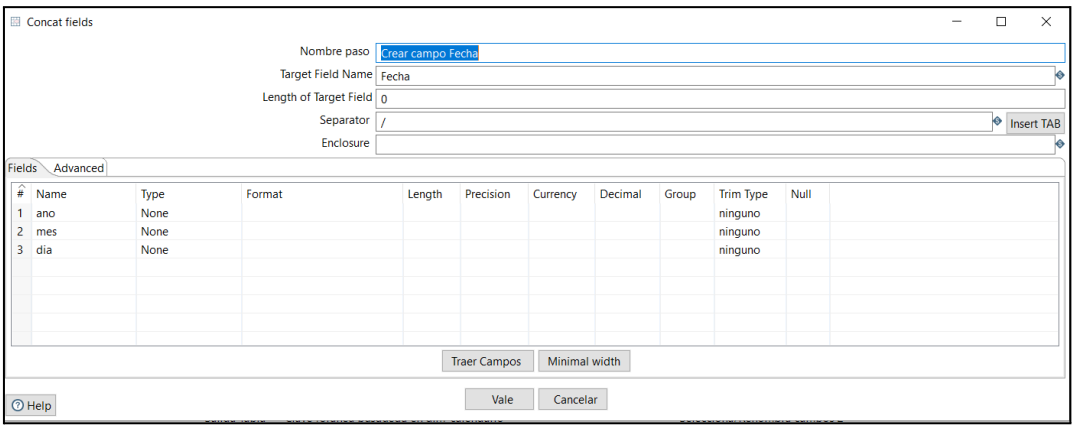
Las tablas de hechos son el núcleo de los procesos ETL, puesto que es en estas tablas donde sucede el auténtico refinamiento de los datos obtenidos en el origen, para así, poder ser analizados a posteriori.



6.5.1. Tabla de hechos de la calidad del aire

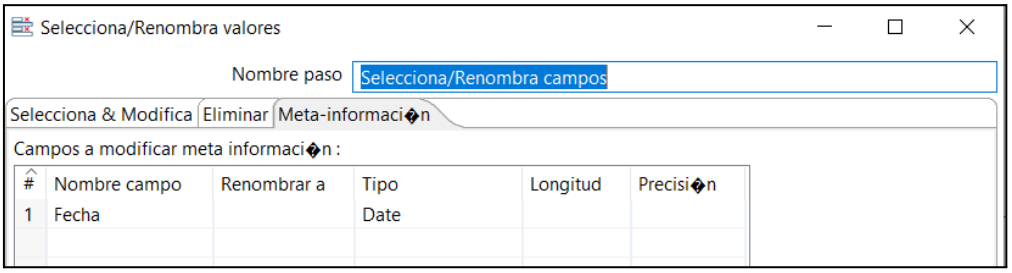
Todos los datos que hemos utilizado son de tipo csv, lo cual nos facilita en gran medida el proceso de transformación.

En primer lugar, observamos que la fecha viene dividida en 3 campos: día, mes, año. Por lo tanto hemos utilizado la actividad “Concat fields” para unir estos tres parámetros en uno solo.



6.5.2. Configuración de la actividad de concatenación

Posteriormente, usaremos la actividad de renombrado para dar nombre a este nuevo atributo.



6.5.3. Seleccionar y renombrar datos

En el siguiente paso, nos encontramos el principal problema de estos datos, las horas del dia vienen desnormalizadas en formato columnas para un solo día, lo cual para trabajar con este tipo de datos mediante tablas de hechos resulta imposible, además, estas columnas vienen con una nomenclatura del estilo: “h01”, “h02”, ”h03”,etc.. Por lo que es necesario realizar un proceso de normalización de esas columnas a filas. En el primer paso cambiaremos la nomenclatura mencionada por su homólogo en formato fecha.

También aprovecharemos este paso para cambiar el formato del campo fecha que hemos creado anteriormente a un formato de tipo “date”.

h01	01:00:00
h02	02:00:00
h03	03:00:00
h04	04:00:00
h05	05:00:00
h06	06:00:00
h07	07:00:00
h08	08:00:00
h09	09:00:00
h10	10:00:00
h11	11:00:00
h12	12:00:00
h13	13:00:00
h14	14:00:00
h15	15:00:00
h16	16:00:00
h17	17:00:00
h18	18:00:00
h19	19:00:00
h20	20:00:00
h21	21:00:00
h22	22:00:00
h23	23:00:00
h24	00:00:00

6.5.4. Creación y formateo del campo fecha

De esta manera pasamos a multiplicar cada fila que teníamos por 24, que es su correspondencia por horas del día.

Una vez que ya tenemos esto realizado, vamos a crear el campo de fecha que de verdad nos interesa, es decir, la concatenación de los campos “Fecha y Hora”.
Seguidamente, transformamos el formato de este campo para que se almacene como un string y renombramos las claves foráneas de las dimensiones que tenemos en los datos para que se sepa que son identificadores

#	Nombre campo	Renombrar a
1	magnitud	Id_magnitud
2	estacion	Id_estacion

6.5.5. Renombrar magnitudes

Para finalizar, relacionamos la tabla con la dimensión calendario mediante el campo “Fecha y Hora” creado anteriormente. Esto lo realizamos mediante la actividad “lockup” en la base de datos en la que nos traemos el identificador de la dimensión calendario.

La clave(s) para realizar b►squeda de valor(es):

#	Campo de tabla	Comparador	Campo1	Campo2
1	DateTime	=	fecha_horas	

Valores a devolver de la tabla de b►squeda :

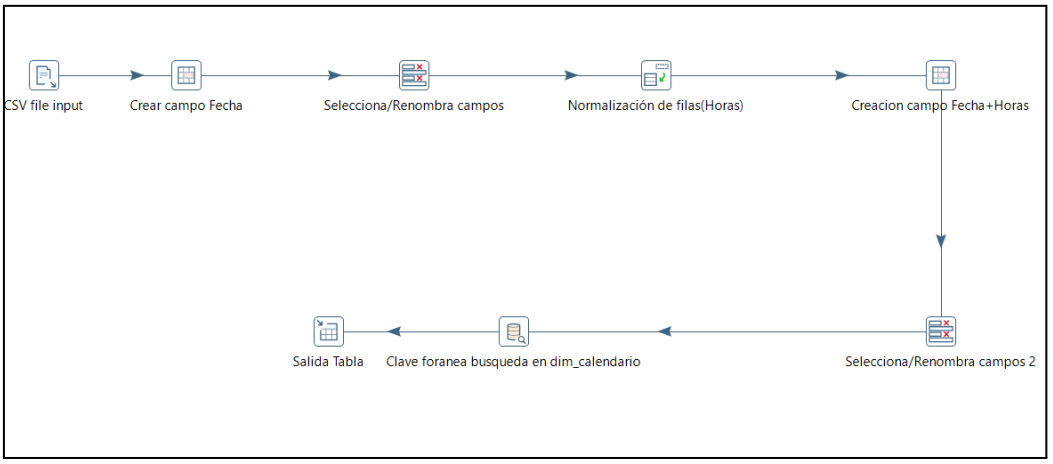
#	Campo	Nuevo nombre	Defecto	Tipo
1	Id_Fecha			None

6.5.6. Asignación del campo de búsqueda

De esta forma, ya tenemos la tabla relacionada con todas las dimensiones. Hemos transformado los datos para poder analizarlos, posteriormente, en la fase de reporting, mejorando de esta manera la calidad de los mismos y corrigiendo las ambigüedades provenientes del origen de los datos. Por lo que el proceso ya se encuentra preparado adecuadamente para insertar la información en la tabla correspondiente en la BBDD.

6.6. Tabla de hechos Metereología

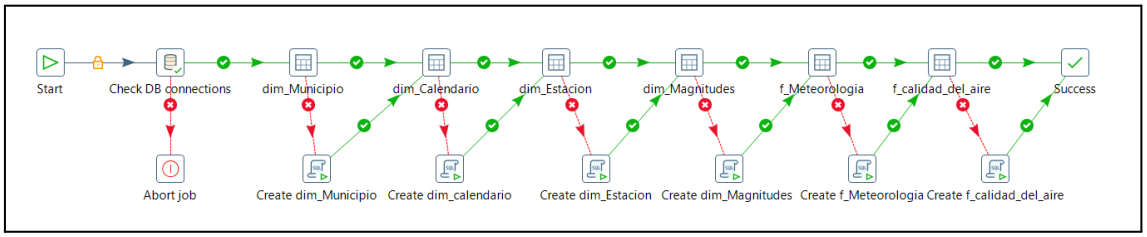
En esta segunda tabla de hechos, el proceso ETL es el mismo que la tabla de hechos explicada en el punto anterior, el unico cambio es el fichero CSV al inicio del proceso con los datos.
Se realizan los siguientes pasos como en la anterior tabla de hechos ya descrita:
Se cargará el csv con los datos recogidos de la Comunidad de Madrid, posteriormente se creará adecuadamente el parámetro fecha, para el cual se normalizarán las filas y se añadirá a la fecha las horas correspondientes, como se muestra en la siguiente imagen.



6.6.1. Tabla de hechos de metereología

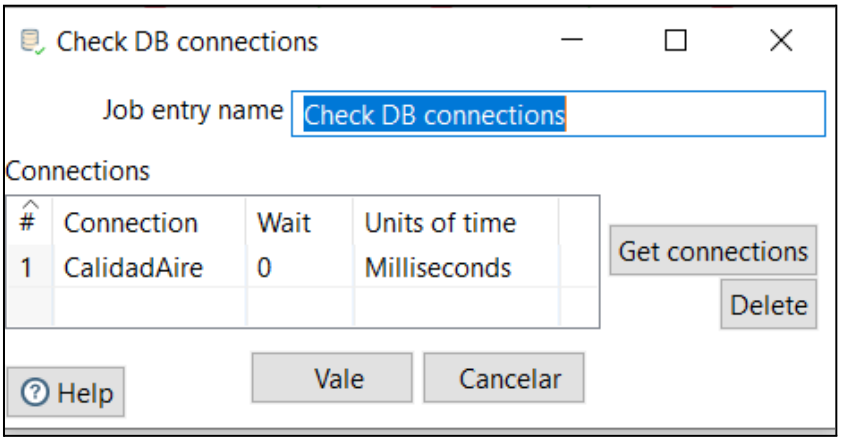
6.7. Validación del Data Warehouse

En el siguiente job se hará previamente la comprobación de los correctos datos de acceso a la base de datos, de esta manera nos aseguramos tener una conexión correcta antes de realizar toda la carga de información.



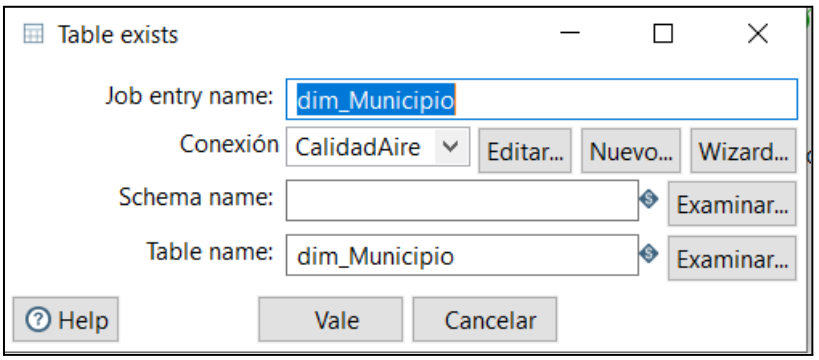
6.7.1. Job de carga y validación de datos

En la siguiente imagen se pueden observar los datos de la conexión a la base de datos.



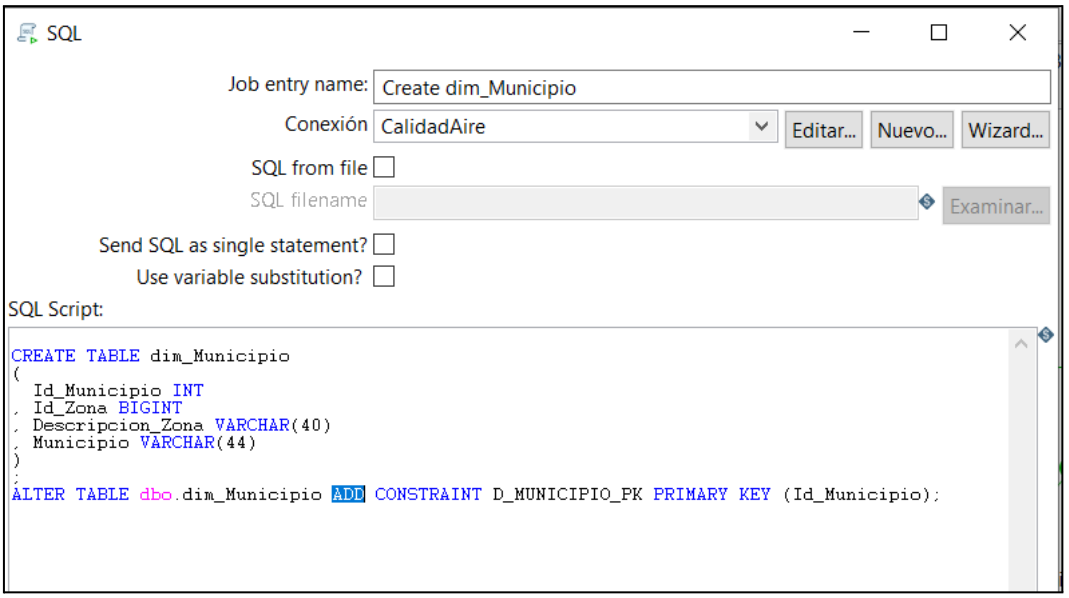
6.7.2. Conexión a la base de datos

Posteriormente, realizaremos las comprobaciones referentes a la existencia de todas las tablas de base de datos en donde van a ser cargados los datos del proceso.



6.7.3. Comprobación de la existencia de las tablas

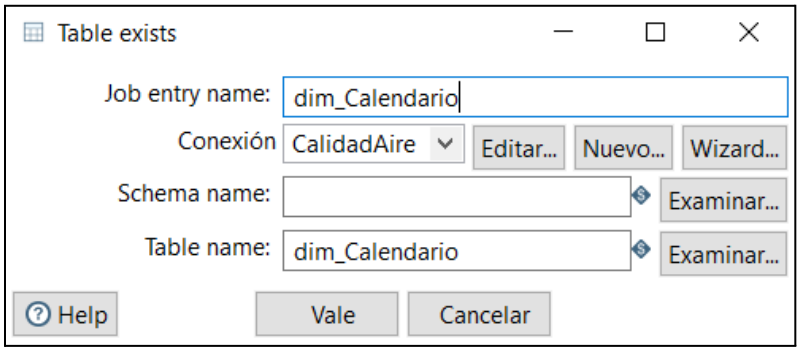
En caso de no existir alguna de las tablas en el sistema, el automatismo se encargará de crear la tabla, para así poder añadir los datos de manera correcta.



