“Análisis de datos - Obra Social de la Universidad Nacional del Litoral ”

Facultad de Ingeniería

Universidad Nacional de Entre Ríos

Tecnicatura Universitaria en Procesamiento y Explotación de Datos.

Docente: Walter Ricardo ELIAS

Autor: Mauricio Ruben Goette

Coordinador de la Empresa: Gastón Monzón

Índice

Introducción

En diversos campos como los negocios, la medicina y la ciencia, la información histórica desempeña un papel fundamental para comprender el pasado, interpretar el presente y anticipar el futuro. No obstante, los datos por sí solos poseen un valor relativo en estos contextos. Lo que resulta verdaderamente interesante es el conocimiento que puede ser extraído a partir de los datos y, aún más importante, la capacidad de utilizar ese conocimiento de manera efectiva.

La minería de datos se centra en analizar los datos con el objetivo de extraer conocimiento significativo. Este conocimiento puede manifestarse en forma de relaciones, patrones o reglas que se derivan de los datos y que previamente eran desconocidos. También puede presentarse como una descripción más concisa o un resumen de la información analizada.

Los modelos predictivos se enfocan en estimar valores futuros o desconocidos de variables de interés, denominadas variables objetivo o dependientes. En contraste, los modelos descriptivos tienen como objetivo identificar patrones que explican o resumen los datos analizados. Estos modelos son valiosos para explorar las propiedades de los datos, pero no están diseñados para predecir nuevos datos.

En este sentido, tanto la minería de datos como la utilización de modelos predictivos y descriptivos se convierten en herramientas esenciales para aprovechar al máximo la información disponible y tomar decisiones fundamentadas en diversos campos de estudio.

Situación Problemática

INTEGRAL SOFTWARE SRL es una empresa comprometida con el diseño, desarrollo, integración e implementación de soluciones informáticas y tecnológicas, cuyo objetivo principal es crear un equilibrio en cada proyecto, combinando estrategia, creatividad, tecnología y, sobre todo, la convicción de brindar un servicio profesional de alta calidad.

Dicha empresa se dedica a satisfacer las necesidades y requerimientos de los clientes, quienes confían en recibir productos y soluciones adaptadas a sus necesidades. La principal misión es contribuir al éxito de los clientes, permitiéndoles aprovechar las oportunidades que ofrece la innovación tecnológica. Colaborando estrechamente para entender sus objetivos y desafíos específicos, se utiliza el conocimiento y experiencia para desarrollar soluciones a medida, que impulsen el crecimiento y la eficiencia de sus negocios.

La empresa Integral Software SRL enfrenta un desafío en la forma en que brinda información generada a partir de bases de datos a sus clientes. Actualmente, la información se entrega en formato crudo, lo que requiere que los usuarios realicen ellos mismos tareas como filtrar los datos, seleccionar variables y crear gráficos para visualizar en el correspondiente dashboard de cada usuario.

Planteamiento del problema

Al analizar dicha situación en el que se encuentra la empresa, nos encontramos con ciertos desafíos ya que la falta de una entrega de información procesada y correctamente visualizada podría generar insatisfacción entre los clientes que esperan obtener resultados listos para su análisis y realizar la toma de decisiones.

¿Existe cierta ineficiencia en la utilización de la información disponible? ya que al requerir que los usuarios realicen tareas de análisis de datos por sí mismos y posibles errores y la falta de uniformidad puede generar una pérdida de tiempo y recursos, lo que podría limitar el aprovechamiento efectivo de la información generada.

¿Qué herramientas de procesamiento y explotación de datos podrían utilizarse para filtrar datos, seleccionar variables y crear gráficos para generar conocimiento útil a la empresa en Paraná, 2023?

Objetivos

Para abordar esta situación, la empresa ha establecido objetivos específicos:

- Seleccionar variables e indicadores de interés para cada cliente
- Analizar la homogeneidad de los datos y detectar incongruencias, datos faltantes o erróneos
- Generar gráficos adecuados a las variables de interés

Con la implementación de estos objetivos, Integral Software SRL busca mejorar la experiencia del cliente al brindarles información procesada y visualizada de manera más

intuitiva y lista para su uso. Esto permitirá a los usuarios enfocarse en la interpretación de los datos y en la toma de decisiones estratégicas, optimizando su tiempo y recursos.

Hipótesis

Podríamos plantear en forma de hipótesis que llevada a cabo la implementación de herramientas de Business Intelligence (BI) y la automatización de procesos para procesar y visualizar datos de manera automática mejorará la eficiencia en la utilización de la información disponible para los clientes.

Al seleccionar variables e indicadores de interés para cada cliente y analizar la homogeneidad de los datos, se reducirán los posibles errores y la falta de uniformidad en la información presentada, permitiendo la implementación de dashboards personalizados que permitan a los clientes acceder a información procesada y visualizada de manera clara y comprensible facilitando el uso de la información por parte de usuarios sin conocimientos técnicos, lo que mejorará su capacidad para obtener insights y tomar decisiones informadas.

Marco teórico

En el marco teórico, se presentan las bases conceptuales y fundamentos que sustentan las decisiones y enfoques tomados para abordar el problema planteado por Integral Software SRL. Al incluir conceptos clave relacionados con el procesamiento y explotación de datos, herramientas de Business Intelligence, Data Warehousing y visualización de datos, entre otros justificando la importancia de la implementación de estas tecnologías y metodologías.

Fuente de Datos

En la fase de integración y recopilación de datos, se determinan las fuentes de información útiles y se identifica dónde se obtuvo. Luego, todos los datos se transforman a un formato común, utilizando un almacén de datos que permite unificar de manera operativa toda la información recopilada, detectando y resolviendo las inconsistencias.

Selección, limpieza y transformación

La recopilación de datos debe ir acompañada de una limpieza e integración de los mismos, para que éstos estén en condiciones para su análisis. Los beneficios del análisis y de la extracción de conocimiento a partir de datos dependen, en gran medida, de la calidad de los datos recopilados. Gran parte del éxito de un proceso de minería de datos depende, no sólo de tener todos los datos necesarios (una buena recopilación), sino de que éstos estén íntegros, completos y consistentes (una buena limpieza e integración).

En la fase de selección, limpieza y transformación, se eliminan o corrigen los datos incorrectos y se decide la estrategia a seguir con los datos incompletos. Además, se proyectan los datos para considerar únicamente aquellas variables o atributos que van a ser relevantes, con el objetivo de hacer más fácil la tarea propia de minería y para que los resultados de la misma sean más útiles.

Conjuntamente, la preparación de datos tiene como objetivo la eliminación del mayor número posible de datos erróneos o inconsistentes (limpieza) e irrelevantes (criba), y trata de presentar los datos de la manera más apropiada para la minería de datos.

Bases de datos relacionales

Una base de datos relacional es una colección de relaciones (tablas). Cada tabla consta de un conjunto de atributos (columnas o campos) y puede contener un gran número de tuplas (registros o filas). Cada tupla representa un objeto, el cual se describe a través de los valores de sus atributos y se caracteriza por poseer una clave única o primaria que lo identifica. Una de las principales características de las bases de datos relacionales es la existencia de un esquema asociado, es decir, los datos deben seguir una estructura y son, por tanto, estructurados.

Son la fuente de datos para la mayoría de aplicaciones de minería de datos, muchas técnicas de minería de datos no son capaces de trabajar con toda la base de datos, sino que sólo son capaces de tratar con una sola tabla a la vez. Lógicamente, mediante una consulta (por ejemplo en SQL, en una base de datos relacional tradicional, o con herramientas y operadores más potentes, en los almacenes de datos) podemos combinar en una sola tabla o vista minable aquella información de varias tablas que se requieren para cada tarea concreta de minería de datos. Por tanto, la presentación tabular, también llamada atributo-valor, es la más utilizada por las técnicas de minería de datos.

Bases de datos multidimensionales

Las bases de datos multidimensionales son sistemas de almacenamiento de datos que están diseñados para trabajar con modelos multidimensionales, donde los datos se organizan en múltiples dimensiones y se pueden analizar desde diferentes perspectivas.

En una base de datos multidimensional, los datos se representan en forma de cubos o hipercubos, donde cada dimensión representa una característica o atributo de los datos. Por ejemplo, en una base de datos de ventas, las dimensiones pueden ser el tiempo, el producto y la ubicación. Las bases de datos multidimensionales permiten realizar análisis complejos y consultas OLAP (Online Analytical Processing) de manera eficiente.

Las bases de datos multidimensionales (BDM) son un tipo de base de datos optimizada para Data Warehouse que se utilizan principalmente para crear aplicaciones OLAP, una tecnología asociada al acceso y análisis de datos en línea.
A diferencia del modelo relacional, el modelo de datos más extendido-donde la información se almacena a través de campos y registros, las bases de datos multidimensionales se caracterizan por los siguientes atributos:

- Se basan en la creación de aplicaciones OLAP y pueden verse como bases de datos contenidos en una sola tabla.
- En las tablas multivaluadas se almacenan registros referidos bien a las dimensiones de la misma o a las métricas que se desean analizar, adoptando un campo o columna por cada dimensión y otro campo por cada métrica o hecho.

- Las tablas del modelo multidimensional se asimilan a un hipercubo o, si usamos herramientas OLAP, a un cubo OLAP. En ambos casos, las dimensiones de los cubos se corresponden con la de la tabla y el valor almacenado en cada celda equivale al de la métrica.

Datawarehouse

Por otro lado, un data warehouse (almacén de datos) en el marco de las bases de datos multidimensionales es un repositorio centralizado que integra datos de diversas fuentes operativas de una organización. El propósito principal de un data warehouse es proporcionar una vista consolidada e histórica de los datos de la organización, lo que facilita el análisis de tendencias, la generación de informes y la toma de decisiones basada en datos.

Un almacén de datos es un tipo de sistema de gestión de datos diseñado para tareas analíticas y de toma de decisiones en Business Intelligence (BI). Los DW suelen contener grandes cantidades de datos históricos y a menudo, la información almacenada proviene de una amplia gama de fuentes, como los archivos de registro de aplicaciones o las aplicaciones de transacciones, esto permite a los usuarios analizar datos que antes estaban dispersos y eran difíciles de acceder.