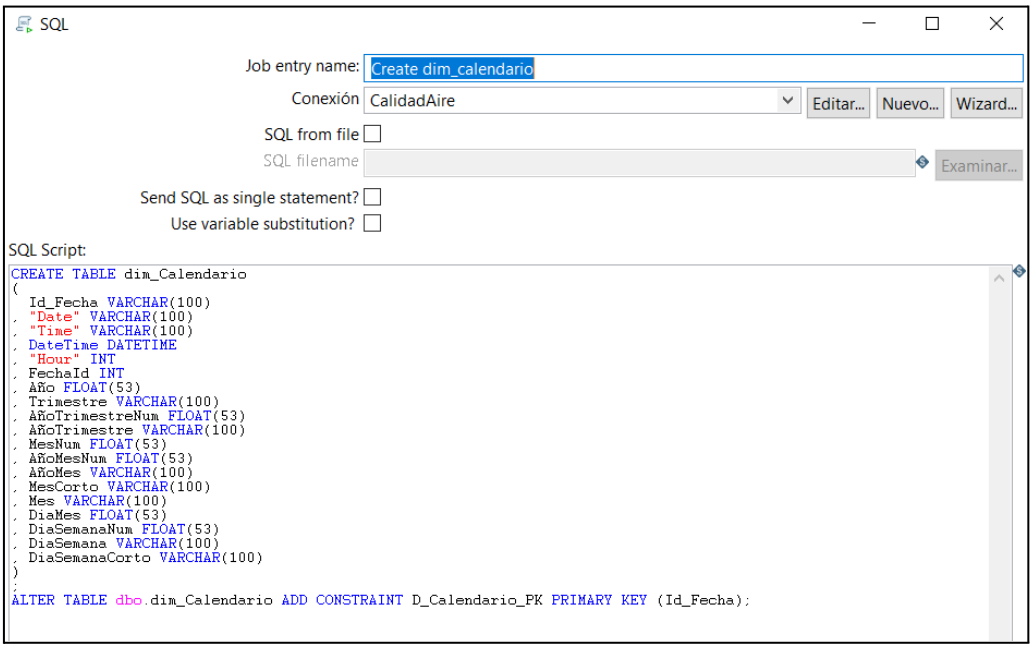
6.7.4. Script de creación de la tabla Municipio

Este proceso se repetirá para cada una de las dimensiones y tablas de hechos.

- Creación de la tabla Calendario

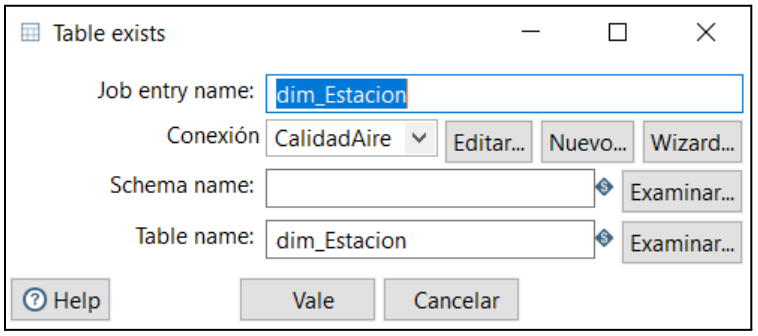


6.7.5. Comprobación de la existencia de las tablas

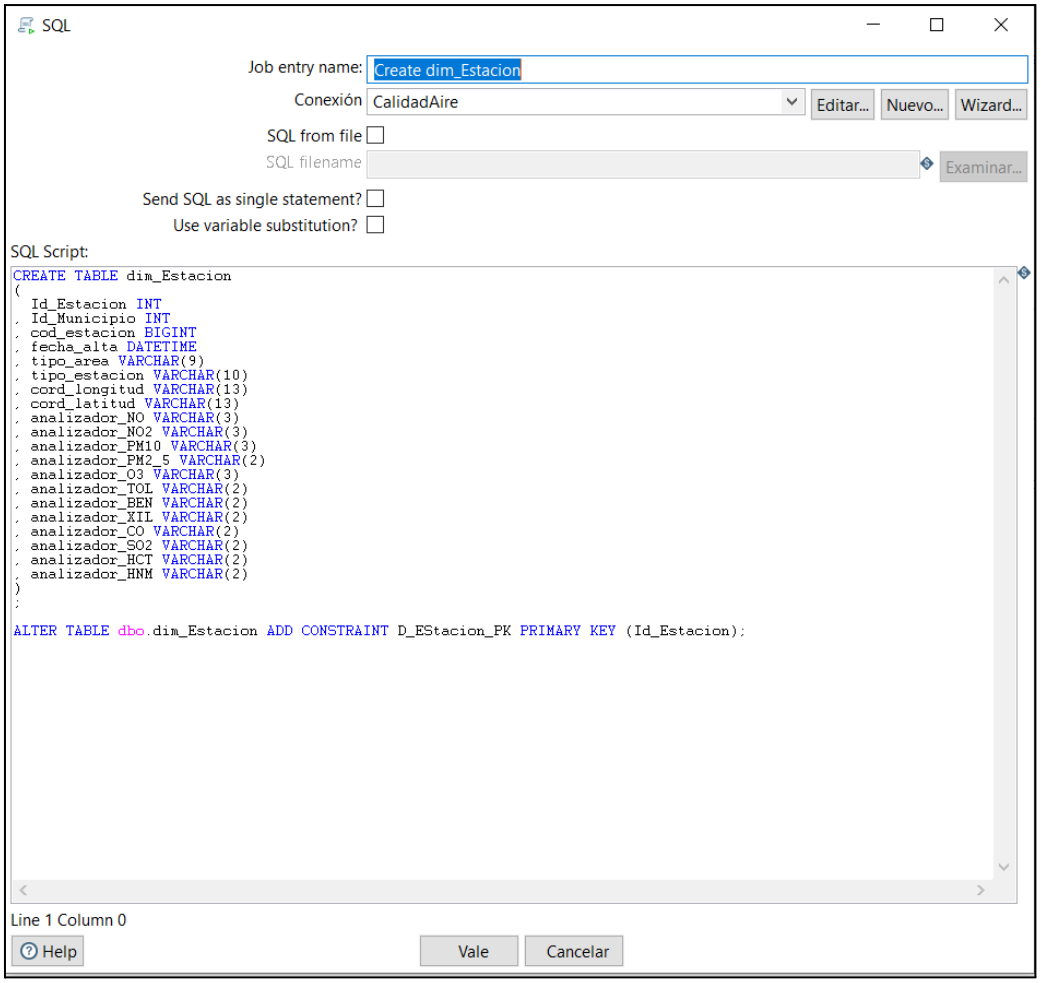


6.7.6. Script de creación de la tabla Calendario

- Creación de la tabla Estación

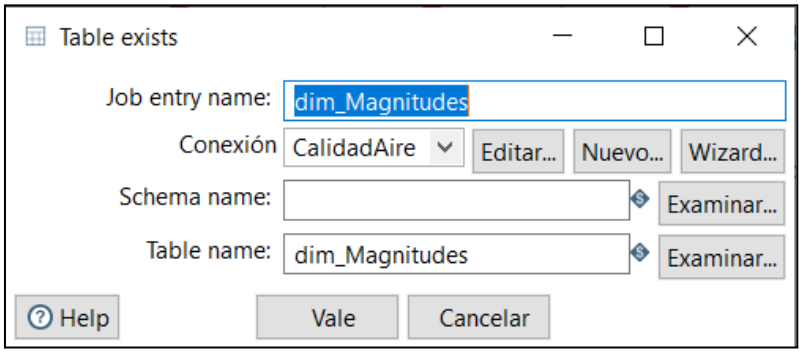


6.7.7. Comprobación de la existencia de las tablas

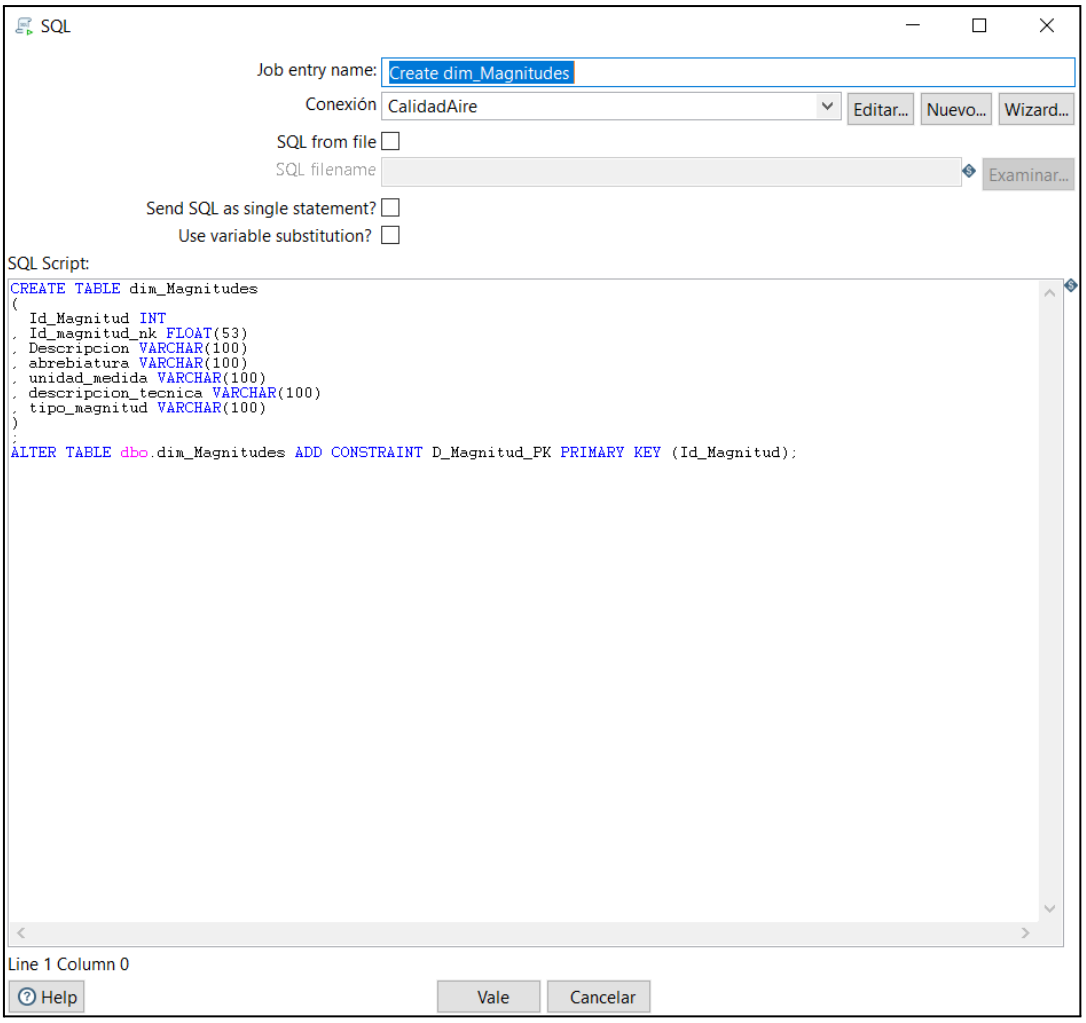


6.7.8.Script de creación de la tabla Estación

- Creación de la tabla Magnitudes

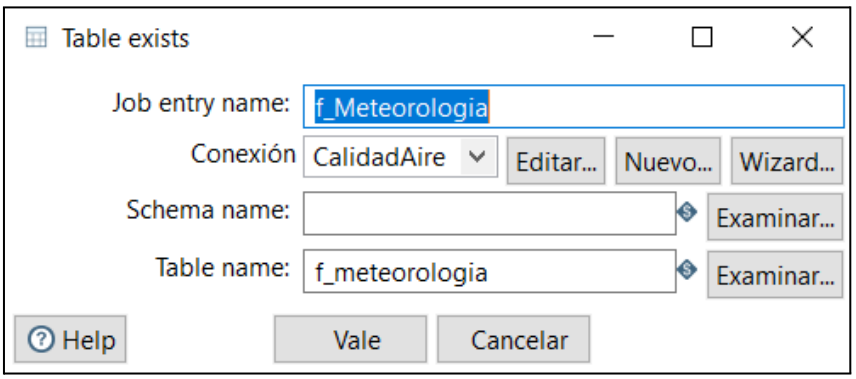


6.7.9. Comprobación de la existencia de las tablas

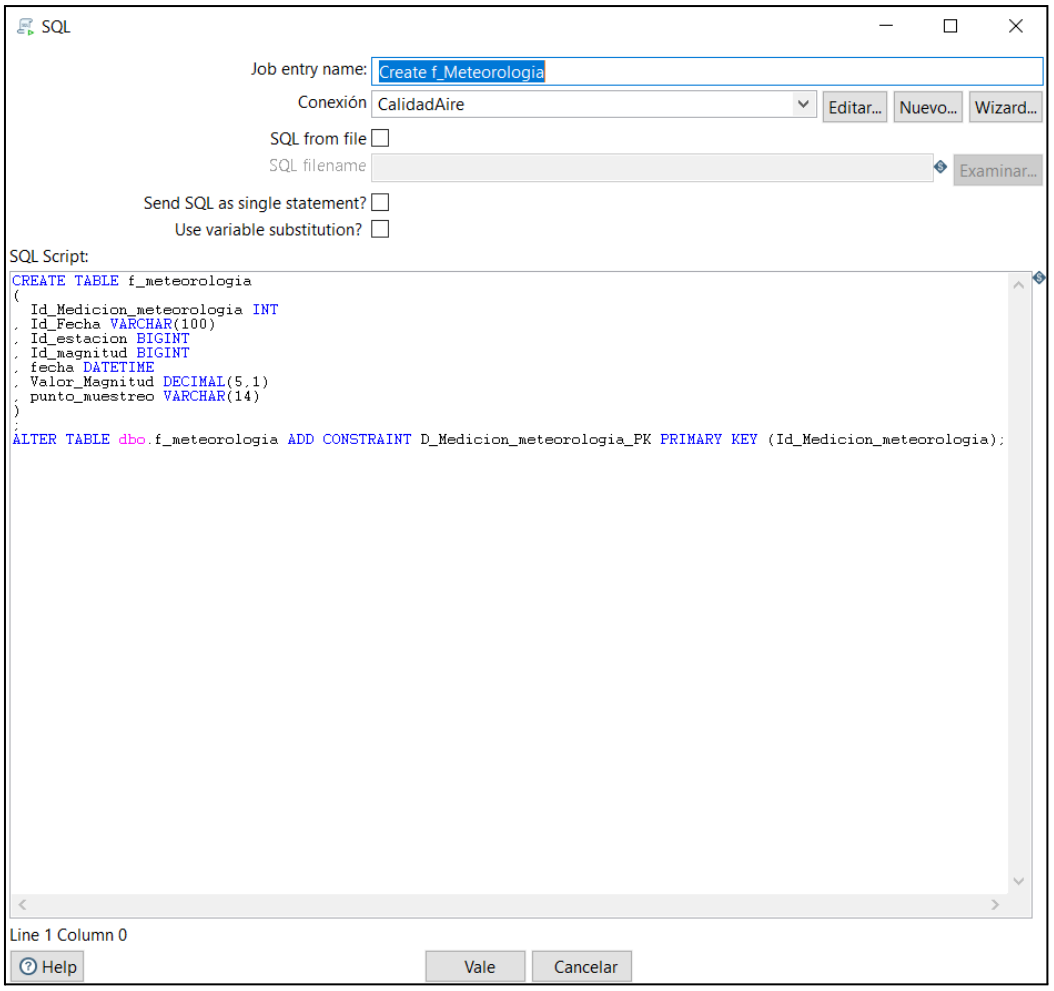


6.7.10.Script de creación de la tabla Magnitudes

- Creación de la tabla Meteorlogía

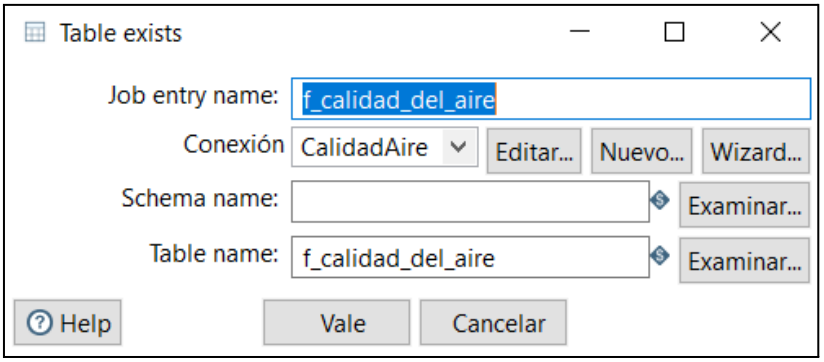


6.7.11. Comprobación de la existencia de las tablas

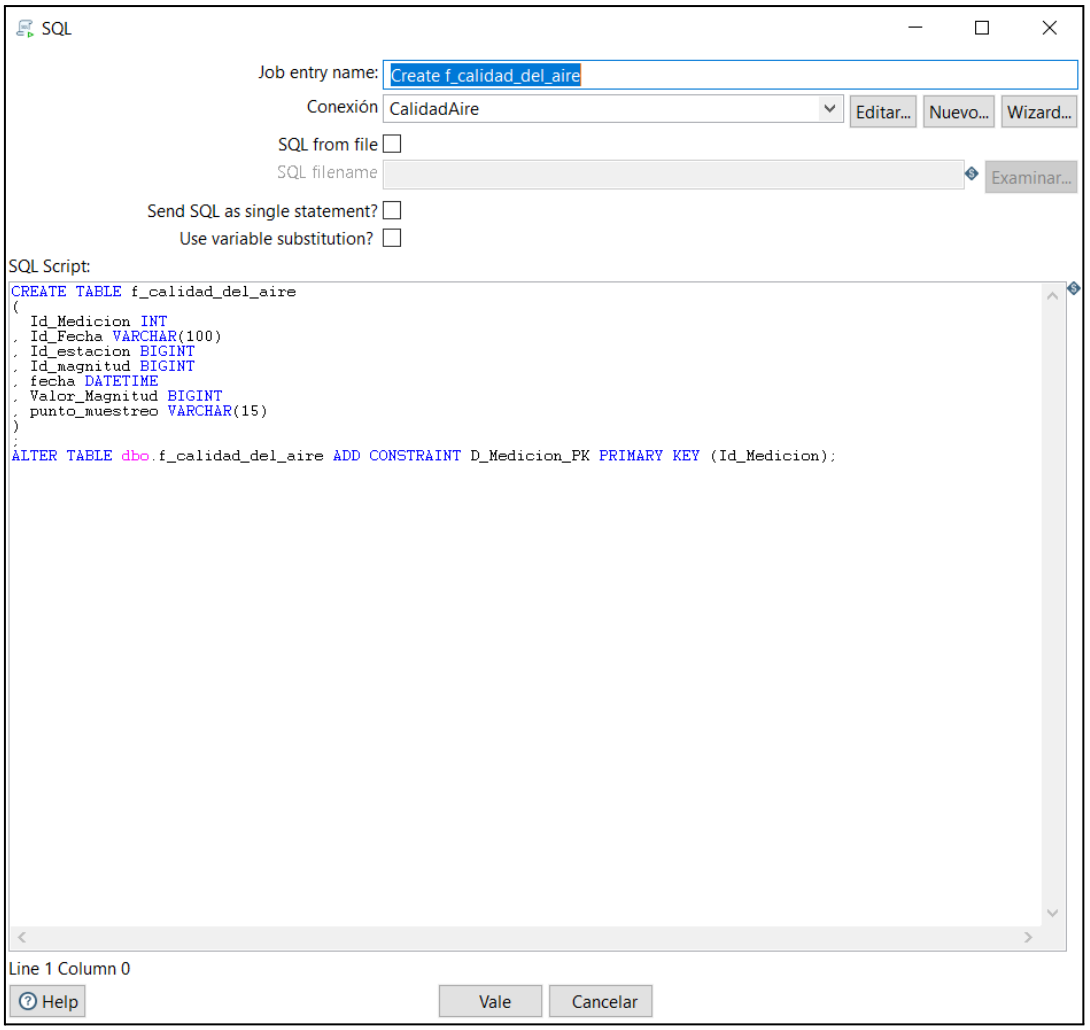


6.7.12.Script de creación de la tabla Meteorología

- Creación de la tabla Calidad del aire



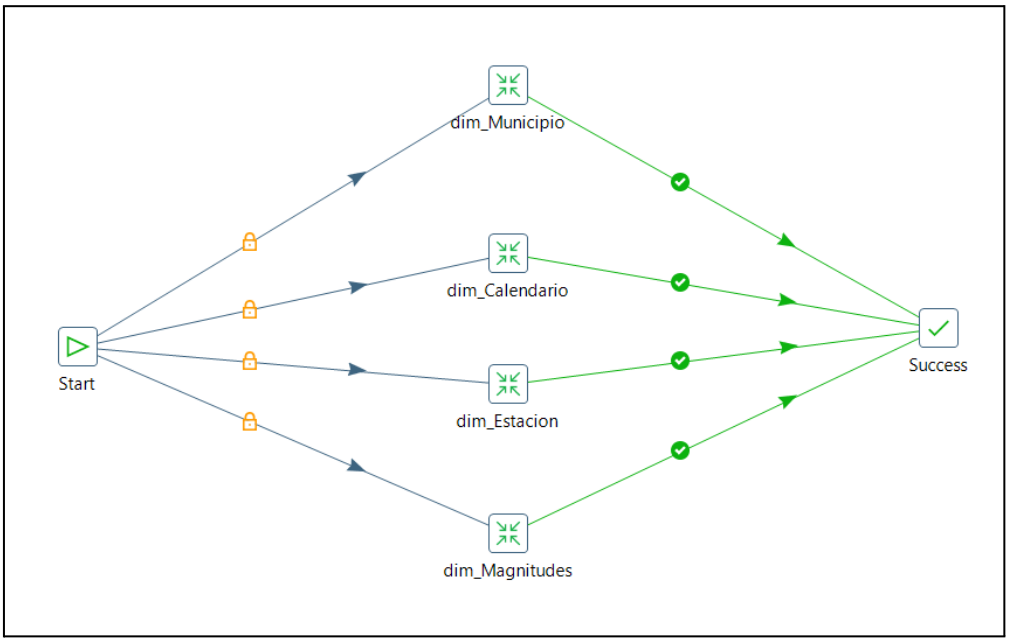
6.7.13. Comprobación de la existencia de las tablas



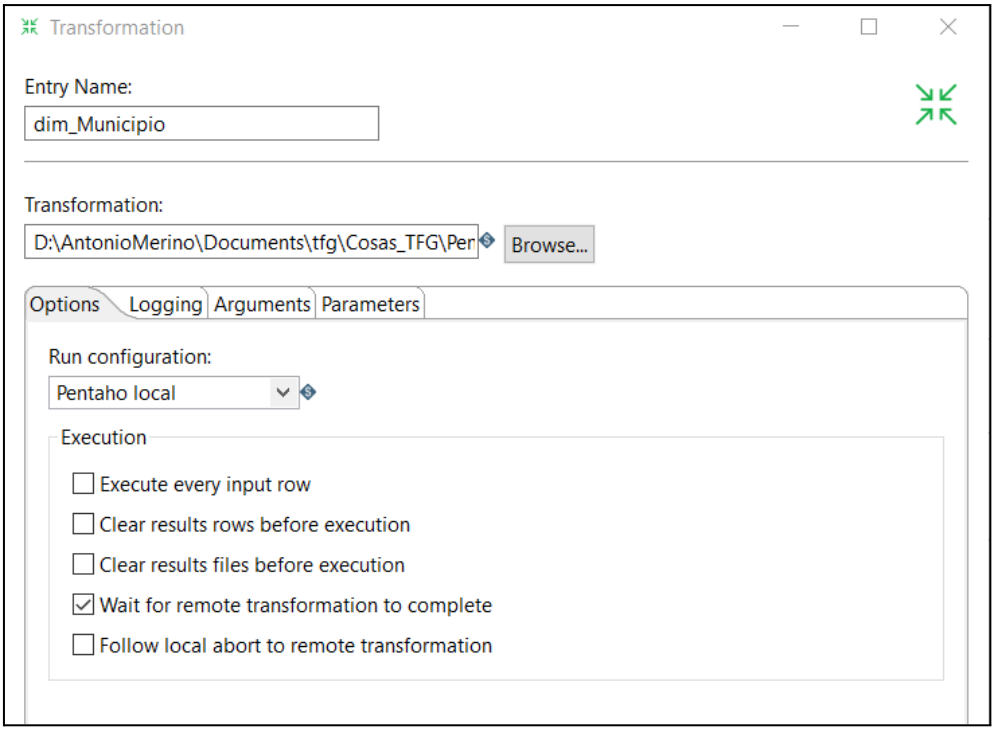
6.7.14.Script de creación de la tabla Calidad del aire

6.8. Carga de las dimensiones

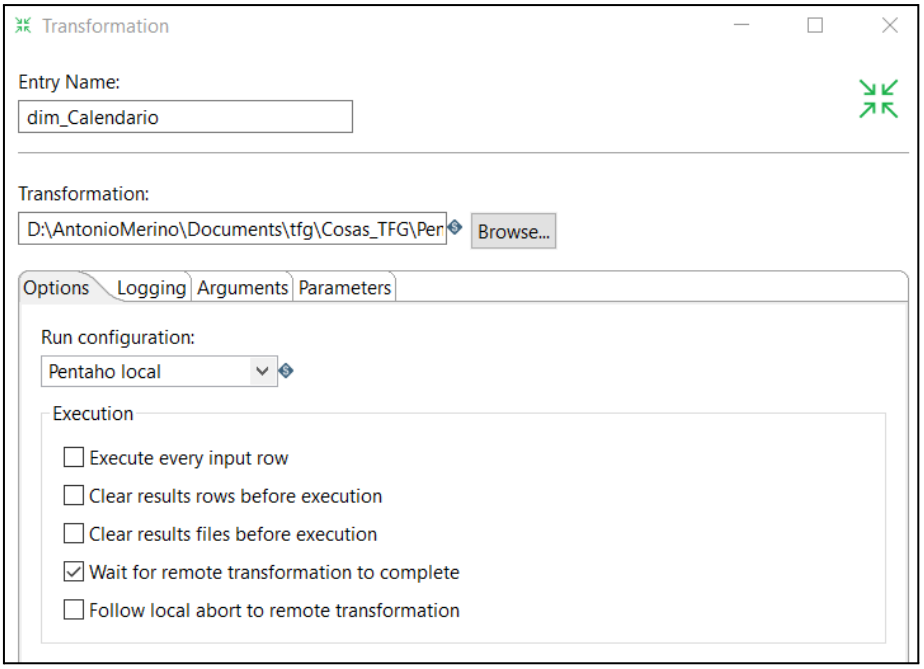
En este job, realizaremos la carga de las dimensiones, dicha carga será paralela, pues los datos de las diferentes dimensiones son independientes entre sí y conseguiremos optimizar el tiempo de carga del conjunto de las dimensiones.



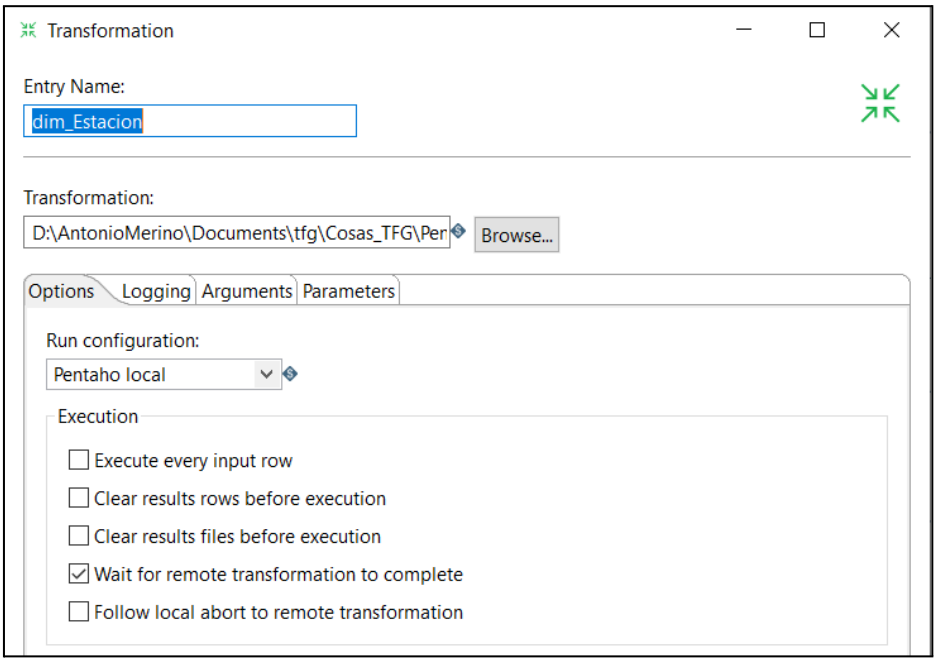
6.8.1. Carga de las dimensiones



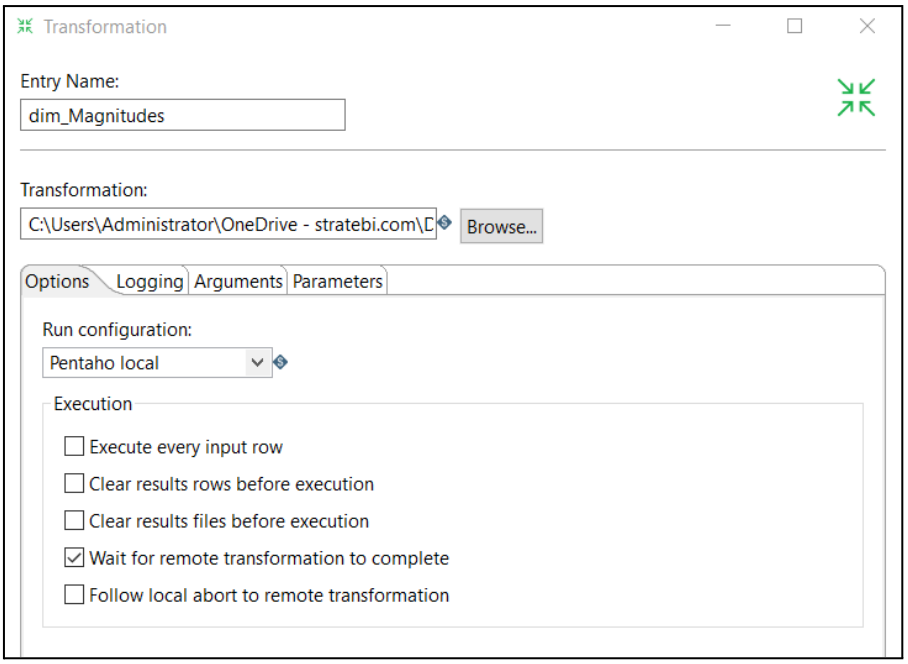
6.8.2. Configuración de la carga de la dimensión Municipio



6.8.3. Configuración de la carga de la dimensión Calendario



6.8.4. Configuración de la carga de la dimensión Estación



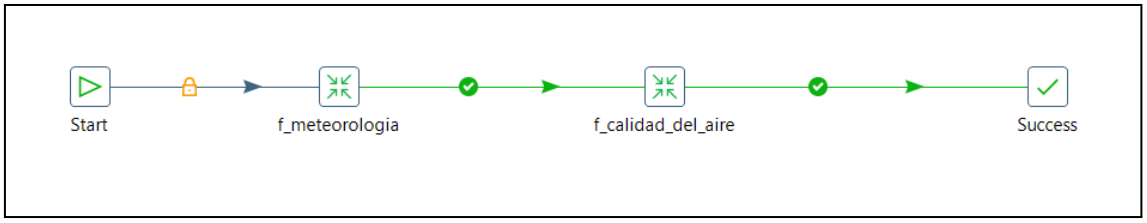
6.8.5. Configuración de la carga de la dimensión Magnitudes

Este paso realiza una llamada a la transformación creada anteriormente para que pase a ejecutarse.