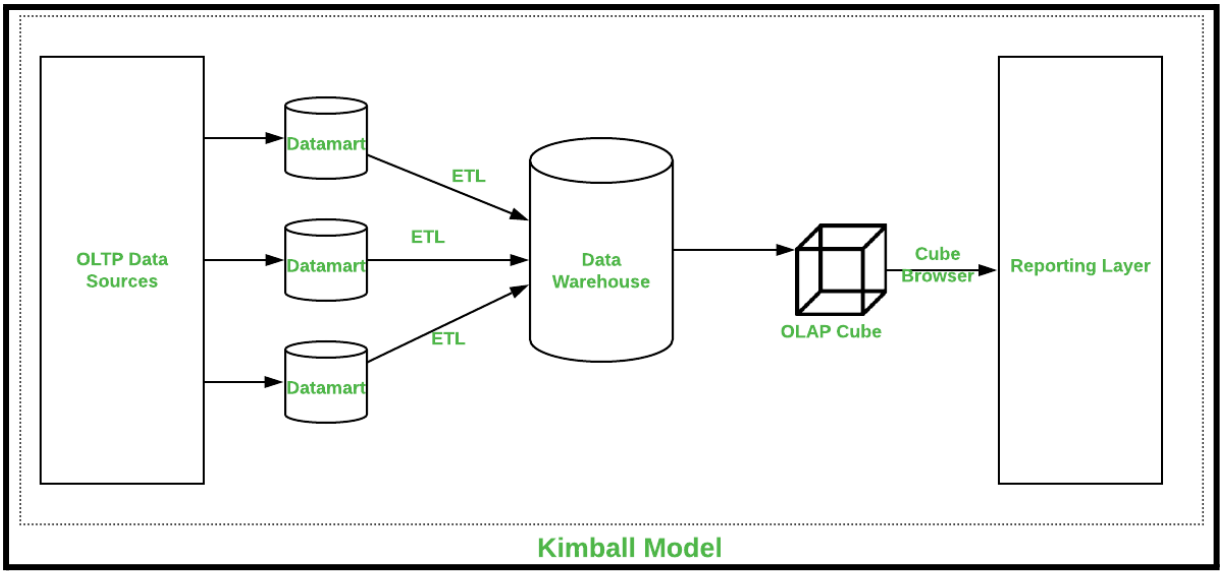
Al tener información qué proviene de diferentes fuentes de datos es necesario hacer una integración, esta se realiza mediante procesos de ETL(extracción, transformación y carga), integrando las diferentes fuentes y dándoles un formato para posteriormente cargarlos en la base de datos.

Diseño datawarehouse

Diseñar un almacén de datos (Data Warehouse) es una parte esencial del desarrollo empresarial. Para el diseño, existen dos arquitecturas más comunes llamadas Kimball e Inmon, en este caso en particular se elige la opción de la metodología Kimball, ya que este enfoque comienza cumple con el proceso de negocio y las preguntas que el Data Warehouse debe responder.

Aplicaciones de la metodología Kimball:

- La configuración y construcción son rápidas.
- Generar informes a partir de múltiples esquemas estrella es muy exitoso.
- Las operaciones de base de datos son muy efectivas.
- Ocupa menos espacio en la base de datos y su administración es fácil.



Al utilizar el modelo de datos de Kimball seguimos un enfoque de abajo hacia arriba para almacenamiento de datos diseño de arquitectura en el que los data marts se forman primero en función de los requisitos comerciales.

Algunos de los principales beneficios del concepto de almacenamiento de datos de Kimball incluyen:

- El modelado dimensional de Kimball es rápido de construir ya que no implica normalización, lo que significa una rápida ejecución de la fase inicial del almacenamiento de datos de procesos.
- Una ventaja del esquema en estrella es que la mayoría de los operadores de datos pueden comprenderlo fácilmente debido a su estructura desnormalizada, que simplifica las consultas y el análisis.
- La huella del sistema de almacenamiento de datos es trivial porque se centra en áreas y procesos comerciales individuales en lugar de en toda la empresa. Por lo tanto, ocupa menos espacio en la base de datos, lo que simplifica la administración del sistema.
- Permite la recuperación rápida de datos del almacén de datos, ya que los datos se segregan en tablas de hechos y dimensiones.
- Estructura dimensional conformada en un marco de calidad de datos. El enfoque de Kimball para el ciclo de vida del almacén de datos también se conoce como el enfoque de estilo de vida dimensional empresarial porque permite que las herramientas de inteligencia empresarial profundicen en varios esquemas en estrella y generen información confiable.

Esquema estrella

El esquema en estrella es el elemento fundamental en este modelo de almacén de datos dimensional. La combinación de una tabla de hechos con varias tablas dimensionales a menudo se denomina esquema en estrella. El modelado dimensional de Kimball permite a los usuarios construir varios esquemas en estrella para satisfacer diversas necesidades de

generación de informes. La ventaja del esquema en estrella es que las consultas de tablas dimensionales pequeñas se ejecutan instantáneamente.

Cada dimensión tiene una estructura jerárquica pero no necesariamente lineal. Esto permite diferentes niveles y caminos de agregación para las diferentes dimensiones, posibilitando la definición de hechos agregados con mucha facilidad. La forma elegida para estos conjuntos

de hechos y sus dimensiones es la de “estrella simple” (cuando no hay caminos alternativos en las dimensiones).

Luego de elegir la metodología a utilizar necesitamos transformar los datos en un esquema estrella, para ello contemplamos los siguientes pasos:

- Identificar el proceso de negocio: El primer paso consiste en identificar el proceso de negocio que se desea analizar. Esto ayuda a determinar la tabla de hechos (fact table) y las tablas de dimensiones correspondientes.
- Identificar la tabla de hechos: La tabla de hechos contiene las medidas o métricas que se desean analizar. Por lo general, estas medidas son datos numéricos, como ventas, ingresos o cantidad de unidades vendidas.
- Identificar las tablas de dimensiones: Las tablas de dimensiones contienen los datos descriptivos que proporcionan contexto a las medidas de la tabla de hechos. Estas tablas suelen incluir información sobre dimensiones como cliente, producto, tiempo, ubicación, etc.
- Normalizar los datos: Es importante normalizar los datos para evitar redundancia y garantizar la integridad de la información. Esto implica organizar los datos en tablas separadas, cada una centrada en un tema específico y evitando la duplicación de información.
- Crear la tabla de hechos: La tabla de hechos se crea combinando las claves de las tablas de dimensiones y las medidas correspondientes. Cada fila de la tabla de hechos representa una combinación única de dimensiones y contiene los valores de las medidas asociadas.
- Crear las tablas de dimensiones: Las tablas de dimensiones se crean extrayendo los atributos descriptivos de las tablas normalizadas. Cada tabla de dimensiones contiene una lista única de valores para una dimensión específica.
- Agregar claves sustitutas: Se agregan claves sustitutas a cada tabla para simplificar las operaciones de unión entre las tablas. Estas claves sustitutas son identificadores únicos asignados a cada fila de datos.
- Establecer relaciones: Se establecen relaciones entre la tabla de hechos y las tablas de dimensiones utilizando las claves primarias y foráneas correspondientes. Esto permite realizar análisis multidimensional y agregar medidas por atributos de las dimensiones.
- Crear índices: Se pueden crear índices en las claves de las tablas para mejorar el rendimiento de las consultas y acelerar el acceso a los datos.
- Probar el esquema: Finalmente, se realiza una prueba del esquema para verificar su correcto funcionamiento. Se ejecutan consultas y se verifican los resultados para asegurarse de que la estructura de datos sea efectiva para el análisis y la generación de informes.

Métricas e Indicadores (KPI)

El término KPI, siglas en inglés, de Key Performance Indicator, cuyo significado en castellano vendría a ser Indicador Clave de Desempeño o Medidor de Desempeño, hace referencia a una serie de métricas que se utilizan para sintetizar la información sobre la eficacia y productividad de las acciones que se lleven a cabo en un negocio con el fin de poder tomar decisiones y determinar aquellas que han sido más efectivas a la hora de cumplir con los objetivos marcados en un proceso o proyecto concreto.

Los KPI también son conocidos como indicadores de calidad o indicadores clave de negocio que pueden ser utilizados y aplicables en cualquier área de negocio y sector productivo, aunque son utilizados de una forma muy habitual en el marketing online.

El objetivo último de un KPI es ayudar a tomar mejores decisiones respecto al estado actual de un proceso, proyecto, estrategia o campaña y de esta forma, poder definir una línea de acción futura.

Dashboard

Un dashboard es una herramienta de gestión de la información que monitoriza, analiza y muestra de manera visual los indicadores clave de desempeño (KPI), métricas y datos fundamentales para hacer un seguimiento del estado de una empresa, un departamento, una campaña o un proceso específico.

Podemos pensar en el dashboard como una especie de "resumen" que recopila datos de diferentes fuentes en un solo sitio y los presenta de manera digerible para que lo más importante salte a la vista. Estas son algunas de las características que debe tener este centro de control:

- Personalizado. Un dashboard debe contener únicamente los KPI que sean relevantes para el departamento, campaña o proceso que nos ocupa. Para orientarlo, podemos pensar en las preguntas principales a las que queremos responder. Por ejemplo, cuáles son las principales fuentes de tráfico a nuestra web, cómo está funcionando nuestro embudo de ventas o cuáles son los 5 productos que nos generan más ingresos.

- Visual. La idea de un dashboard es que podamos obtener la información que buscamos a golpe de vista. Por ello, los datos se presentan en forma de gráficos y debemos contar con indicadores rápidos a través de claves de color, flechas hacia arriba o abajo o cifras destacadas, por ejemplo.

- Práctico. La función principal de un dashboard siempre debe ser orientar las acciones de nuestro equipo. Por tanto, debe facilitarnos la información necesaria para que podamos saber cuáles son los siguientes pasos a seguir para mejorar los resultados.

- En tiempo real. A día de hoy, las acciones de marketing digital evolucionan con gran rapidez y aprovechar el momento clave es esencial. Por eso, la

información debería estar actualizada al momento en todas las fuentes y mostrarse en el dashboard en tiempo real.