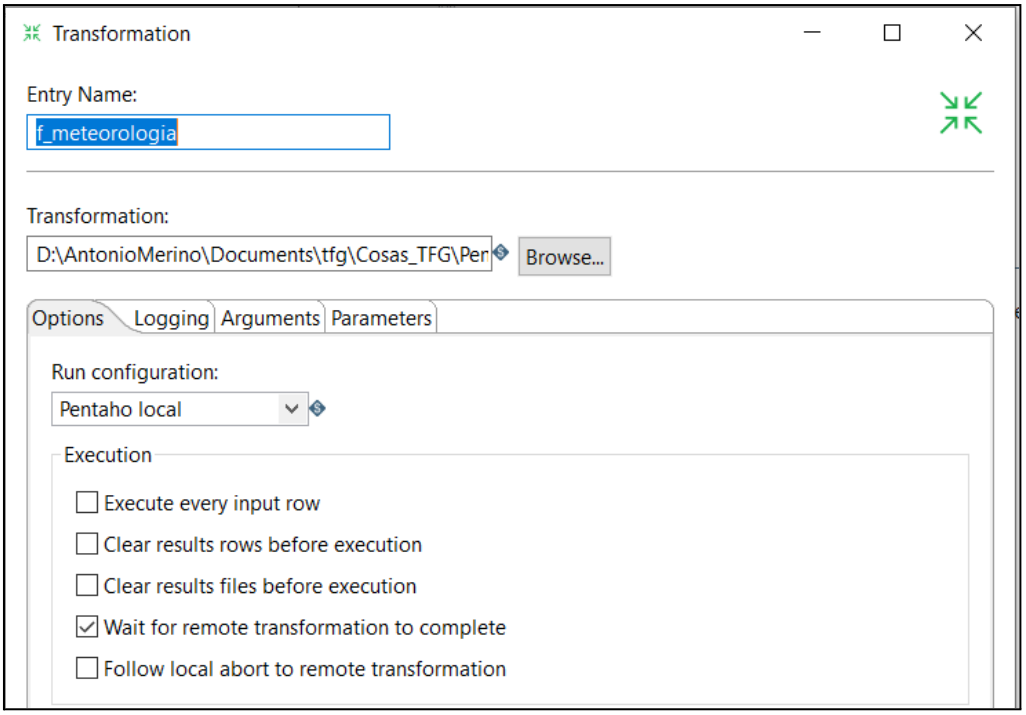
Este proceso se repetirá para cada una de las otras 3 dimensiones restantes.

6.9. Carga de las tablas de hechos

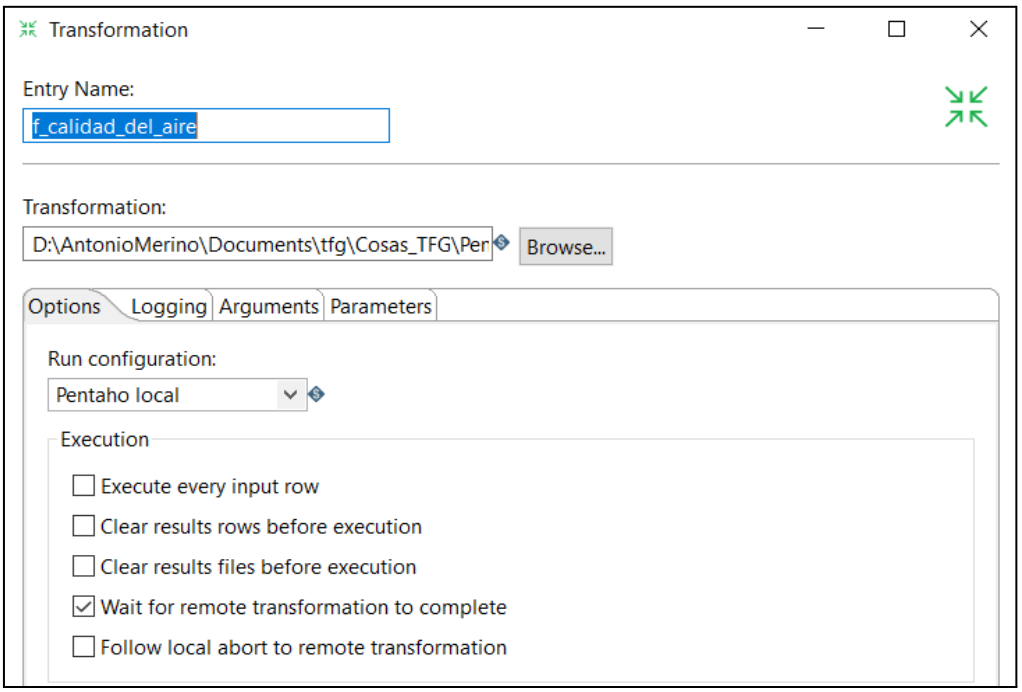
Gracias a este job realizaremos la carga de las dos tablas de hechos, meteorología y calidad del aire. Se trata de un proceso similar a la carga de dimensiones vista en el punto anterior.



6.9.1. Carga de las tablas de hechos



6.9.2. Configuración de la carga de la tabla de hechos metereología

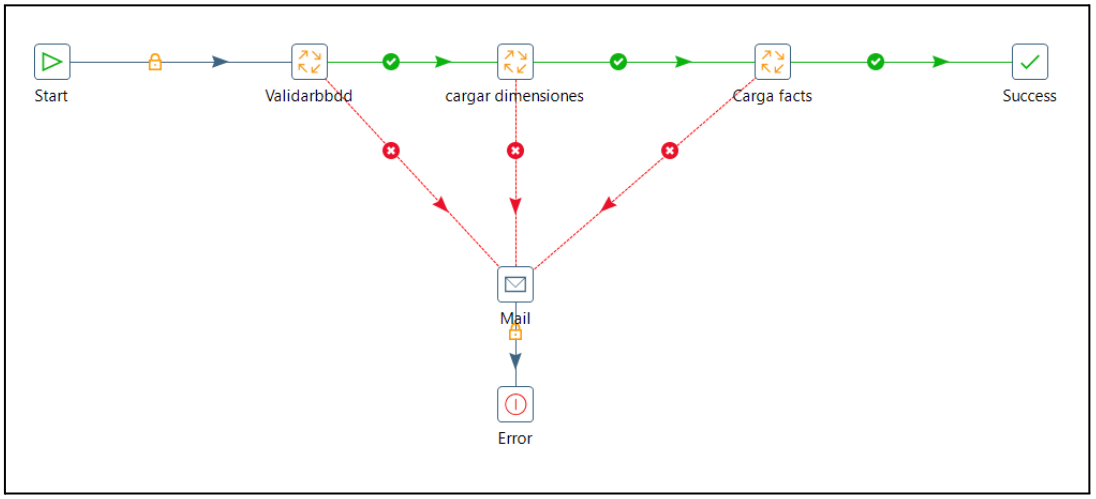


6.9.3. Configuración de la carga de la tabla de hechos de la calidad del aire

6.10. Carga total

Este último job, contiene el conjunto de todos los jobs creados anteriormente, con la configuración de un proceso global como es el que estamos definiendo en este punto optimizamos la ejecución de los procesos, de esta forma, solo será necesario ejecutar este job para que se ejecute el total de procesos que han intervenido en este proyecto.

Debido a que los datos vienen en formato csv/excel este proceso es un poco más tedioso ya que tendríamos que añadir manualmente los archivos a cada una de las transformaciones cada vez que se quieran añadir datos nuevos a la base de datos. Si los datos vinieran de una base de datos este proceso se podría automatizar para que se ejecutará periódicamente y así tener los datos más recientes en cada momento.



6.10.1. Carga total del proceso

Debido a que la herramienta dispone de un log con la ejecución del proceso en caso de que alguno de los 3 jobs fallaran esto quedaría registrado en ese log.
Además si se dispone de un servidor SMTP se podría generar de manera automática un correo electrónico que avisa de manera remota de si el proceso ha sufrido algún problema, de esta manera el proceso no requiere que nadie supervisando in situ.

7. Cuadro de Mandos

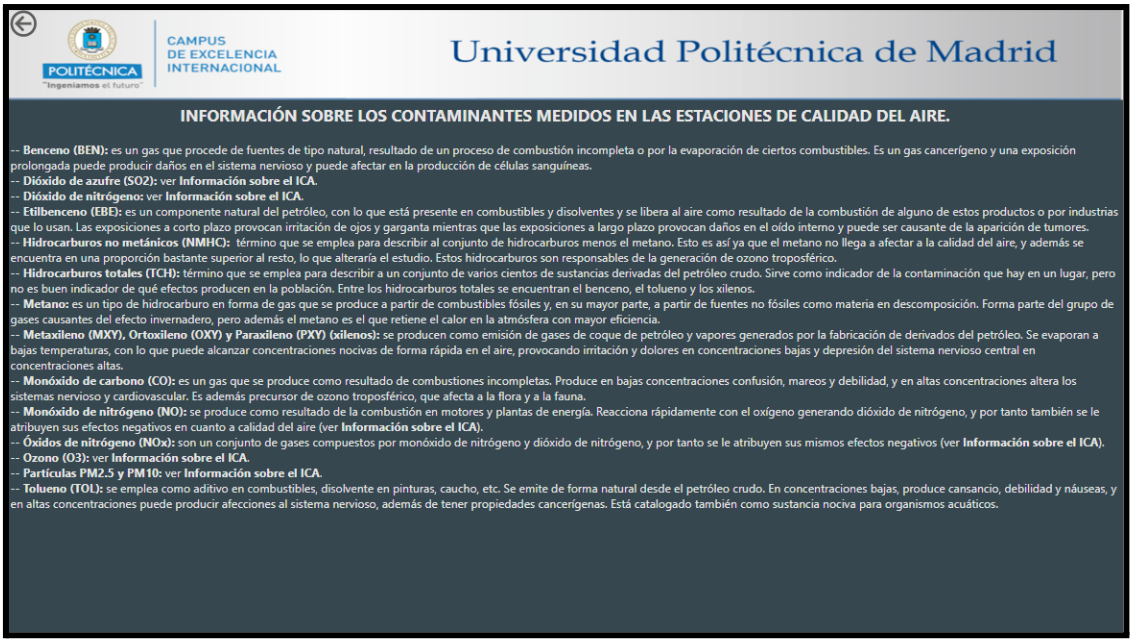
En este capítulo vamos a proceder a explicar el informe creado con la herramienta PowerBi, este informe una vez que ha sido creado, se publicará en el servicio de PowerBi en la web para así poder compartirlo con los usuarios que vayan a hacer uso del mismo. Adicionalmente, hemos realizado una guía de uso para que cualquier usuario pudiera utilizar este informe y obtener datos de el.

En primer lugar, contamos con una página principal a modo de portada, la cual servirá también de menú de navegación entre las diferentes páginas del informe, de esta manera se hace mucho más gráfica y visual la interacción con el informe para que aun sin saber manejar el sitio web de PowerBi donde esta alojado el informe se pueda hacer uso de este.

7.1. Portada del informe Power BI

En la imagen podemos ver lo comentado anteriormente y los 4 elementos del menú que nos dirigirán a cada una de las 4 páginas del informe.

Empezaremos por la primera página que nos aparece en el menú “Información Contaminantes” para navegar a ella solo tendremos que pulsar sobre el nombre y nos encontraremos la siguiente página:



7.2. Descripción de las partículas contaminantes I

En esta tenemos una descripción de todos los contaminantes que pueden llegar a ser medidos en la Comunidad de Madrid en todas las estaciones, a su vez tenemos la descripción de todos ellos para que el usuario entienda en todo momento qué es lo que se está midiendo en los gráficos.

También podemos observar que ciertos contaminantes tienen una referencia a otra página del informe en la que son explicados con mayor detalle.

En la esquina superior izquierda encontramos una flecha que nos permite volver al menú principal con un solo click para poder ir a otra de las páginas.

Si seleccionamos la segunda página “Información sobre el ICA” nos encontraremos con esta imagen:

7.3. Descripción de las partículas contaminantes II

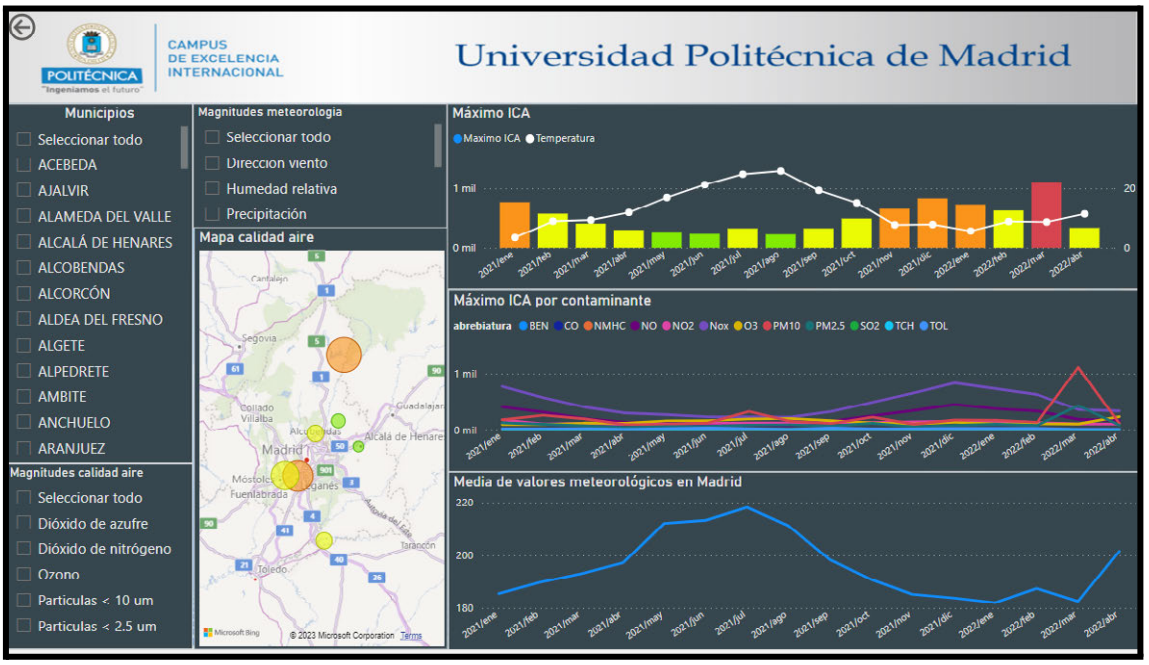
En la mitad superior tenemos los 4 contaminantes que tienen más impacto en el índice de calidad del aire, estos están definidos en detalle ya que son lo que se analizará principalmente en los gráficos más adelante.

En la mitad inferior tenemos una gama de colores, la cual, va relacionada al grado de peligro que supone la calidad del aire en ese momento del tiempo, esto es útil ya que ciertos gráficos utilizan esta gama de colores y nos permite identificar de manera muy rápida y visual cuales son los momentos en los que hay mejor o peor calidad del aire.

Como en la página anterior disponemos del botón para volver al menú principal.

Describiremos a continuación las páginas principales del informe, en las que tenemos las tablas y gráficos que analizan la información, así como los filtros que nos permiten centrarnos en conjuntos de datos específicos.