Objetivos

Para abordar esta situación, la empresa ha establecido objetivos específicos:

1. Seleccionar variables e indicadores de interés para cada cliente:

Se busca identificar las variables relevantes para cada cliente, considerando sus necesidades y requerimientos particulares como también la búsqueda de indicadores o KPIs(indicador clave de rendimiento) que se realizará en función de los objetivos y metas específicas de cada cliente, de manera que se puedan monitorear y evaluar de forma efectiva los resultados. Esto permitirá personalizar la información y brindarles datos significativos para la toma de decisiones.

2. Analizar la homogeneidad de los datos y detectar incongruencias, datos faltantes o erróneos:

Es importante garantizar la calidad de los datos proporcionados a los clientes. Se realizará un análisis exhaustivo para verificar la coherencia de los datos, identificar posibles inconsistencias y corregir errores, asegurando la fiabilidad de la información entregada.

3. Generar gráficos adecuados a las variables de interés:

Con base en las variables seleccionadas, se crearán gráficos relevantes y claros. Estos gráficos se integrarán en un dashboard completo y listo para presentar al usuario final. El objetivo es proporcionar una visualización efectiva de los datos, facilitando su comprensión y análisis.

Con la implementación de estos objetivos, Integral Software SRL busca mejorar la experiencia del cliente al brindarles información procesada y visualizada de manera más intuitiva y lista para su uso. Como nota, es importante considerar que al analizar las necesidades de la organización y definir el problema, debemos tener en cuenta que el proceso de extracción de conocimiento es iterativo e interactivo.

Es iterativo en el sentido de que el avance en una fase determinada puede requerir volver a etapas anteriores, ya que el análisis de datos y la comprensión del problema pueden evolucionar a medida que se obtiene más información. Además, es común que sean necesarias varias iteraciones para obtener un conocimiento de alta calidad y precisión.

Por otro lado, el proceso es interactivo, lo que implica que el usuario o un experto en el dominio del problema desempeña un papel fundamental. Su participación es requerida para ayudar en la preparación de los datos, validar el conocimiento extraído y ofrecer información relevante que permita mejorar el análisis.

Esta interacción entre el experto y el análisis de datos es fundamental para garantizar que el conocimiento extraído sea preciso, relevante y útil para la toma de decisiones. La retroalimentación constante y la colaboración entre el experto y el proceso de extracción de conocimiento contribuyen a obtener resultados más sólidos y confiables.

Metodología - Desarrollo

Obtención de Datos

El conjunto de datos brindados para el análisis contiene información de seguros de salud que incluye datos sobre los pacientes, sus prestaciones, los profesionales intermediarios y los procedimientos médicos realizados. Cada fila representa una prestación realizada por un paciente para un procedimiento médico específico.

Variables del dataset

Las columnas del dataset suministrado son las siguientes:

- Afiliado: id del paciente
- EDAD: edad del paciente
- convenio prestacion: Código para el tipo de acuerdo de seguro de salud
- Fecha Emisión: fecha de reclamo
- Profesional Efector: ID del profesional que realizó el procedimiento médico
- Profesional Solicitante: ID del profesional que solicitó el procedimiento médico
- convenio facturación: Código para el tipo de acuerdo de seguro de salud utilizado para facturación
- Orden: Número de orden del procedimiento médico
- Prestación: Código del procedimiento médico realizado
- Estado_prestacion: Estado del procedimiento médico
- tipo_componente: Código para el tipo de procedimiento médico
- Grupo_prestaciones: Código para el grupo de procedimientos médicos
- Capitulo_prestaciones: Código del capítulo de procedimientos médicos

La empresa en cuestión ha proporcionado los datos en formato CSV, que contienen principalmente números utilizados como categorías en lugar de información descriptiva. Esto se debe a motivos de confidencialidad, ya que se ha eliminado cualquier dato personal identificable de los pacientes, como sus nombres o información médica detallada. En cambio, la información de los pacientes se ha reducido únicamente a su número de afiliado, que se utiliza para identificar de manera única a cada paciente, y a su edad en el momento en que se solicitó la prestación de servicios. Esta medida garantiza la protección de la privacidad de los pacientes y cumple con las normas de confidencialidad y protección de datos.

Selección, limpieza y transformación

En el proyecto actual, se ha desarrollado la fase de selección, limpieza y transformación de los datos utilizando el software Pentaho Data Integration. Esta herramienta permite realizar diferentes tareas de extracción, transformación y carga de datos de manera muy eficiente.

Sin embargo, como parte del proceso de garantizar la calidad de los datos, se realiza un análisis previo utilizando las bibliotecas de Python como pandas, numpy y matplotlib. Estas bibliotecas proporcionan funcionalidades avanzadas para trabajar con datos, realizar operaciones de limpieza y detección de posibles errores o inconsistencias en los datos.

Valores faltantes

Una primera inspección del dataset nos muestra las columnas con la respectiva cantidad de filas en cada una de estas columnas, podemos determinar que ya tenemos valores faltantes al comparar el número de cada una de las columnas con el valor total de filas que es de “919964”. Dichas filas con valores faltantes serán eliminadas ya que no representan un número muy grande de datos perdidos, con la excepción de las columnas de los profesionales, donde le asignaremos un valor de “0” a los datos faltantes, tomando este valor como una categoria “GENERICA”

```
RangeIndex: 919964 entries, 0 to 919963
Data columns (total 13 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   Afiliado                919963 non-null  float64
 1   EDAD                    919962 non-null  float64
 2   convenio_prestacion     919964 non-null  int64
 3   Fecha_Emision           919964 non-null  object
 4   Profesional_efector     918947 non-null  object
 5   Profesional_solicitante 598313 non-null  object
 6   convenio_facturacion    919964 non-null  int64
 7   Orden                   919964 non-null  int64
 8   Prestacion              919963 non-null  object
 9   Estado_prestacion       919958 non-null  object
 10  tipo_componente         919964 non-null  int64
 11  Grupo_prestaciones      919964 non-null  int64
 12  Capitulo_prestaciones   919964 non-null  int64
dtypes: float64(2), int64(6), object(5)
memory usage: 91.2+ MB
```

Outliers

En la variable edad podemos observar como la edad cubre un rango desde 0 años hasta 7084 años de edad. En este caso al analizar las edades optamos por realizar un recorte de datos, donde fijamos los límites desde 0 años ya que en esta edad se encuentra incluidos los niños de pocos meses de edad y una edad máxima de 120 años que es una cifra más realista de la edad máxima a la que podría llegar un paciente.

```
Column Name :  EDAD
col_0   freq
EDAD
0.0     9269
1.0     8775
2.0     5765
3.0     5726
4.0     5448
...      ...
162.0      1
269.0      3
3663.0     1
5913.0     2
7084.0     2

[111 rows x 1 columns]
```

En cuanto al caso de las fechas utilizaremos la misma estrategia donde fijamos el rango entre la primera fecha del dataset y la última fecha que no sobrepasa la fecha actual al momento de realizar el proyecto.

```
Column Name :  Fecha_Emision
col_0                freq
Fecha_Emision
2017-05-01 00:00:00   117
2017-05-01 09:49:23     1
2017-05-01 09:50:58     1
2017-05-01 10:07:47     1
2017-05-01 10:08:46     1
...                   ...
2109-06-24 00:00:00     1
2201-12-18 00:00:00     3
5618-06-11 00:00:00     1
7860-09-04 00:00:00     2
9034-08-07 00:00:00     2

[294061 rows x 1 columns]
```

Valores erróneos

En cuanto a los profesionales podemos observar unas cuantas variables con valor categórico, dado que cada profesional posee un número de identificación se decide por transformar estos datos a un solo valor genérico con el valor de “0”

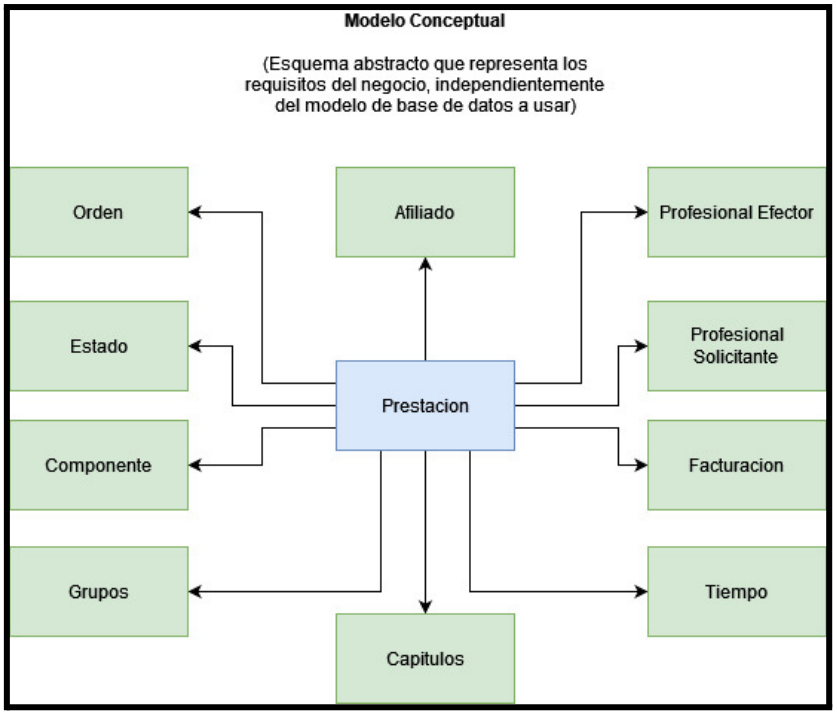
```
Column Name :  Profesional_efector
col_0                      freq
Profesional_efector
                              1
06287840                     34
10058055                    122
10061796                    136
10062607                    118
...                         ...
97935705                      5
99921116                    196
CD0002                       10
GENERICO                 117486
generico                      1
```

Al utilizar las bibliotecas de Python donde detectamos los posibles errores o inconsistencias mediante el análisis y visualización de los datos y combinamos esta información recolectada con el programa Pentaho Data Integration proporcionamos una solución integral para el proceso ETL, garantizando la calidad y confiabilidad de los datos utilizados en el proyecto.

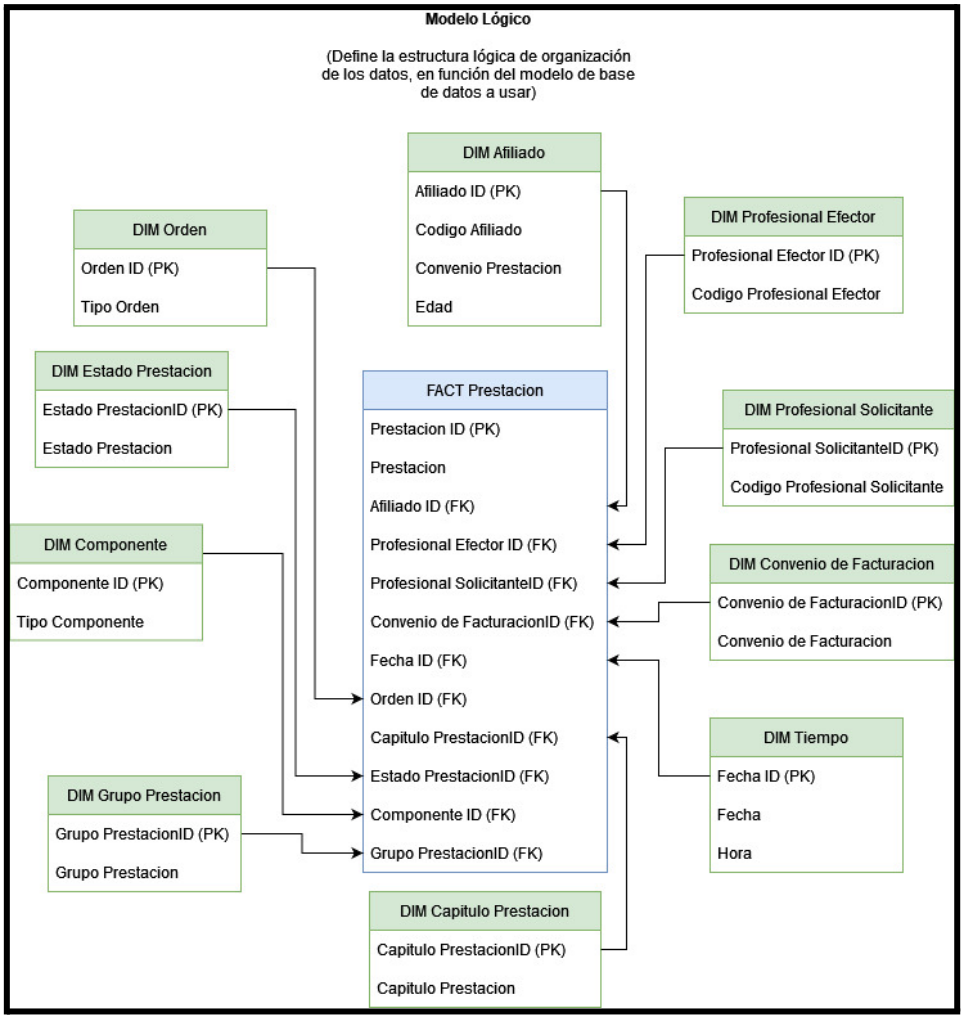
DataWarehouse

Diagrama datawarehouse

En el modelo multidimensional utilizado los datos se organizan en torno a los hechos, que tienen unos atributos o medidas que pueden verse en mayor o menor detalle según ciertas dimensiones. Esto nos permite, de una manera sencilla, obtener información sobre hechos a diferentes niveles de agregación.



Objetivo del modelo conceptual: Facilitar la comunicación entre los diseñadores y los usuarios sin conocimientos técnicos sobre implementación.



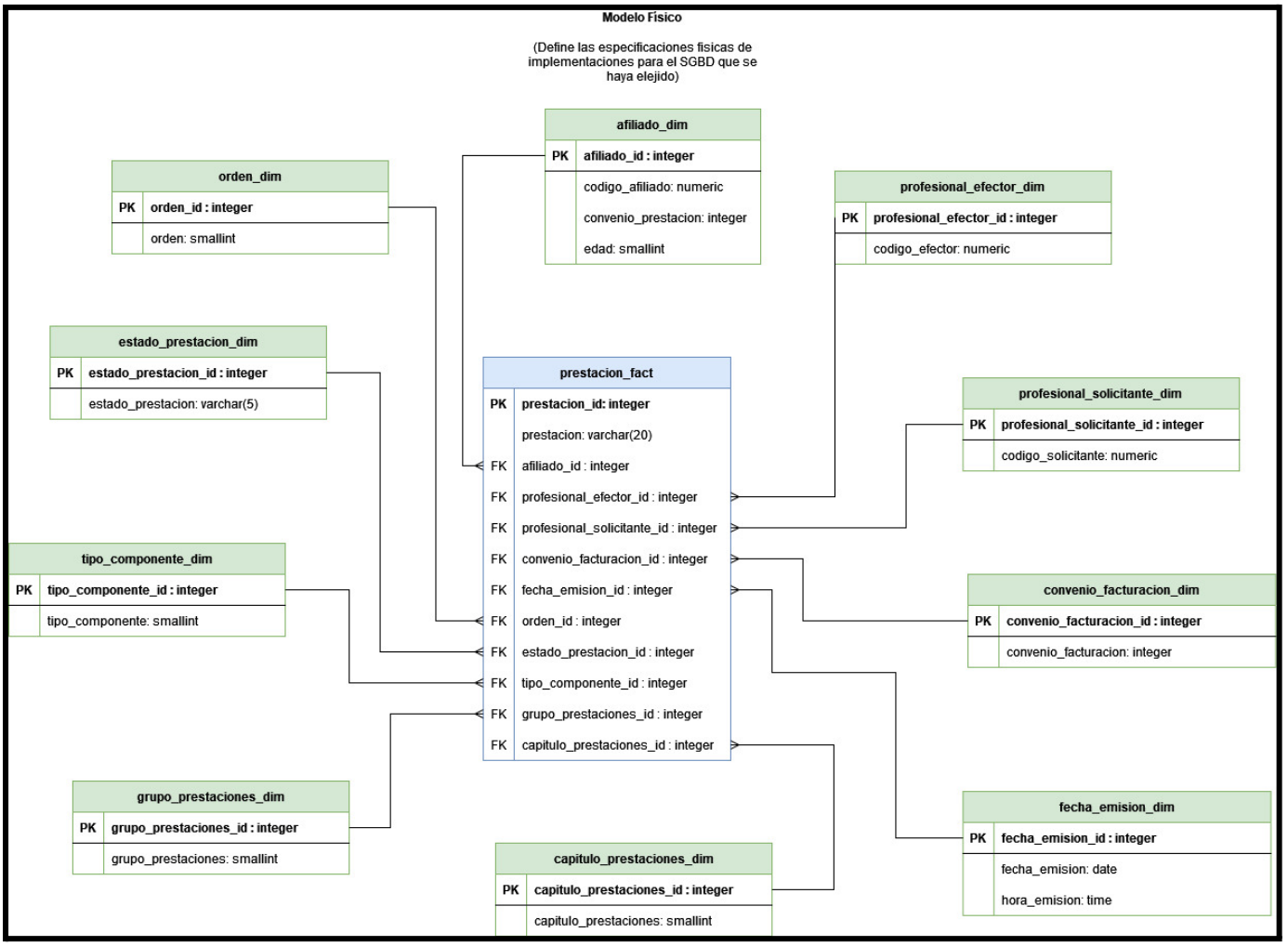


Tabla de hechos, mediciones, dimensiones y niveles

Hay que tener en cuenta que dentro del diseño de nuestro Data Warehouse, deberemos determinar la tabla de hechos, las dimensiones, los niveles y las mediciones, dado que estos son conceptos fundamentales que se utilizan para organizar y analizar los datos de manera estructurada.

La tabla de hechos es la tabla central en un esquema de diseño dimensional. Contiene las claves foráneas de las dimensiones y las mediciones cuantitativas que se analizan. Cada fila en la tabla de hechos representa una instancia o evento que se está analizando. Nuestra tabla de hechos representa por fila una prestación individual y única, que incluye las claves foráneas vinculando las dimensiones. La tabla de hechos está compuesta por:

Fact Table:

- prestacion_fact
    - prestacion_id (primary key)
    - prestacion
    - afiliado_id (foreign key)
    - profesional_efector_id (foreign key)

    - profesional_solicitante_id (foreign key)
    - convenio_facturacion_id (foreign key)
    - fecha_emision_id (foreign key)
    - orden_id (foreign key)
    - estado_prestacion_id (foreign key)
    - tipo_componente_id (foreign key)
    - grupo_prestaciones_id (foreign key)
    - capitulo_prestaciones_id (foreign key)

Teniendo en cuenta que las dimensiones representan las características o atributos descriptivos de los datos, estos elementos clave se pueden analizar, filtrar y agrupar en el Data Warehouse. Cada dimensión tiene sus propios miembros o valores únicos, que se organizan en jerarquías de niveles. El diseño actual de la base de datos implementada para los datos cuenta con 10 dimensiones las cuales son:

Dimension Tables:

- afiliado_dim
    - afiliado_id (primary key)
    - edad
    - Codigo Afiliado
    - convenio_prestacion
- profesional_efector_dim
    - profesional_efector_id (primary key)
    - codigo_efector
- profesional_solicitante_dim
    - profesional_solicitante_id (primary key)
    - codigo_solicitante
- convenio_facturacion_dim
    - convenio_facturacion_id (primary key)
    - convenio_facturacion
- fecha_emision_dim
    - fecha_emision_id (primary key)
    - fecha_emision
    - hora_emision
- orden_dim
    - orden_id (primary key)
    - orden
- estado_prestacion_dim
    - estado_prestacion_id (primary key)
    - estado_prestacion
- tipo_componente_dim
    - tipo_componente_id (primary key)
    - tipo_componente
- grupo_prestaciones_dim
    - grupo_prestaciones_id (primary key)
    - grupo_prestaciones
- capitulo_prestaciones_dim
    - capitulo_prestaciones_id (primary key)
    - capitulo_prestaciones