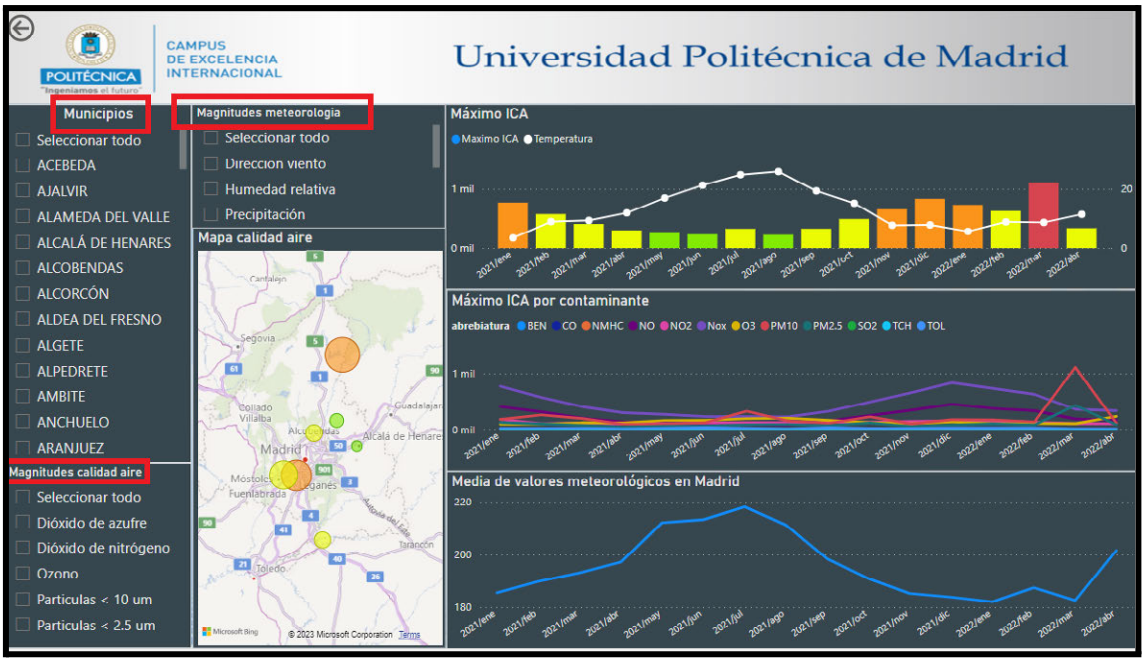
Comenzamos por la pagina “Indice de calidad del aire por estación” , al pinchar en el botón nos dirigirá a una página de este estilo:



7.4. Calidad del aire por estación

En esta página tenemos que hacer una diferenciación entre los diferentes elementos que nos encontramos, estos son las gráficas y tablas y los filtros.

Vamos a comenzar explicando los filtros y cuales son sus usos, identificamos tres diferentes filtros: Municipios, Magnitudes de calidad del aire y Magnitudes meteorológicas.

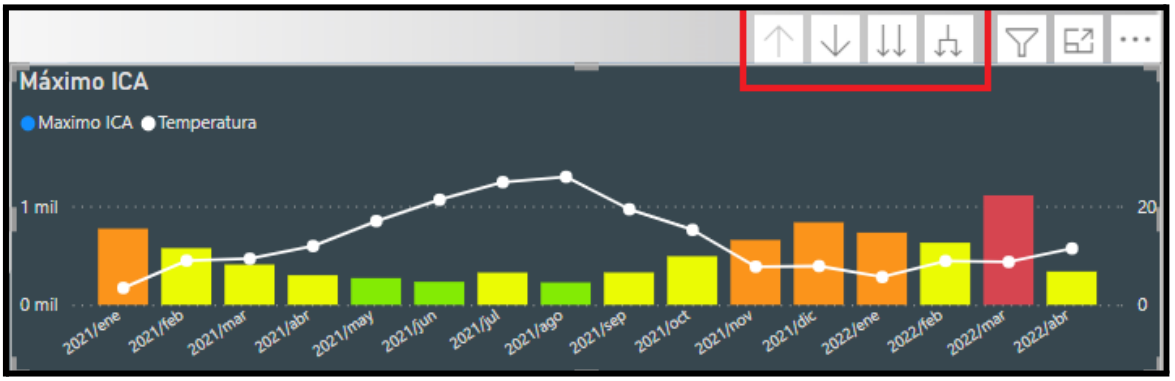


7.5. Calidad del aire por estación, filtros

Estos filtros son del tipo Lista vertical, y permiten modificar las tablas que tenemos en la página para así acotar los datos a lo que necesitemos. Esta selección se realiza pinchando encima de cualquiera de las casillas, permitiendo seleccionar varias casillas en el mismo filtro manteniendo pulsado la tecla CTRL mientras las seleccionamos, en este caso tenemos 3 posibles filtros que son : Municipios, Magnitudes meteorología y Magnitudes calidad del aire.

Como es lógico los filtros solo afectan a las tablas con las que tengan sentido lo que estamos filtrando, como por ejemplo si filtramos una magnitud de meteorología observaremos que la tabla de calidad del aire no se verá afectada ya que lo hemos diseñado de esa manera para el correcto funcionamiento de la herramienta.

Continuando con esta página vamos a tratar la primera tabla que nos encontramos, “Máximo ICA”.



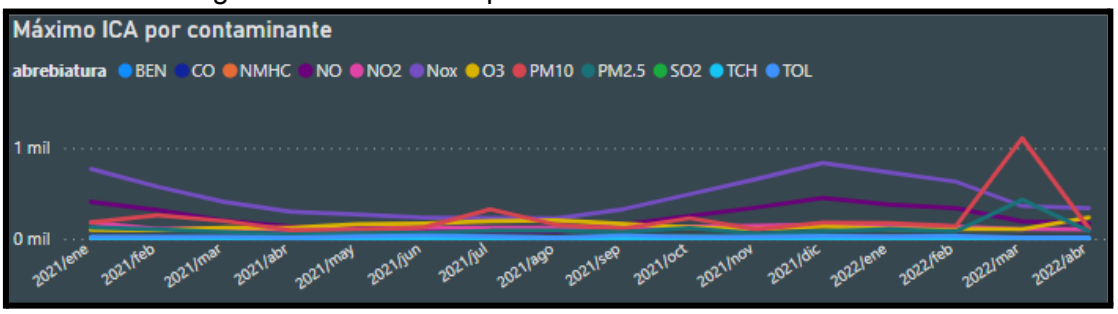
7.6. Máximo ICA

En esta tabla vamos a medir cuál fue el valor máximo del índice de calidad del aire en las barras verticales y a su vez estamos midiendo la temperatura mediante una línea superpuesta a las barras, todo esto lo medimos en un rango temporal variable, ya que tenemos la opción

de variar el eje x entre 3 rangos temporales, los cuales serían el mes, el dia y la hora. Para así poder hacer un análisis exhaustivo de los datos a todos los niveles. Esto es posible con las opciones que nos aparecen al pasar con el cursor por encima del gráfico, mostradas en el recuadro rojo en la imagen.

Como hemos mencionado anteriormente esta tabla incorpora el rango de colores para poder identificar de manera rápida si la calidad del aire es buena o mala con un simple vistazo.

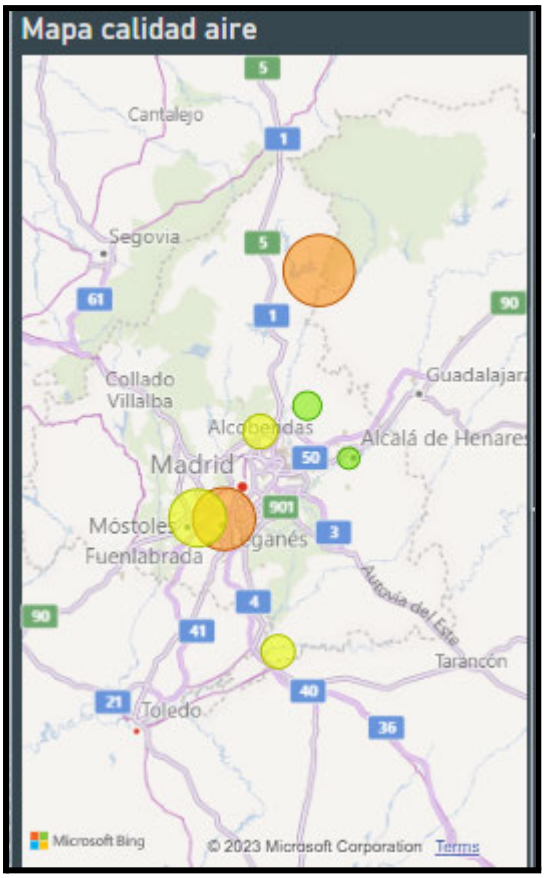
Continuando con la gráfica “Máximo ICA por contaminante”



7.7. Máximo ICA, por contaminante

En esta gráfica vamos a medir cual es el máximo valor del ICA de cada uno de los contaminantes, volvemos a tener esa flexibilidad en el eje x para poder variar entre los 3 rangos temporales.

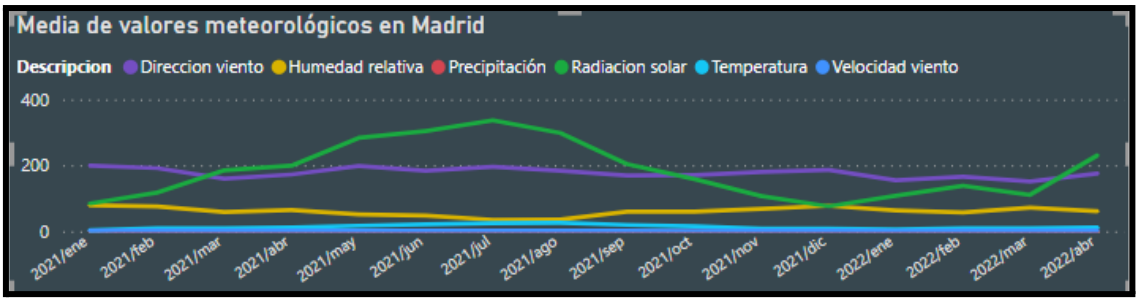
Encontramos otra grafica un poco peculiar, que es un mapa



7.8. Mapa de calidad del aire

Este mapa tiene las estaciones de la comunidad de madrid de las que disponemos de datos, y mide el promedio de su ICA, cuanto más alto sea, el círculo que las identifica será mayor, además volvemos a contar con la gama de colores para ayudarnos a identificar si es buena o mala la calidad del aire en esa zona.

Por último, en la página encontramos la gráfica “Media de valores meteorológicos en Madrid”.

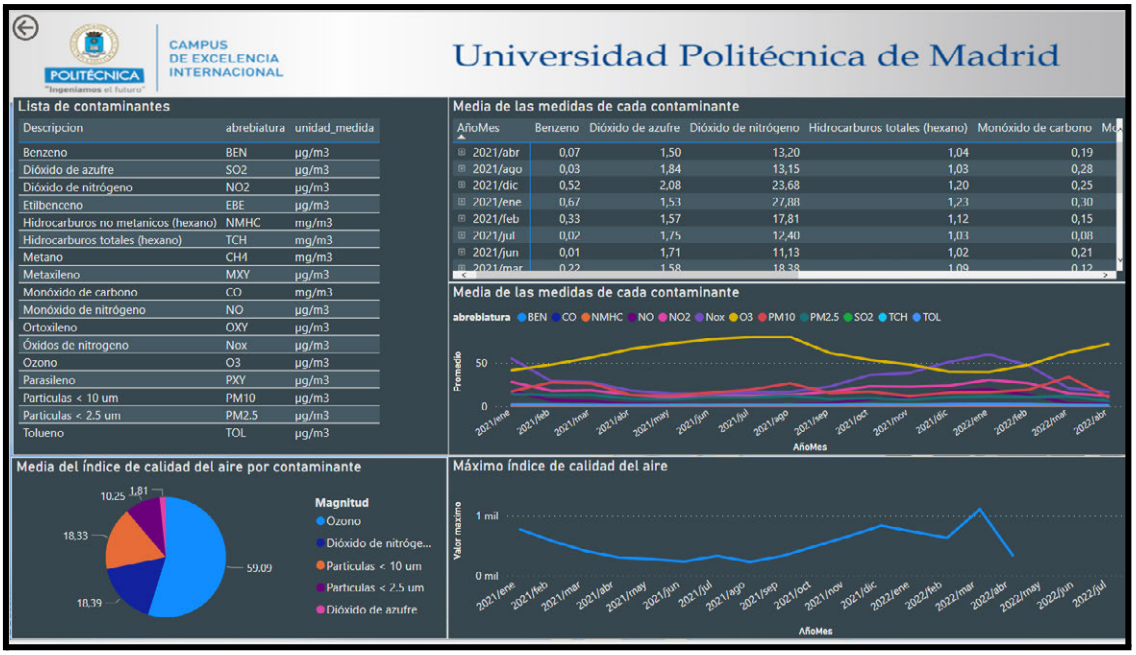


7.9. Media de valores meteorológicos en Madrid

En esta gráfica vamos a poder analizar la media de los factores meteorológicos para así poder ver la relación directa que tiene con la calidad del aire. Esta tabla se puede filtrar por el filtro que hemos visto anteriormente para así centrarnos en una sola magnitud, comparte con las otras 2 tablas anteriores la posibilidad de jugar con el eje x.

Cabe destacar otra forma de filtrar que nos ofrece la herramienta y esta es pinchando encima de alguno de los elementos que se encuentran dentro de las tablas, por ejemplo con esta última tabla si pinchamos encima de precipitación en la leyenda, todas las tablas de la página se filtraron con esta magnitud, lo que nos ofrece una cantidad de opción de filtrado muy elevada.

Ya por ultimo nos dirigimos a la pestaña “Medidas por contaminantes” donde nos encontraremos la siguiente página:



7.10. Contaminantes

Como podemos observar, esta página se compone mayoritariamente de gráficas, solo disponemos de un filtro, el cual se encuentra en la parte superior izquierda “Lista de contaminantes” que tiene 2 funciones, una es la de tener un resumen con las medidas su abreviatura y en qué unidad se miden y la otra nos sirve para poder filtrar la página entera por alguna o varias de las medidas en específico simplemente haciendo click encima de ellas.

En esta página lo que vamos a medir principalmente son los valores de las medidas y así como el índice de calidad del aire representado en un formato temporal.

Con las 4 gráficas que vemos en la página podemos hacer un estudio detallado de los valores de los contaminantes, ya que vamos a poder saber la media en cualquier instante del tiempo de los valores e incluso el valor máximo que se ha alcanzado del índice de calidad del aire.