Los niveles son las diferentes granularidades o niveles de detalle dentro de una dimensión. Representan diferentes niveles de resumen o desglose de los datos. Por ejemplo, en una dimensión de tiempo, los niveles pueden incluir año, trimestre, mes, día, etc. En nuestro

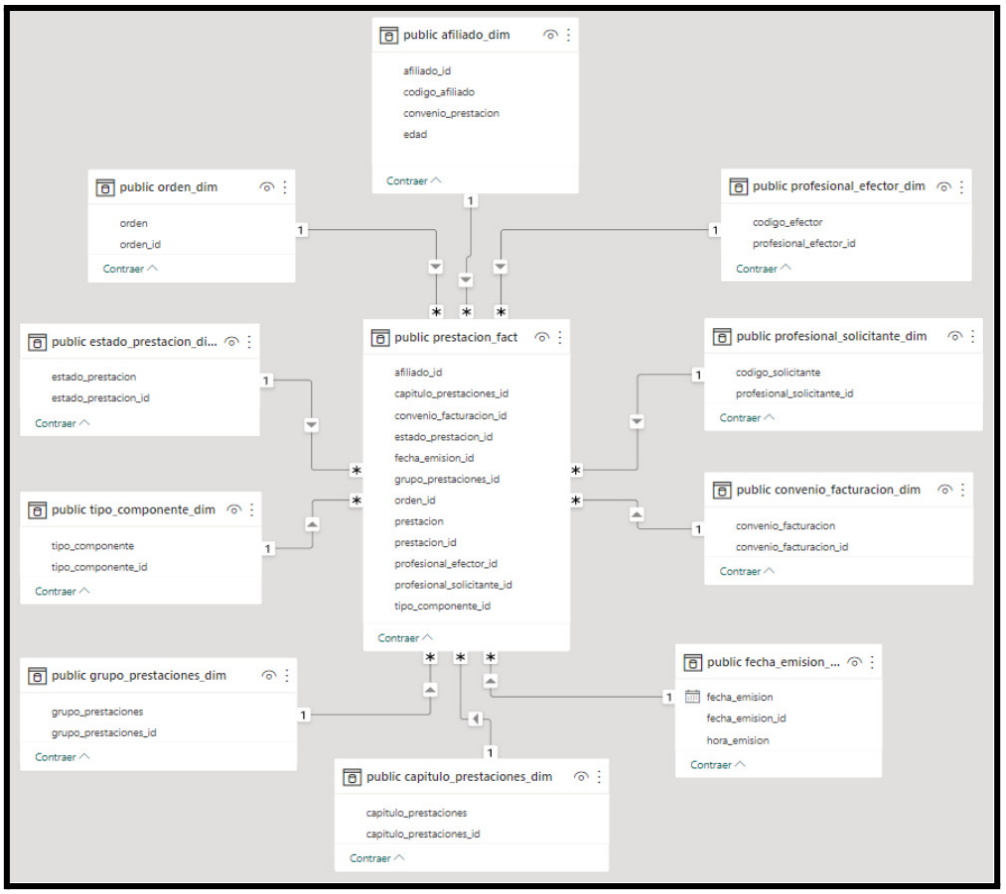
caso en particular al haber seleccionado un diseño de estrella simple, solamente contamos con un solo nivel de granularidad en nuestras dimensiones.

Las mediciones, también conocidas como medidas o métricas, son los valores numéricos que analizaremos en el Data Warehouse una vez cargados los datos. Representan los datos cuantitativos que se desean analizar calculado mediante operaciones de agregación, como suma, promedio, máximo, mínimo, etc.

En nuestro proyecto en particular dichas mediciones se realizan utilizando diferentes funciones y herramientas disponibles en la herramienta Power BI con la cual realizaremos nuestro dashboard. Power BI se encarga de Agregar columnas calculadas, crear nuevas medidas o funciones de resumen utilizando el lenguaje exclusivo de PowerBI llamado DAX, y también utilizando visualizaciones con agregaciones predefinidas que se realizan automáticamente en los datos subyacentes, como por ejemplo, los gráficos de barras, los gráficos de líneas y los gráficos circulares agregan automáticamente los datos según las categorías y proporcionan mediciones visuales.

En conjunto, la tabla de hechos, las dimensiones, los niveles y las mediciones permiten organizar y analizar los datos de manera significativa en nuestro Data Warehouse. Las dimensiones proporcionan el contexto y los atributos descriptivos, los niveles ofrecen diferentes niveles de detalle y los datos numéricos se analizan a través de las mediciones. Estos conceptos son fundamentales para la modelización y consulta de datos en un entorno de Data Warehouse.

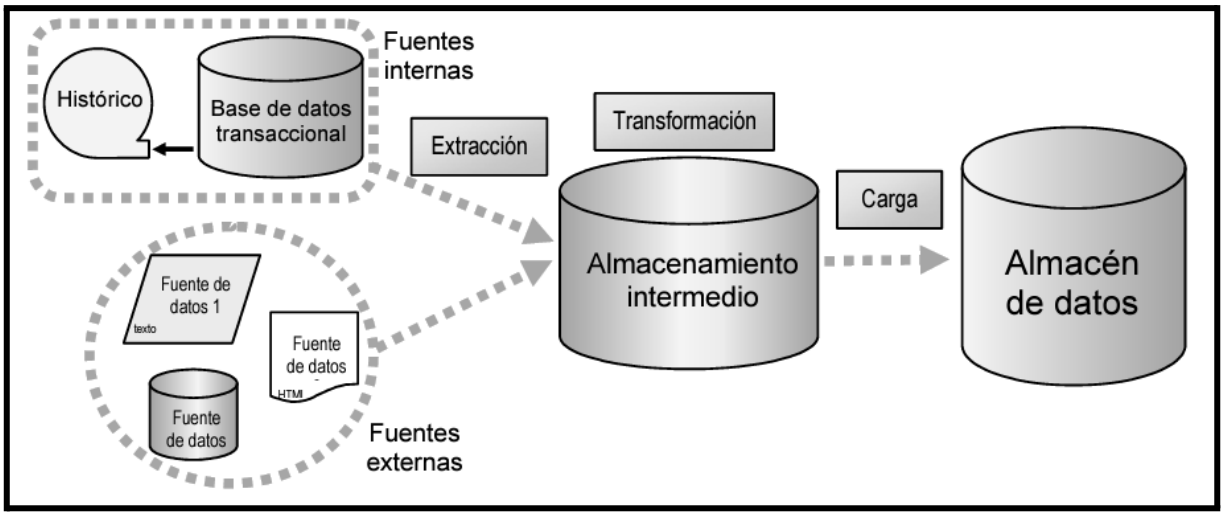
ERD(PowerBi):



ETL

El proceso ETL (Extract, Transform, Load) es una metodología utilizada en el ámbito de la gestión de datos para la integración y preparación de datos provenientes de diversas fuentes con el objetivo de cargarlos en un destino final, como un almacén de datos o un data warehouse.

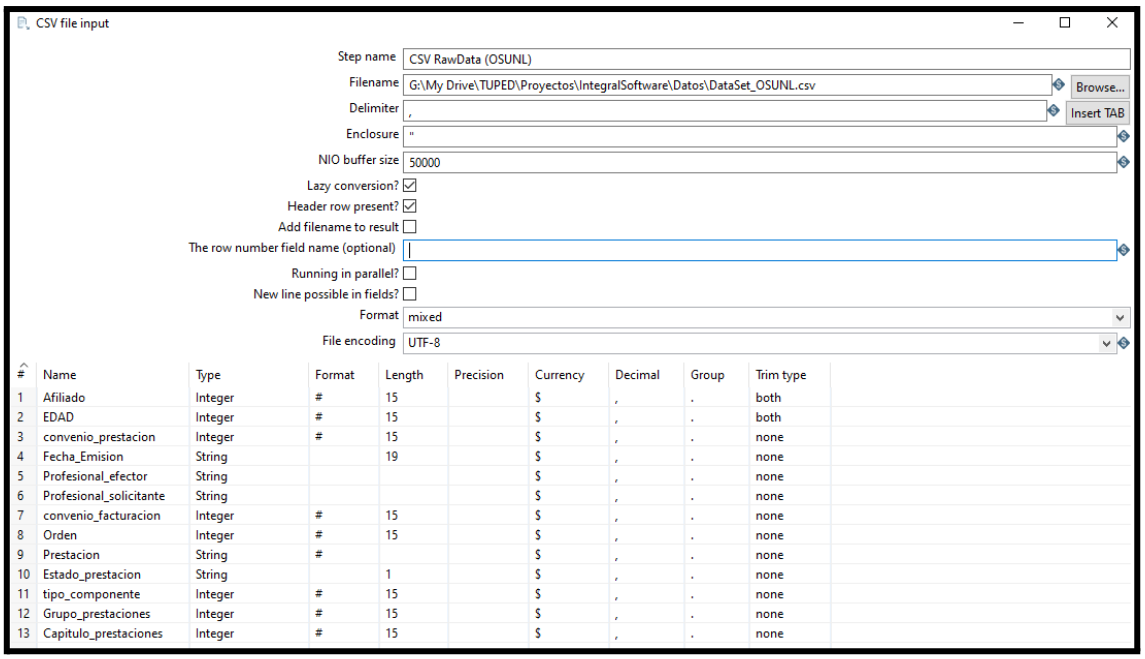
En el proyecto se utiliza la herramienta de ETL de código abierto Pentaho Data Integration(PDI). Esta nos permite tomar la fuente de los datos y cargarlos en un área de preparación antes de cargarlos en el servidor de base de datos relacional. Una vez que los datos se cargan en el área de ensayo del almacén de datos, la siguiente fase incluye la carga de datos en un modelo de almacén de datos dimensional que no está normalizado por naturaleza. Este modelo divide los datos en la tabla de hechos, que son datos transaccionales numéricos o tabla de dimensiones, que es la información de referencia que respalda los hechos.



Extracción (Extract)

En esta etapa, los datos se obtienen de diferentes fuentes de datos, como bases de datos, archivos CSV, APIs, sistemas externos, entre otros. El objetivo es extraer los datos necesarios para el análisis o almacenamiento posterior.

Nuestro proyecto en particular parte tomando la extracción de los datos de un archivo CSV brindado por la empresa , PDI nos permite especificar la ubicación del archivo, definir el delimitador utilizado para separar los valores y configurar varias opciones más para leer y procesar los datos. También podemos observar las columnas que serán cargadas como también su tipo de dato, formato, cantidad de caracteres entre otras cosas. Un detalle en la carga del archivo es seleccionar correctamente la codificación del formato del lenguaje, en este caso es UTF-8 que es compatible con ASCII, lo que significa que los caracteres ASCII (los primeros 128 caracteres en el conjunto Unicode) se pueden representar de la misma forma que en ASCII estándar, lo que garantiza la compatibilidad con sistemas y aplicaciones existentes que utilizan ASCII.

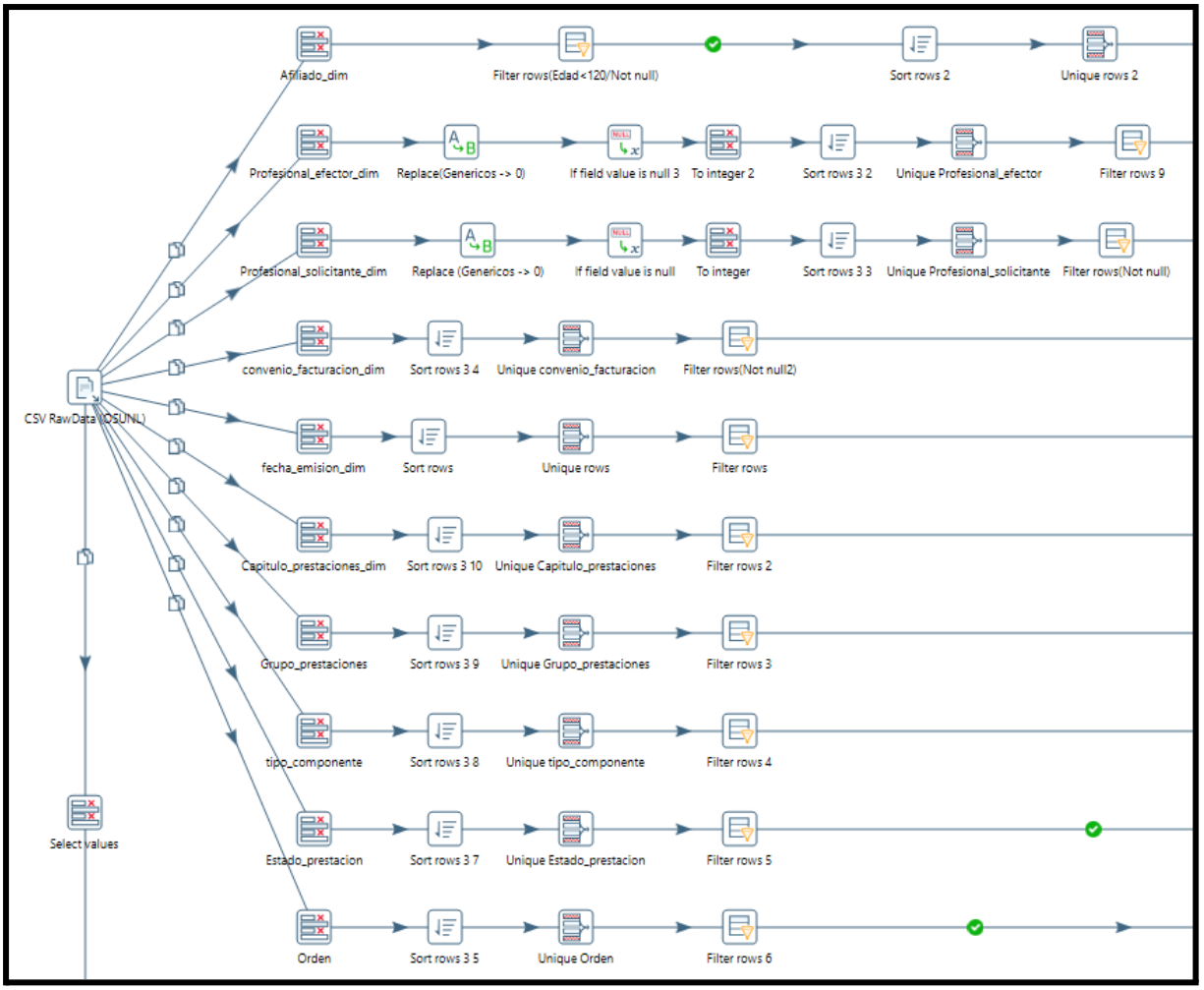


#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	Afiliado	Integer	#	15		$	,	.	both
2	EDAD	Integer	#	15		$	,	.	both
3	convenio_prestacion	Integer	#	15		$	,	.	none
4	Fecha_Emision	String		19		$	,	.	none
5	Profesional_efector	String				$	,	.	none
6	Profesional_solicitante	String				$	,	.	none
7	convenio_facturacion	Integer	#	15		$	,	.	none
8	Orden	Integer	#	15		$	,	.	none
9	Prestacion	String	#			$	,	.	none
10	Estado_prestacion	String		1		$	,	.	none
11	tipo_componente	Integer	#	15		$	,	.	none
12	Grupo_prestaciones	Integer	#	15		$	,	.	none
13	Capitulo_prestaciones	Integer	#	15		$	,	.	none

Transformación (Transform)

En esta etapa, los datos extraídos se someten a una serie de transformaciones para limpiar, filtrar, integrar y dar formato adecuado a los datos. Esto puede incluir la eliminación de datos duplicados o inconsistentes, la corrección de errores, la normalización de valores, la creación de nuevas columnas o cálculos derivados, entre otras operaciones. El objetivo es preparar los datos para su posterior carga y análisis.

En nuestro proyecto en particular, la primera parte de la transformación consiste en separar las columnas del dataset con las cuales crearemos nuestras dimensiones, y así poder trabajarlas individualmente. Estas columnas extraidas (Parte derecha de la imagen) reciben primero una limpieza de valores erróneos o nulos, reemplazando valores en el caso de que sea posible y filtrando outliers, como en el caso de la edad en particular. Luego estos valores son ordenados y filtrados para obtener solamente valores si repetirse, de esta manera obtendremos un único valor de cada variable a las cuales se le aplica un proceso de serialización y así de esta manera obtener un número entero para cada variable categórica, así de esta manera logramos reducir la redundancia y optimizar la eficiencia de almacenamiento de los datos.

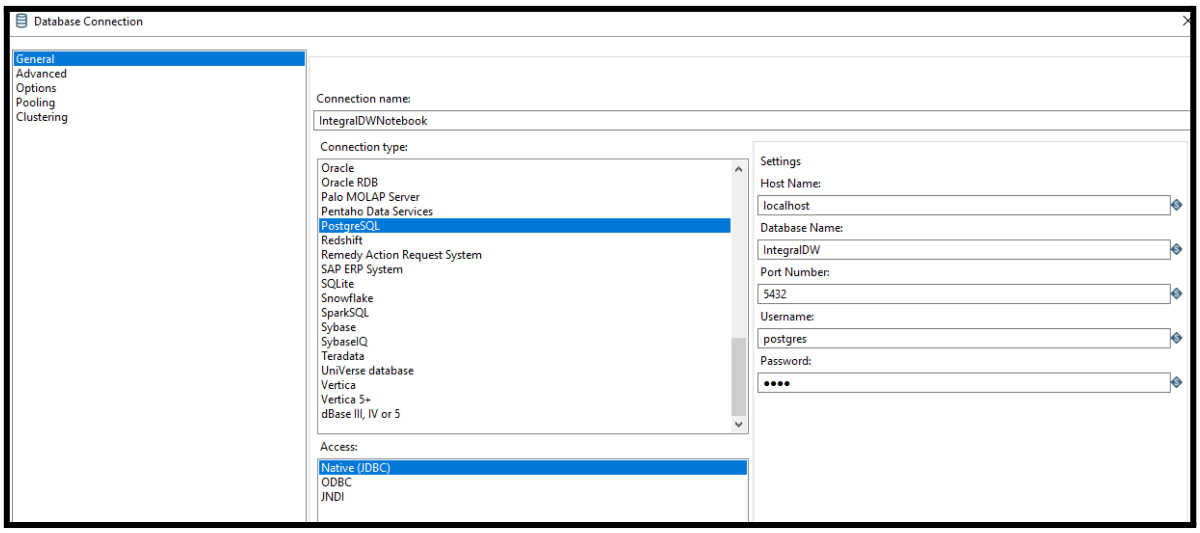


Carga (Load)

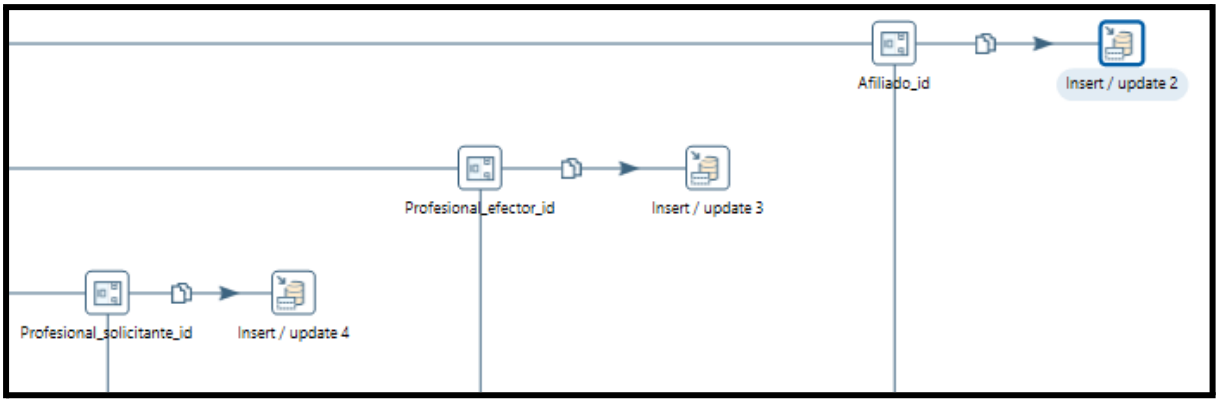
En esta etapa, los datos transformados se cargan en el destino final, que puede ser un almacén de datos, un data warehouse, una base de datos relacional u otro sistema de almacenamiento. Los datos se organizan y estructuran de acuerdo con el esquema de destino y se realizan las operaciones de carga necesarias, como la inserción, actualización o eliminación de registros. El objetivo es asegurar que los datos estén disponibles y listos para su uso en análisis o reportes.