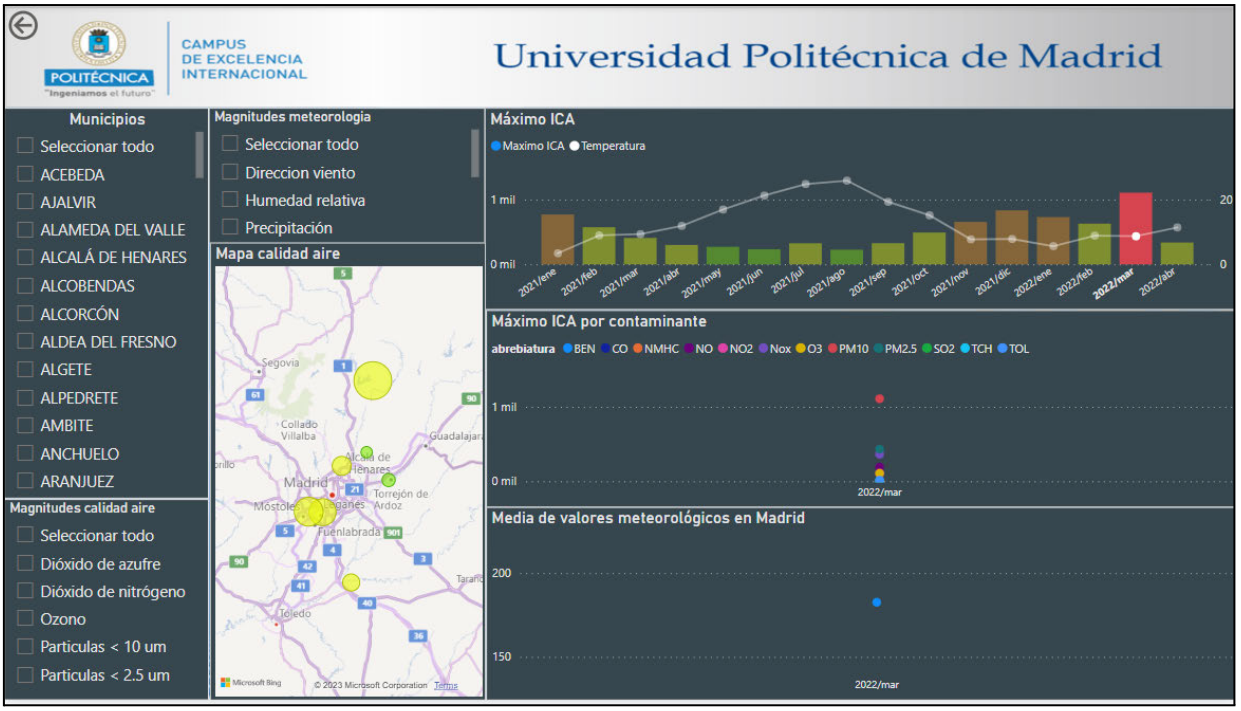
Los datos podemos observarlos en un pie chart en formato tabla y en formato gráfica para con pocos clicks poder tener visualmente un estudio tanto específico del dato como una evolución más visual a lo largo de la línea de tiempo, así como el dato general.

Esta página tiene al igual que la anterior las mismas posibilidades de filtrado al hacer click encima de cualquier elemento de las gráficas para así modificar las demás en función de lo requerido.

Esto nos permite que jugando con las 2 páginas en las que disponemos de gráficas poder hacer un estudio minucioso de la calidad del aire para poder obtener todo tipo de conclusiones.

8. Análisis de datos

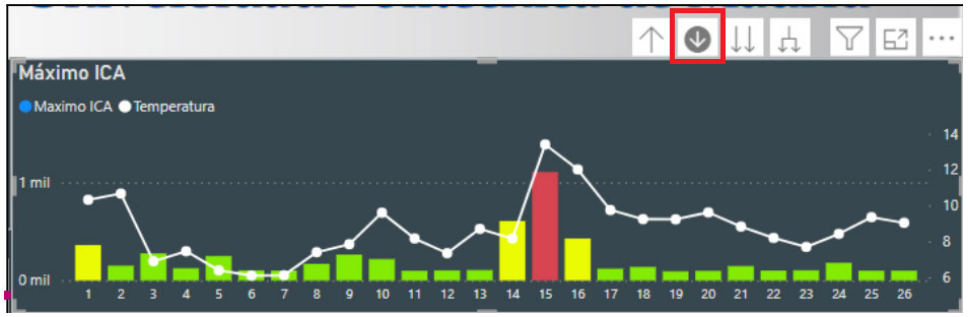
Comenzaremos con el análisis del mes en el que ha habido un valor máximo en cuanto a alguno de los parámetros medibles de la calidad del aire, esto no representa que el mes en cuestión, Marzo del 2022 haya sido el peor en comparación con el resto de meses..
Representa que ha sufrido un pico máximo en alguno de los contaminantes evaluados.



8.1. Mes con máximo ICA

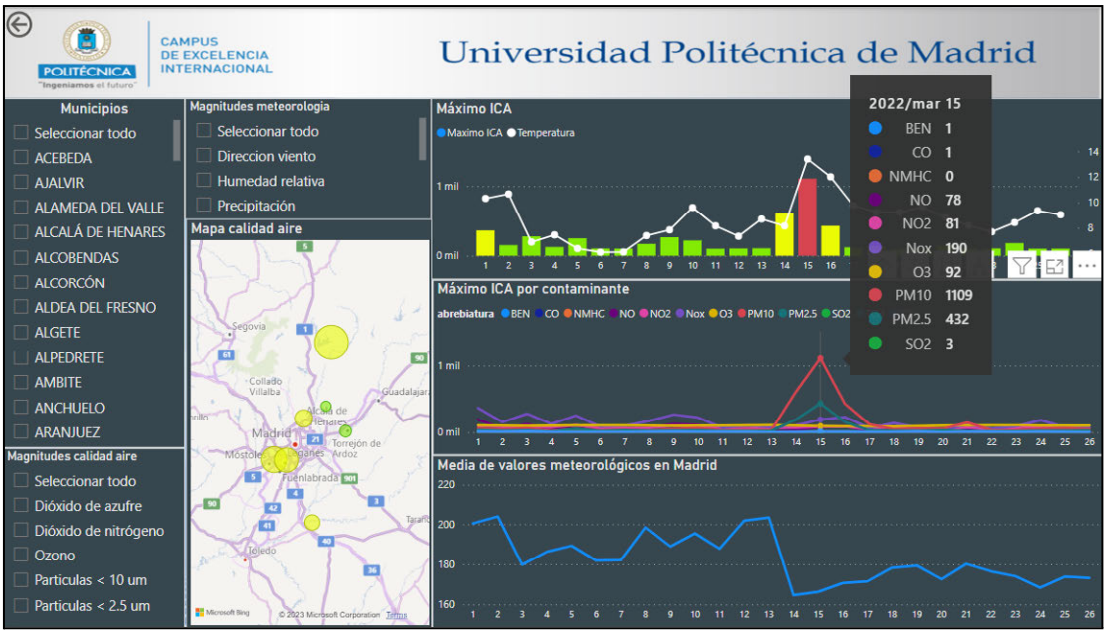
A continuación, pulsando en la opción de ampliación podemos acceder al detalle del mes y así ver de manera detallada el transcurso de los días.

El día con mayor ICA se corresponde con el 15 de Marzo.



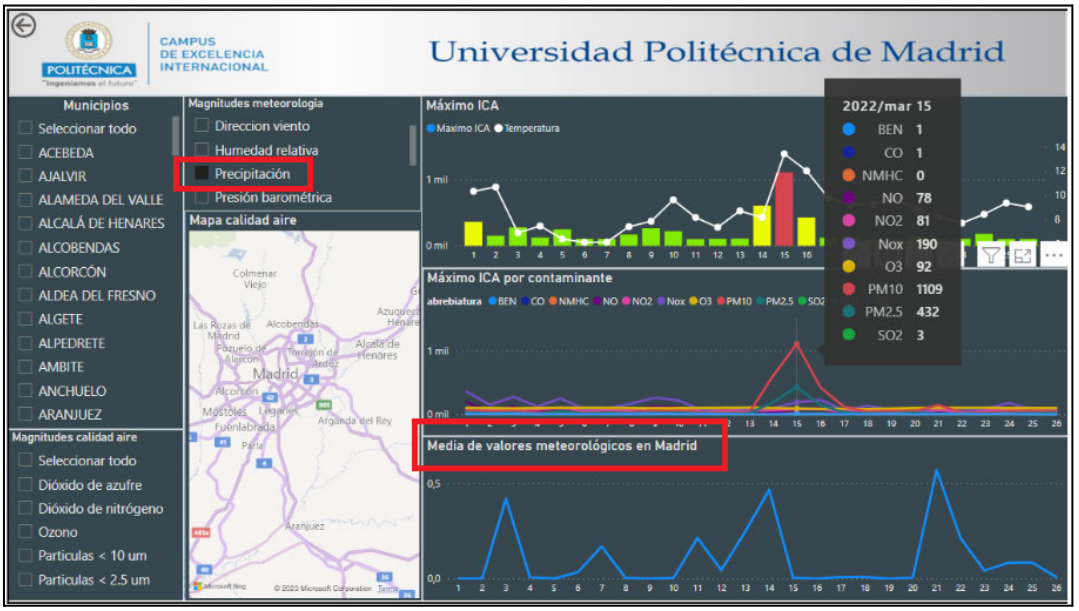
8.2. Mes con máximo ICA ampliación de detalle

Observamos que el pico se debe al contaminante PM10, también tenemos un pico del contaminante PM2.5.



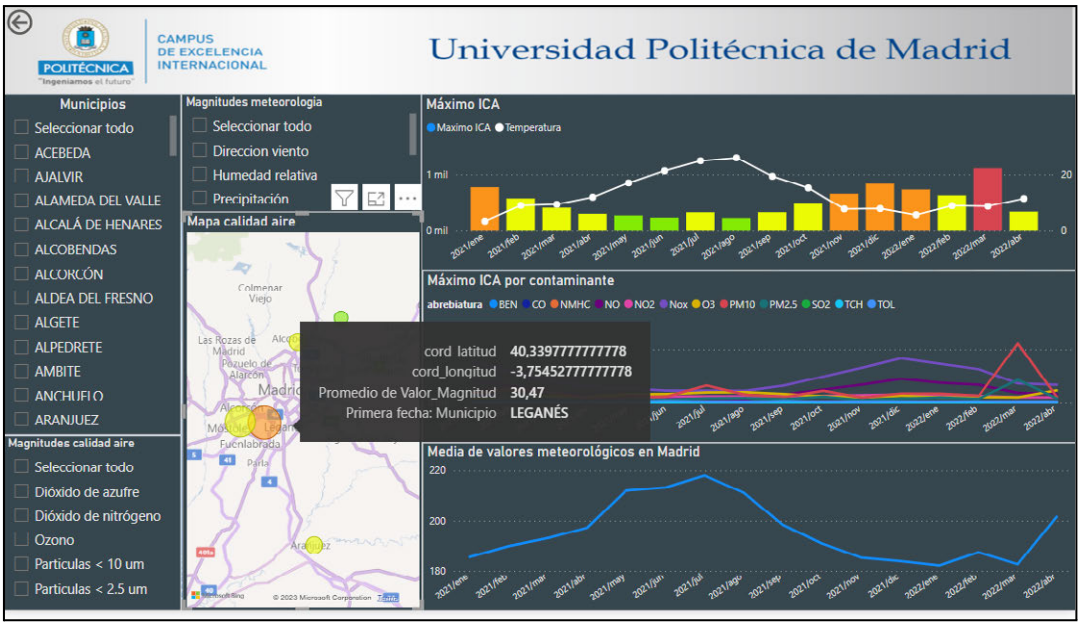
8.3. Mes con máximo ICA ampliación de detalle II

Filtrando por la magnitud "Precipitación" obtenemos el gráfico de precipitaciones por día, es decir, litros por metro cuadrado al día.



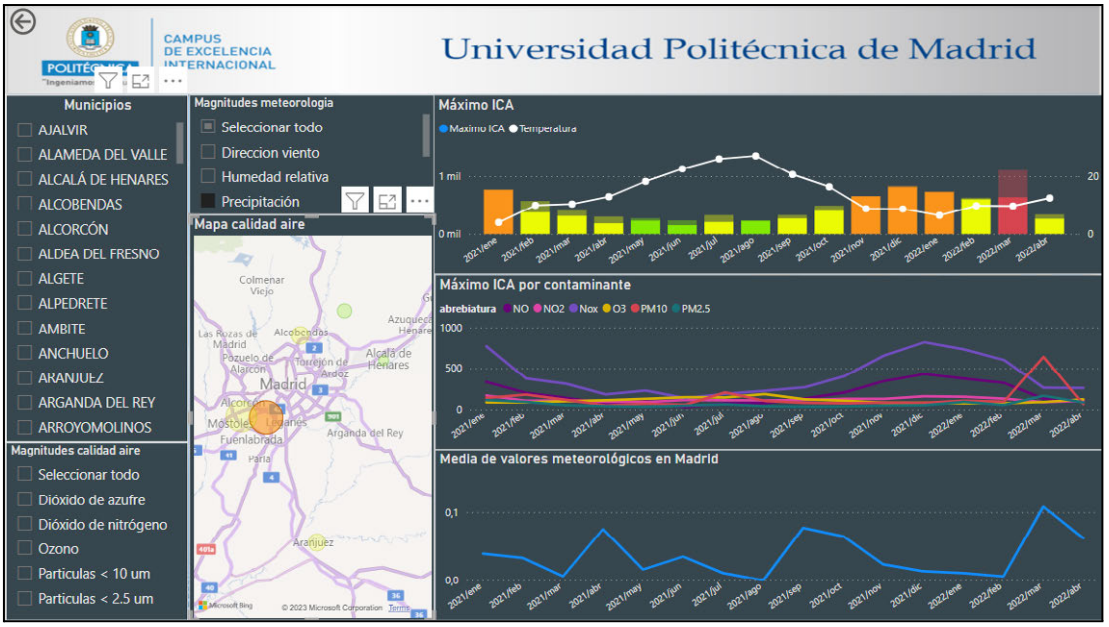
8.4. Mes con máximo ICA ampliación de detalle, filtrando por precipitaciones

Posteriormente, nos centramos en la zona con peor promedio de calidad del aire, en el mapa podemos seleccionar la posición que nos interesa.
En este caso vamos a analizar el municipio de Leganés que es el que cuenta con peor promedio de ICA en comparación con los demás municipios a lo largo del tiempo.



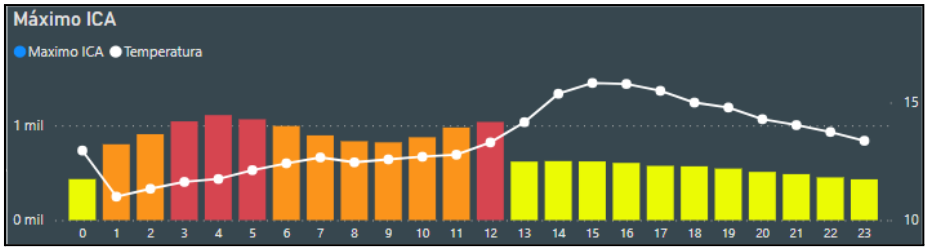
8.5. Mes con máximo ICA ampliación de detalle, coordenadas

Observando los gráficos, vemos claramente la supremacía del promedio de calidad del aire de este municipio en comparación con el resto



8.6. Mes con máximo ICA ampliación de detalle, filtrando por magnitudes

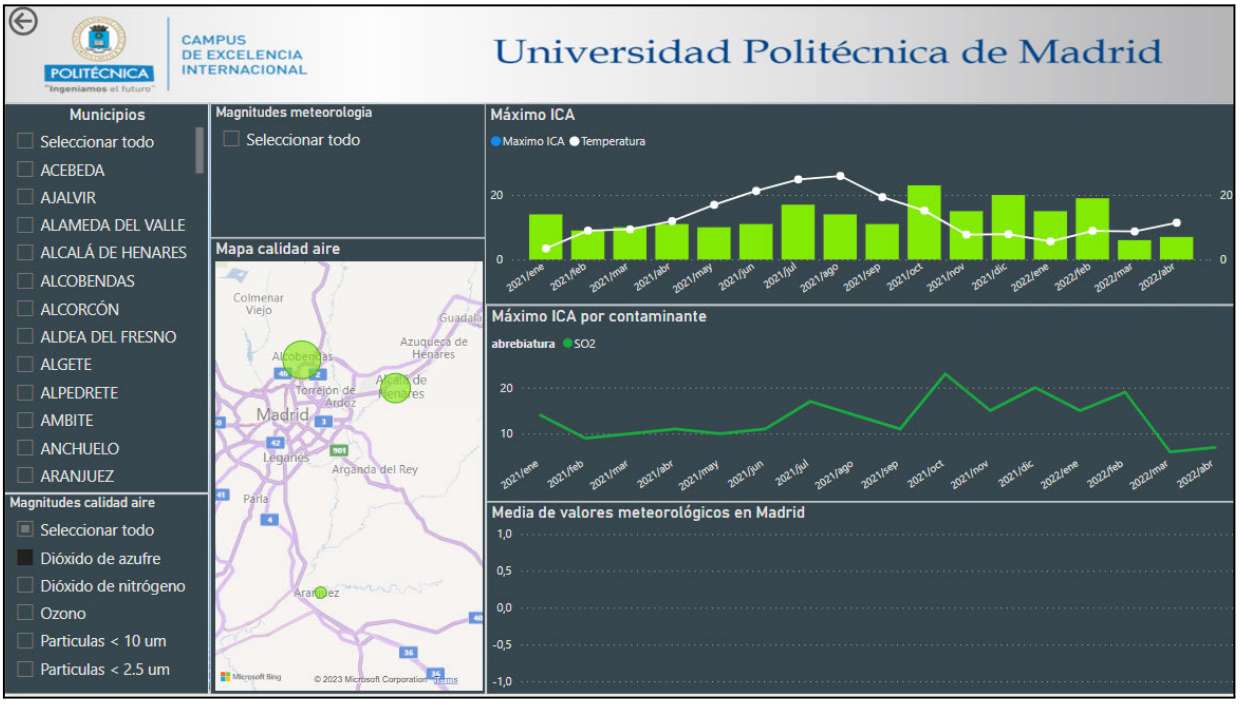
Ampliando el mes de Marzo podemos observar los contaminantes en función de la hora y temperatura del día.
Observamos que cuanta más temperatura tenemos menor ICA tenemos.



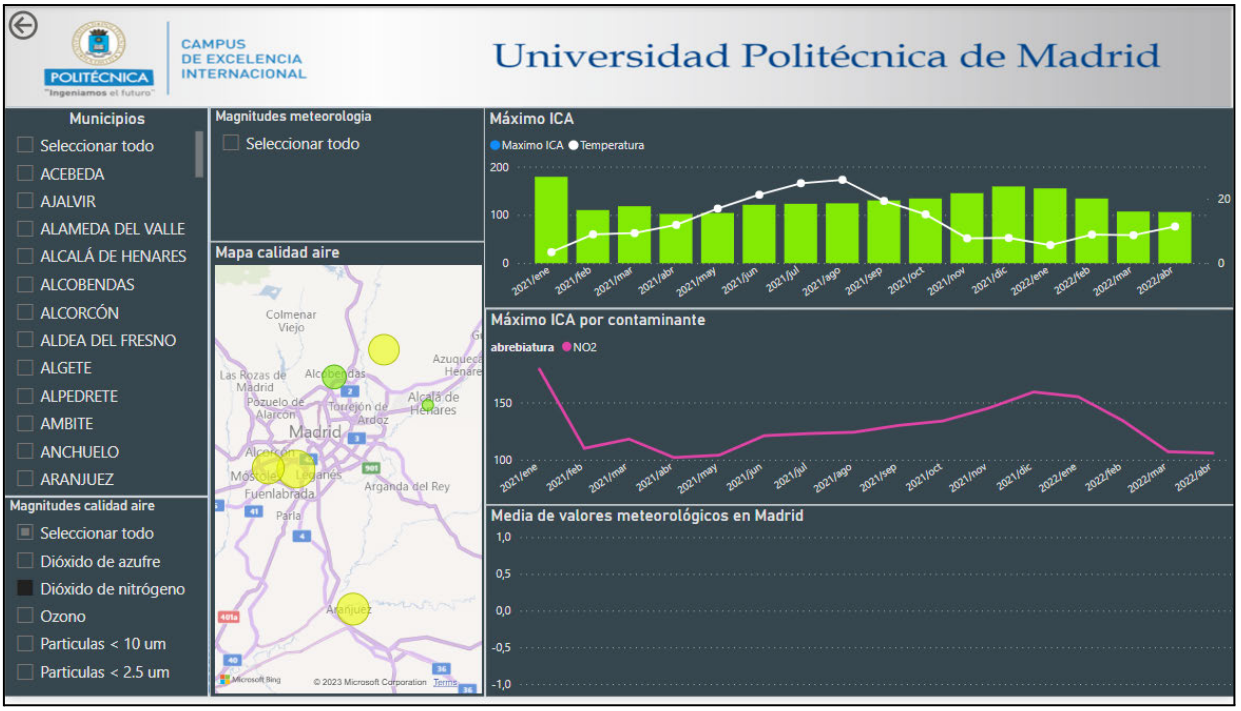
8.7. Mes con máximo ICA, detalle por día

Al filtrar por las magnitudes del aire se suprime la visibilidad de las magnitudes de metereología.

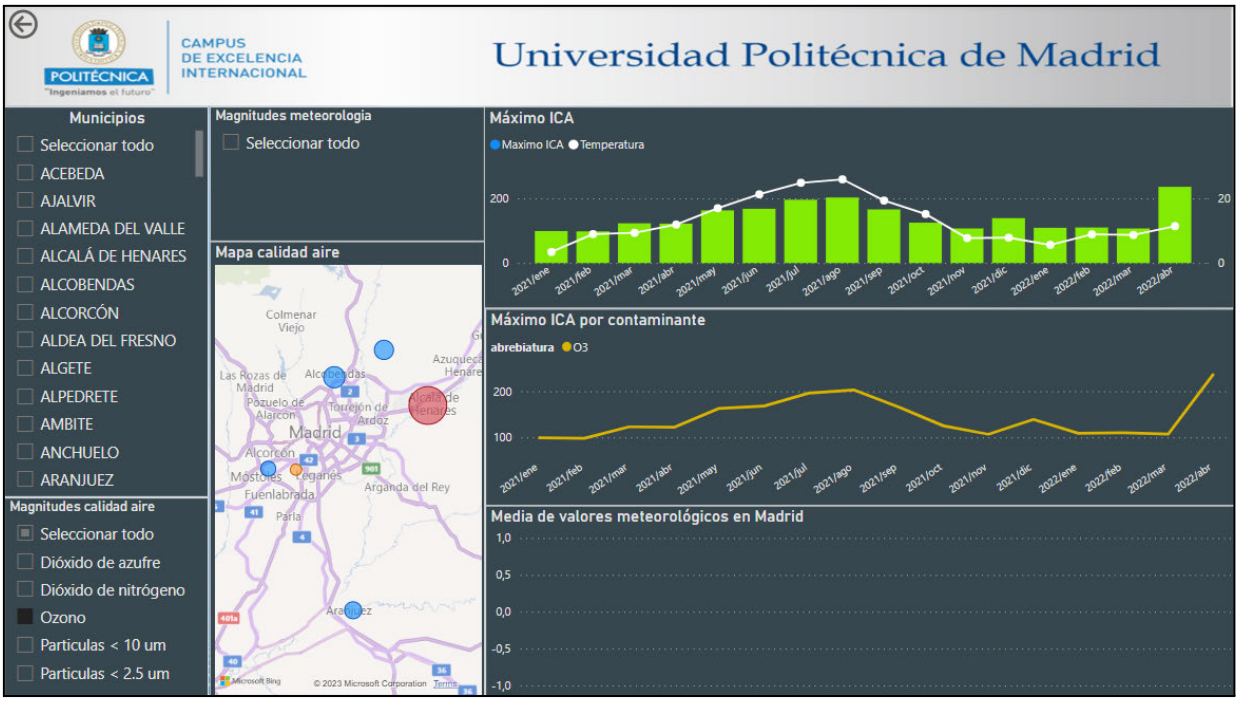
Filtrando por las principales magnitudes obtenemos el valor máximo de cada contaminante.



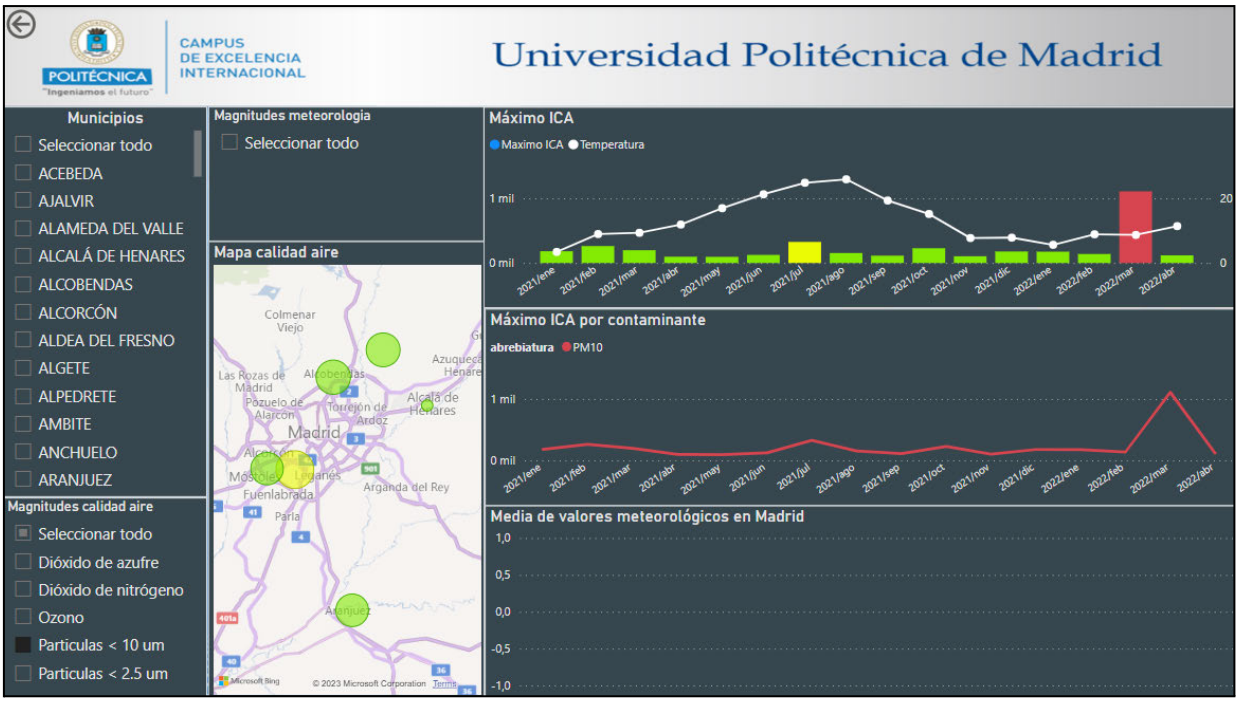
8.8. Mes con máximo ICA ampliación de detalle, filtrando por dióxido de azufre



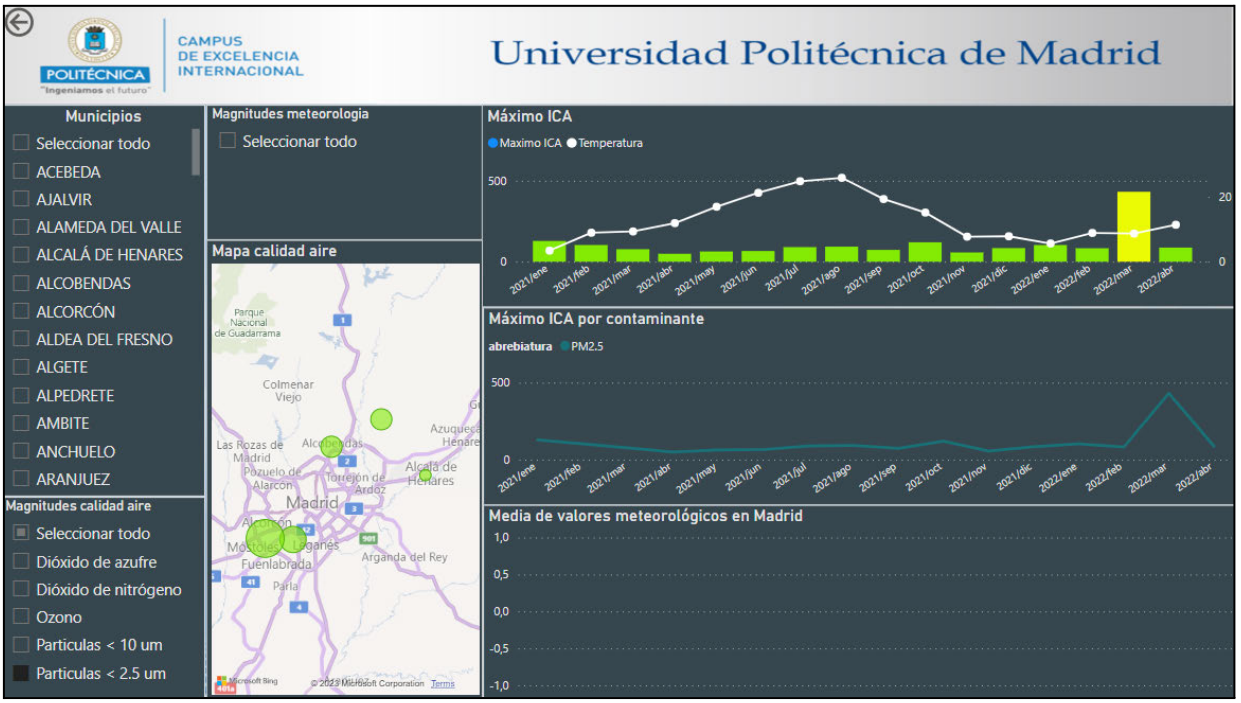
8.9. Mes con máximo ICA ampliación de detalle, filtrando por dióxido de nitrógeno



8.10. Mes con máximo ICA ampliación de detalle, filtrando por ozono



8.11. Mes con máximo ICA ampliación de detalle, filtrando por Partículas < 10 um

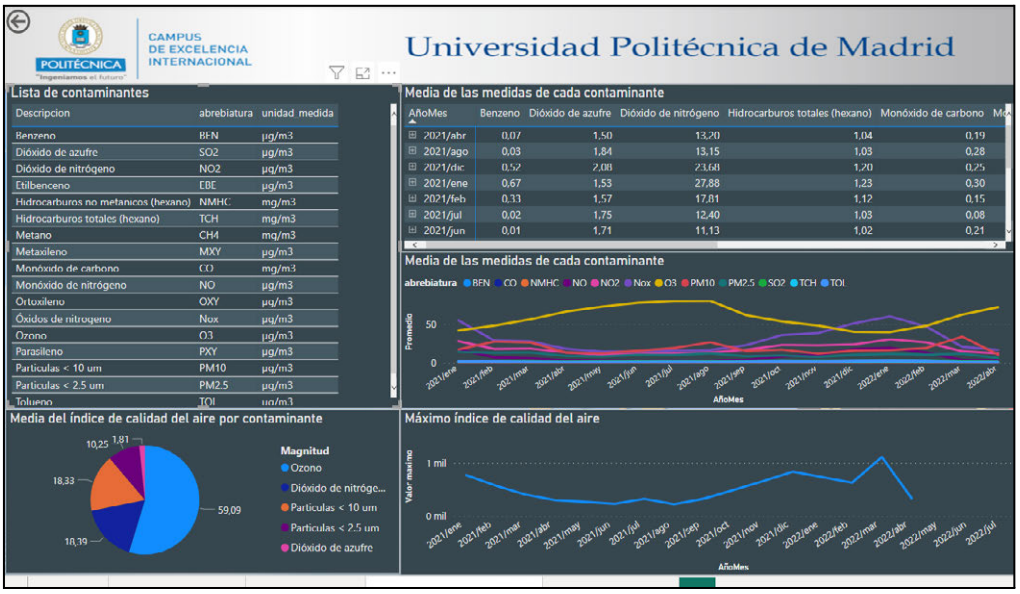


8.12. Mes con máximo ICA ampliación de detalle, filtrando por Partículas < 2.5 um

En la siguiente página del cuadro de mandos podemos filtrar también por magnitudes y ver los datos de manera más representativa, así como, la media por contaminante.

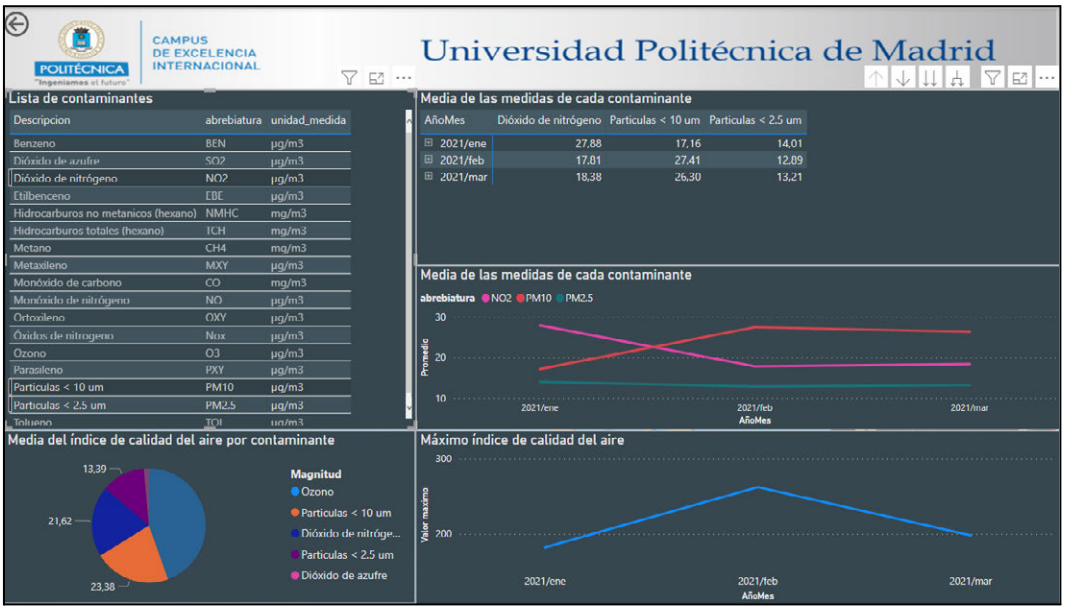
En el apartado “Media de las medidas de cada contaminante” podremos ver la media de las medidas por fecha. Y posteriormente en la gráfica de abajo podemos observar visualmente de manera lineal en el tiempo la evolución de la media de cada contaminante.

Esta página del cuadro de mandos también nos ofrece la información en forma de gráfico de tarta, como se muestra a continuación.



8.13. Media por contaminantes

De esta forma, si filtramos por las partículas NO2, PM10 y PM2,5 obtenemos la siguiente representación para los meses de enero hasta marzo.
Podemos comprobar que las partículas NO2 y PM2,5 tienen un efecto inversamente proporcional.



8.14. Media por contaminantes

Este análisis se puede ampliar por cada uno de los contaminantes en contraposición al resto para realizar las comparativas deseadas.

La calidad del aire en la Comunidad de Madrid puede variar según el mes del año debido a diferentes factores, como la temperatura, la humedad, la dirección y la intensidad del viento, la actividad económica y la movilidad. A continuación se describe la calidad del aire en Madrid según el mes del año, tomando como referencia los datos proporcionados por la red de estaciones de monitoreo de la Comunidad de Madrid.

Enero a marzo: Durante estos meses, los niveles de contaminación en la Comunidad de Madrid suelen ser más altos debido a la mayor actividad económica y al mayor uso de calefacción. Los niveles de dióxido de nitrógeno (NO2) y las partículas en suspensión (PM10 y PM2.5) pueden ser particularmente altos durante este período.

Abril y mayo: Durante estos meses, la calidad del aire en la Comunidad de Madrid suele mejorar debido a la disminución de la actividad económica y al aumento de la ventilación y la lluvia. Los niveles de ozono (O3) pueden ser más altos durante este período debido a la mayor intensidad de la radiación solar.

Junio a septiembre: Durante estos meses, los niveles de ozono (O3) pueden ser particularmente altos debido a la mayor intensidad de la radiación solar y a la menor ventilación. También pueden producirse episodios de contaminación por partículas (PM10 y PM2.5) debido a incendios forestales o a episodios de polvo africano.

Octubre a diciembre: Durante estos meses, los niveles de contaminación en la Comunidad de Madrid pueden aumentar debido a la mayor actividad económica y al aumento del uso de calefacción. Los niveles de dióxido de nitrógeno (NO2) y las partículas en suspensión (PM10 y PM2.5) pueden ser particularmente altos durante este período.

En general, la calidad del aire en la Comunidad de Madrid es peor durante los meses de invierno y de verano debido a la mayor actividad económica y a las condiciones meteorológicas. Se han implementado medidas para reducir la contaminación atmosférica en la región, como restricciones al tráfico, promoción del transporte público y la movilidad sostenible, y fomento de la energía limpia.

En Madrid, existen zonas con mayor contaminación del aire debido a una combinación de factores como el tráfico, la actividad industrial, la densidad de población y la ubicación de las estaciones de monitoreo de la calidad del aire. A continuación, se describen algunas de las zonas más afectadas por la contaminación del aire en la ciudad:

La zona centro: Es una de las áreas más afectadas por la contaminación del aire en Madrid debido a la alta densidad de tráfico y la gran cantidad de actividad económica y comercial. Los niveles de dióxido de nitrógeno (NO2) y partículas en suspensión (PM10 y PM2.5) suelen ser particularmente altos en esta zona.

La zona de la M-30: La M-30 es una de las vías más transitadas de Madrid y su tráfico es una de las principales fuentes de contaminación del aire en la ciudad. Los niveles de dióxido de nitrógeno (NO2) y partículas en suspensión (PM10 y PM2.5) son muy altos en esta zona.
El Corredor del Henares: Es una de las áreas más industrializadas de la Comunidad de Madrid y se encuentra al este de la ciudad. Los niveles de dióxido de nitrógeno (NO2) y

partículas en suspensión (PM10 y PM2.5) son particularmente altos en esta zona debido a la actividad industrial.

La zona sur de la ciudad: Los niveles de ozono (O3) son particularmente altos en esta zona debido a la mayor intensidad de la radiación solar y la menor ventilación.

A continuación, se describen algunos de los contaminantes más relevantes y cómo varían sus niveles según el mes de exposición:

Dióxido de nitrógeno (NO2): El NO2 es uno de los contaminantes más preocupantes en Madrid debido a su alta concentración en la ciudad. Los niveles de NO2 son más altos durante los meses de invierno, cuando las condiciones meteorológicas favorecen su acumulación en el aire. Los niveles de NO2 también varían según la ubicación de las estaciones de monitoreo, siendo las zonas más afectadas las cercanas a las vías de tráfico con mayor densidad.

Partículas en suspensión (PM10 y PM2.5): Las partículas en suspensión son uno de los contaminantes más peligrosos para la salud humana debido a su capacidad de penetrar en los pulmones y causar enfermedades respiratorias y cardiovasculares. Los niveles de PM10 y PM2.5 son más altos durante los meses de invierno, cuando la calefacción y la combustión de combustibles fósiles son más intensas. Los niveles de PM10 y PM2.5 también varían según la ubicación de las estaciones de monitoreo, siendo las zonas más afectadas las cercanas a las vías de tráfico y las áreas industriales.

Ozono (O3): El ozono es un contaminante que se forma a partir de la reacción entre los contaminantes emitidos por los vehículos y la radiación solar. Los niveles de ozono son más altos durante los meses de verano, cuando la intensidad de la radiación solar es mayor. Los niveles de ozono también varían según la ubicación de las estaciones de monitoreo, siendo las zonas más afectadas las cercanas a las áreas industriales y las zonas con menor ventilación.

A partir del análisis de los datos de calidad del aire de la Comunidad de Madrid, se pueden obtener varias conclusiones:

La calidad del aire en la Comunidad de Madrid varía según el contaminante y la ubicación de las estaciones de monitoreo.
El dióxido de nitrógeno (NO2) es uno de los contaminantes más preocupantes en Madrid debido a su alta concentración en la ciudad, especialmente en las zonas cercanas a las vías de tráfico con mayor densidad.
Las partículas en suspensión (PM10 y PM2.5) son otro contaminante relevante en Madrid, especialmente durante los meses de invierno, cuando la calefacción y la combustión de combustibles fósiles son más intensas.
El ozono (O3) es un contaminante que se forma a partir de la reacción entre los contaminantes emitidos por los vehículos y la radiación solar, y sus niveles son más altos durante los meses de verano.
La exposición a niveles elevados de contaminantes atmosféricos puede tener efectos negativos en la salud, especialmente en las personas más vulnerables, como niños, ancianos y personas con enfermedades respiratorias o cardiovasculares.
La Comunidad de Madrid ha implementado diversas medidas para mejorar la calidad del aire, como la creación de zonas de bajas emisiones, la promoción del uso del transporte público y

de vehículos menos contaminantes, la regulación de las emisiones de las industrias y la promoción de la eficiencia energética en los edificios.

En conclusión, la calidad del aire es un tema relevante en la Comunidad de Madrid debido a sus efectos sobre la salud de la población. Es fundamental seguir monitoreando los niveles de contaminación y tomar medidas para reducirlos y mejorar la calidad del aire en la ciudad.

9. Impactos sociales y medioambientales del proyecto

Realizar un cuadro de mandos para el análisis de la calidad del aire puede tener un impacto socio ambiental significativo.

En cuanto al impacto ambiental, el cuadro de mandos puede ayudar a detectar y monitorear los niveles de contaminación en el aire, lo que puede ser una herramienta valiosa para la protección y conservación del medio ambiente. Esto es especialmente importante dado el impacto negativo que la contaminación del aire puede tener en la salud de las personas, la fauna y la flora, así como en la calidad del agua y la tierra.

En términos de impacto social, el cuadro de mandos puede ayudar a aumentar la conciencia y la comprensión de la importancia de la calidad del aire entre el público en general. También puede ser utilizado por gobiernos y organizaciones para tomar decisiones informadas sobre políticas públicas y medidas de protección ambiental que puedan mejorar la calidad de vida de las personas y el bienestar de la sociedad en general.

Sin embargo, es importante tener en cuenta que la implementación de un cuadro de mandos para el análisis de la calidad del aire también puede tener algunos efectos negativos, especialmente si no se lleva a cabo de manera adecuada. Por ejemplo, si los datos recopilados y analizados no son precisos o confiables, se pueden tomar decisiones incorrectas o ineficaces que pueden empeorar la situación ambiental. Además, la recolección y análisis de datos puede requerir recursos y tecnología que pueden generar emisiones adicionales y contribuir al cambio climático si no se manejan de manera adecuada

Las ventajas socioambientales de tener un cuadro de mandos para el análisis de calidad del aire son las siguientes:

Protección de la salud humana: La calidad del aire tiene un impacto directo en la salud humana. Un cuadro de mandos para el análisis de la calidad del aire puede ayudar a detectar y monitorear los niveles de contaminación en el aire, lo que puede ser una herramienta valiosa para la protección de la salud humana.
Conservación del medio ambiente: La contaminación del aire puede tener un impacto negativo en el medio ambiente, incluyendo la fauna y la flora, así como en la calidad del agua y la tierra. Un cuadro de mandos para el análisis de la calidad del aire puede ayudar a detectar y monitorear la contaminación del aire y, por lo tanto, contribuir a la conservación del medio ambiente.

Toma de decisiones informadas: Un cuadro de mandos para el análisis de la calidad del aire puede proporcionar datos precisos y confiables que pueden ser utilizados por gobiernos y organizaciones para tomar decisiones informadas sobre políticas públicas y medidas de

protección ambiental que puedan mejorar la calidad de vida de las personas y el bienestar de la sociedad en general.
Concientización y educación: La implementación de un cuadro de mandos para el análisis de la calidad del aire puede aumentar la conciencia y la comprensión de la importancia de la calidad del aire entre el público en general, lo que puede ser beneficioso para fomentar un cambio positivo en las actitudes y comportamientos de la sociedad hacia la protección ambiental.

Reducción de emisiones contaminantes: Un cuadro de mandos para el análisis de la calidad del aire puede proporcionar información valiosa sobre las fuentes de contaminación del aire, lo que puede ser utilizado para desarrollar estrategias para reducir las emisiones de contaminantes y, por lo tanto, mejorar la calidad del aire.

10. Conclusión

10.1. Resumen de aportaciones del trabajo

Nuestro objetivo principal ha sido ofrecer al ciudadano de manera más clara el comportamiento de los agentes contaminantes, efectivamente hemos cumplido dicho objetivo, ya que hemos podido analizar de forma clara toda la información que tenemos en la web por medio de csv y json y de forma que cualquier persona pueda entender la información y comprender la situación de la calidad del aire en nuestra comunidad.

Hemos sido capaces de monitorear la contaminación del aire, aportando claridad a la información, y representarla en tablas y cronogramas, para así ser conscientes de la situación anterior y la presente.

Además, hemos representado las principales fuentes de contaminación para dar visibilidad a las causas y consecuencias de la contaminación.

10.2. Vías de continuación

Dentro de este apartado definiremos las vías de continuación del proyecto realizado.

En primer lugar, sería importante añadir una automatización inicial al proyecto para que cada inicio de día obtuviera la información vía API de la web de la Comunidad de Madrid del día anterior.
Para ello, se tendría que configurar un orquestador de eventos, el cúal, a través de un proceso realizaría la conexión, la obtención de datos desde la API de la web para posteriormente realizar una actualización de los mismos en el proyecto diariamente.
Se podría crear un proceso, el cual, se conectaría a la API de Microsoft y actualizaría los datos del informe de manera automática

La siguiente vía de continuación sería aplicar el proceso de este proyecto en los datos de todas las comunidades autónomas de España, de esta forma, tendríamos que ampliar el cuadro de mandos para visualizar la información de todas las comunidades autónomas.
Y posteriormente, obtener toda la información del resto de países.

Otra de las vías de continuación de proyecto sería la compra de la licencia “Embedded” de PowerBI, la cuál nos proporciona la opción de publicar el cuadro de mandos en una página web y/o aplicaciones, gracias a ello el acceso al cuadro de mandos sería libre.
La licencia mencionada es a nivel de capacidad, es decir, se cobraría por consumo y no por usuario que haga uso del informe.

Como idea para el futuro, se podría contemplar la posibilidad de implantar un sistema que recoja los datos directamente de los puntos de medición para tener la información actualizada a tiempo real en el sistema, y no tener que realizar conexiones a los datos de las diferentes comunidades autónomas y/o países.

Bibliografia

[22/05/2022] - Herramienta Spoon

[10/09/2022] - Hitachi(Pentaho)

[15/07/2022] - Power BI

[01/10/2022] - Información sobre Data Warehouse

[2021/2022] - Material proporcionado por el profesor Pedro Pablo Alarcon de la Universidad Politécnica de Madrid en la asignatura Sistemas de Soporte a la toma de decisiones.

[01/02/2023] - Microsoft SQL Server

[01/11/2021] - Datos abiertos de la Comunidad de Madrid

[2013/2020] - ESTRATEGIA DE CALIDAD DEL AIRE Y CAMBIO CLIMÁTICO DE LA COMUNIDAD DE MADRID

[2017/2018] - ESTUDIO PARA LA CUANTIFICACIÓN DE LA CONTRIBUCIÓN DE FUENTES A LOS NIVELES DE CALIDAD DEL AIRE EN EL MUNICIPIO DE MADRID

[2020/2021] - Estrategia de Calidad del Aire -EEA