Antes que nada debemos agregar una conexión a nuestro Data Warehouse definido anteriormente, una vez realizada la conexión exitosamente podemos continuar con la carga de nuestros datos. Para realizar la carga antes que nada debemos cargar los valores que conforman las dimensiones definidas, estas mismas contendrán los valores únicos de cada atributo por dimensión. Cada uno de estos valores es comprobado por una clave que asignamos al momento de carga, en nuestro caso sera el ID generado en el proceso de serialización de las variables.

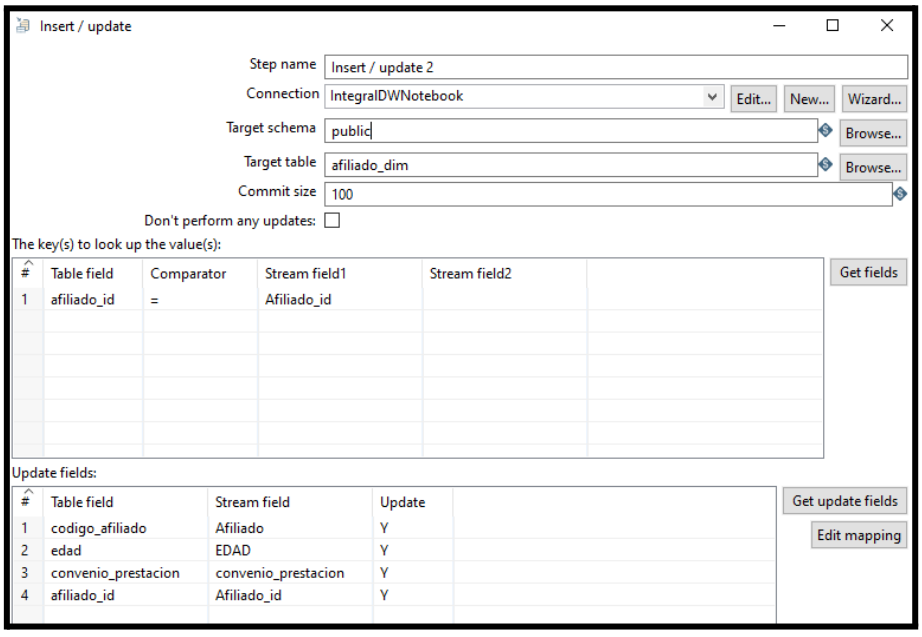
Conexión al servidor:



Ejemplo de carga de los valores a las dimensiones:

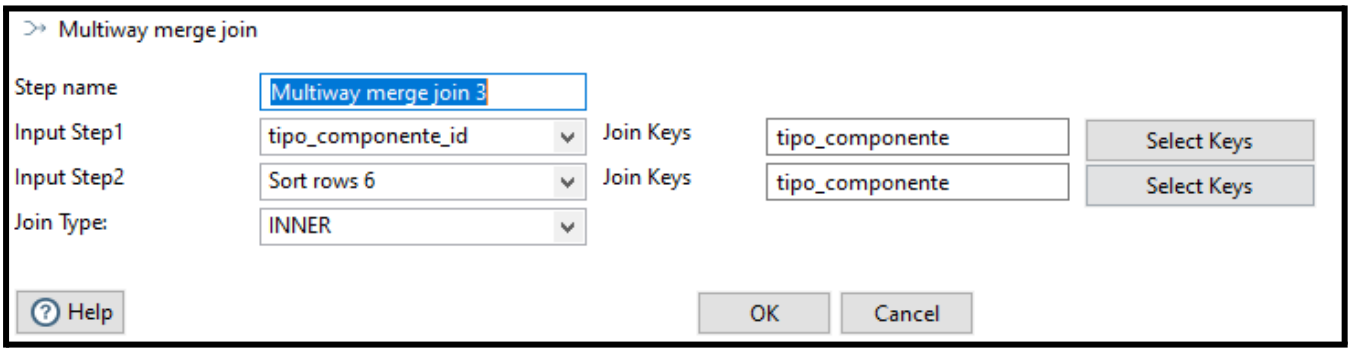


Ejemplo de carga de datos(dimensiones):

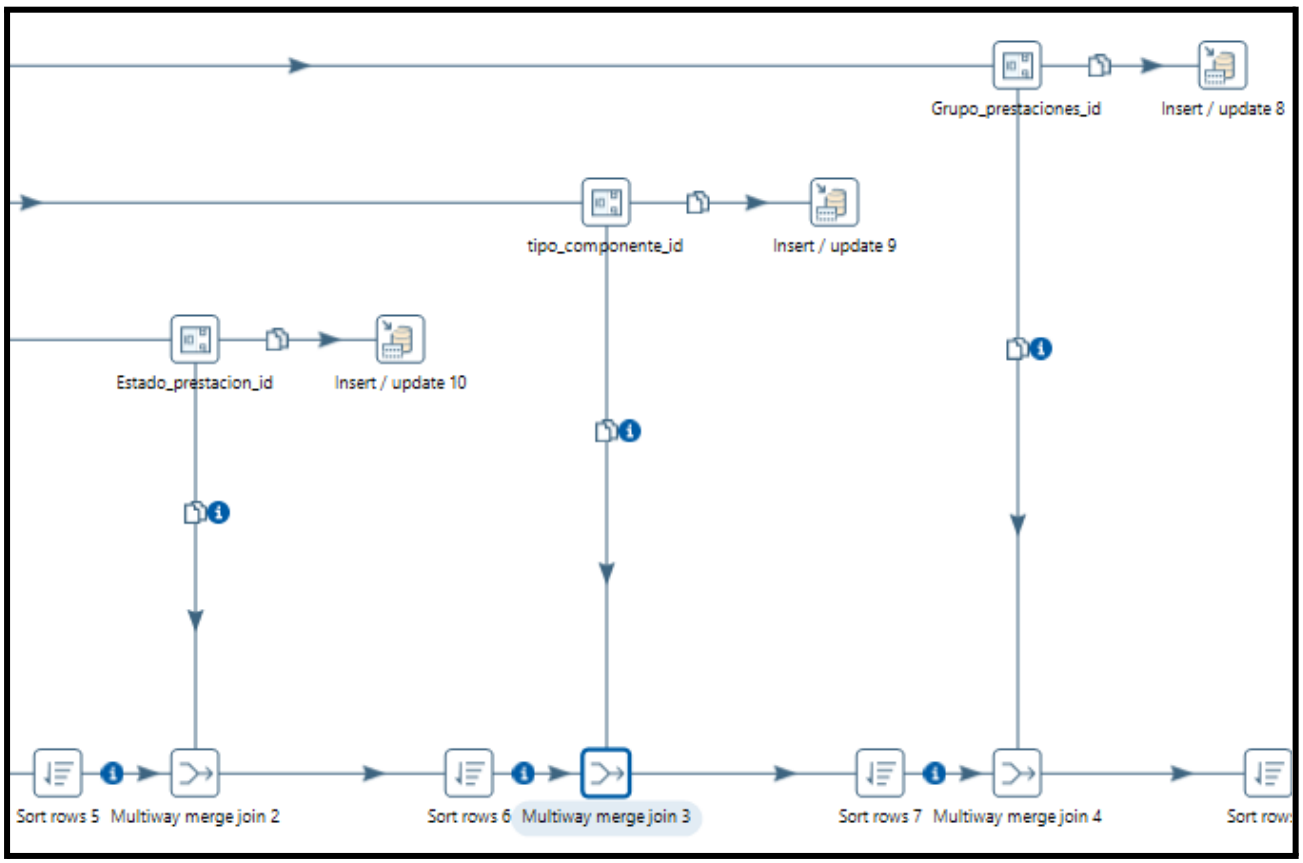


La otra parte del proceso de carga consta en la carga de la tabla de hechos(Fact table), al realizar un INNER JOIN o Multiway merge join, se establece una conexión entre las tablas dimensionales y la tabla de hechos utilizando los IDs correspondientes. Esto permite vincular los valores de las dimensiones con los valores en la tabla de hechos, creando un conjunto completo de datos analíticos.Es importante definir correctamente las relaciones y asegurarse de que los IDs sean únicos y coincidan entre las dimensiones y la tabla de hechos

Ejemplo Multiway merge join(INNER JOIN):



Ejemplos Multiway merge joins:

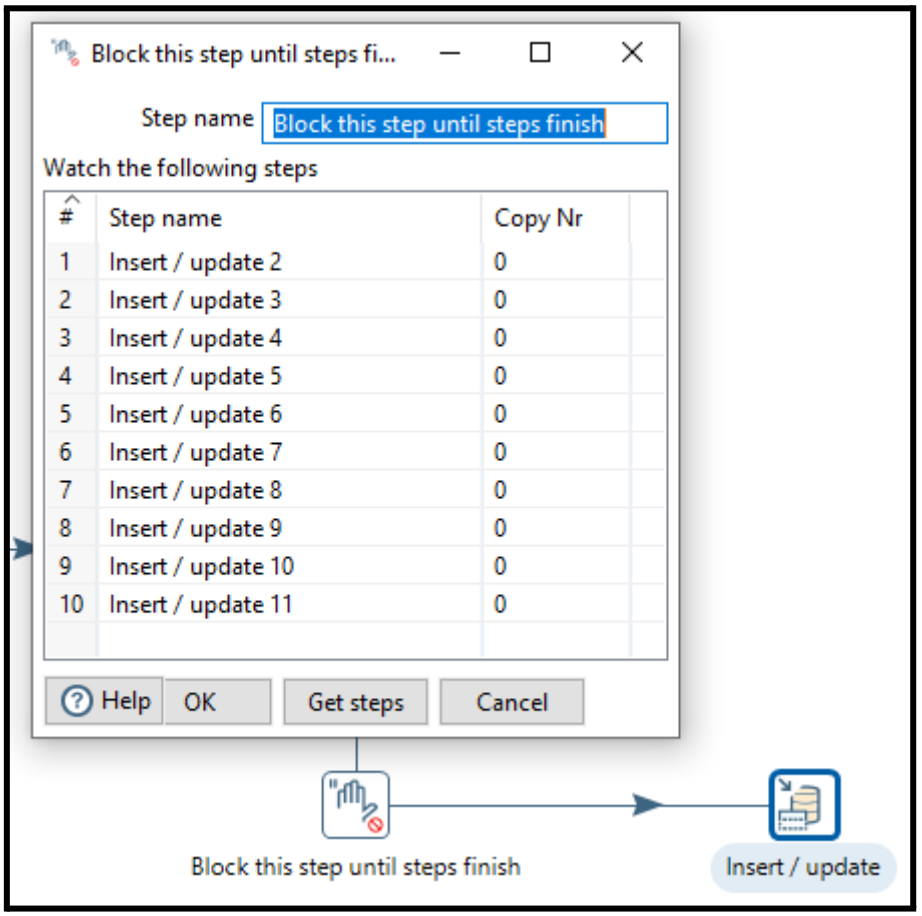


Como últimos pasos, procedemos a completar la carga de la tabla de hechos. Antes de realizar este paso, es importante asegurarse de que todas las dimensiones tengan cargados los valores y sus respectivos IDs. Estos IDs serán los que vinculen el número de la serialización de valores con el nombre correspondiente de cada uno.

Para garantizar que la carga en la tabla de hechos se realice posteriormente a la carga de las dimensiones, utilizamos un paso que nos permite bloquear el avance de la carga a la tabla de hechos hasta que todas las demás cargas hayan sido realizadas con éxito.

Este paso de bloqueo asegura la integridad referencial de los datos, ya que evita que se realice la carga de la tabla de hechos antes de que todas las dimensiones estén completamente cargadas. Una vez que todas las dimensiones han sido cargadas y los IDs están disponibles, podemos proceder con la carga de la tabla de hechos, asegurando la coherencia y consistencia de los datos en el Data Warehouse.

Bloquear la carga de la tabla de hechos hasta completar la carga de las dimensiones:

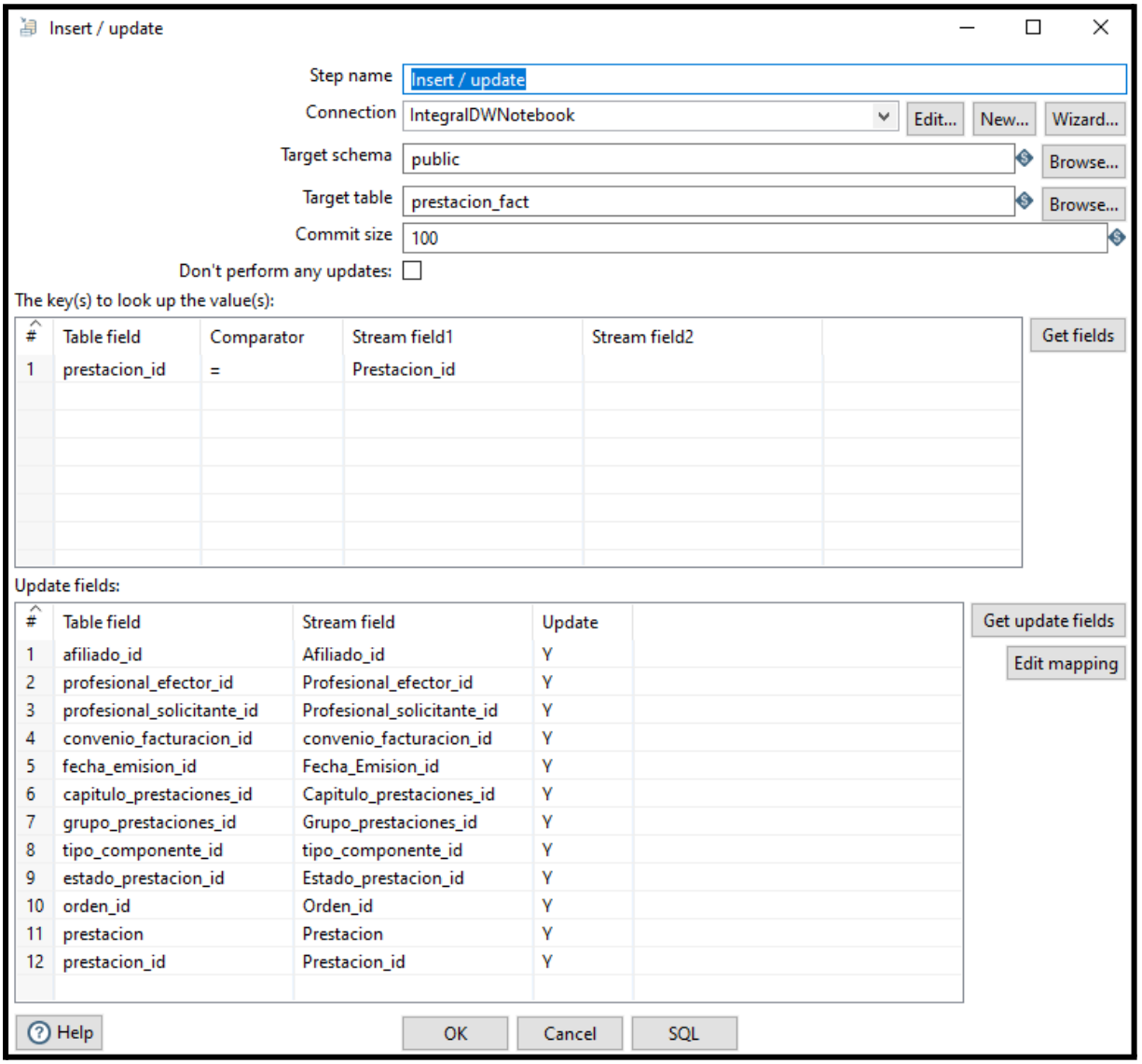


Al finalizar el proceso, procedemos a cargar la tabla de hechos donde asignamos un ID o valor único a cada prestación realizada. Esta tabla de hechos estará vinculada a los demás valores de dicha prestación a través de los IDs de cada variable presentes en las columnas.

Este enfoque nos permite reducir significativamente el tamaño y la carga de información en nuestro esquema estrella finalizado. Al utilizar IDs para representar los valores de las dimensiones, evitamos la duplicación innecesaria de datos en la tabla de hechos. En su lugar, simplemente hacemos referencia a los IDs correspondientes en las columnas de la tabla de hechos.

Esta estrategia de normalización y vinculación de datos nos brinda un esquema más eficiente y optimizado, permitiéndonos almacenar y analizar grandes volúmenes de información de manera más ágil y efectiva en nuestro entorno de Data Warehouse.

Carga de la tabla de hechos(Fact table):



Dashboard

Finalmente, en la fase de difusión se hace uso del nuevo conocimiento y se hace partícipe de él a los usuarios finales, tiene como fin el empleo de forma correcta del modelado de los datos en el contexto de la aplicación real y de los usuarios para los cuales se inició el proceso de extracción de conocimiento. Empresas de todo tamaño y profesionales independientes dependen de la correcta visualización de la información recopilada ya sea en sitios web, redes sociales, campañas de publicidad, correos electrónicos, etc, para así poder entender a sus clientes y así crear estrategias comerciales más eficientes.

Pasos previos

Comprension del usuario

Antes de comenzar una presentación, es importante hacerse preguntas sobre la audiencia. ¿Quiénes son? ¿Cuáles son sus necesidades y cómo puedes abordarlas? ¿Cómo puede la información que presentamos mejorar su rendimiento? ¿Qué decisiones deberían tomar después de que termine la presentación? Preguntas como estas son fundamentales para desarrollar contenido relevante y resonante. Pensar en la audiencia garantizará que estamos poniendo sus necesidades en primer lugar y nos dará un punto de referencia con el cual podemos evaluar el mensaje.

En realidad la comprensión de la información es un factor subjetivo, ya que depende en gran modo de la experiencia y conocimiento de los usuarios de la problemática. Los gráficos y tablas generalmente son considerados como unas de las representaciones que permiten comprender más fácilmente el comportamiento de los datos a través de un dashboard.

Los gráficos permiten que los usuarios puedan identificar fácilmente y de manera directa los patrones más significativos que se han descubierto. Asimismo, muchos de los métodos de visualización tienen funciones para que los propios usuarios modifiquen los informes para refinarlos o adaptarlos según su conocimiento o circunstancias del ámbito de aplicación.

Visualización

La visualización de los datos es una pieza clave para la problemática planteada, ya que muestra lo logrado respaldandose por datos. Este informe o dashboard servirá para guiar las acciones a medida que se incorporan lecciones aprendidas de datos históricos y otras actividades. Es por eso que debemos identificar qué elementos o variables funcionan para responder los problemas planteados, así como indicar las fortalezas y debilidades de las estrategias a utilizar en base a esta información.

Se debe brindar un resumen breve de la información a mostrar y enumerar los datos de mayor impacto. Esto se puede hacer en forma de tarjetas que muestren la información de manera concreta y en un formato fácil de distinguir. Es de importancia usar colores que destaquen dichos datos de importancia, se puede utilizar un degradado del color según la ocurrencia del dato observado.

Los gráficos que se usan en los informes deberán ser acordes a los objetivos y las métricas que se relacionan con lo que se desea mostrar. Una buena técnica al mostrar la información es comparando los resultados actuales con estados pasados para poder evaluar las mejoras o los puntos de oportunidad. De esta manera podemos valernos de gráficos lineales, capaces de mostrar la cantidad de variables a través de un periodo de tiempo.

El uso estratégico del color tiene un impacto significativo en la forma en que presentamos y percibimos los datos, y puede influir en la retención e interpretación de la información.

Cuando utilizamos colores de manera adecuada, podemos resaltar aspectos importantes de los datos, facilitar la identificación de patrones y tendencias, y proporcionar contexto adicional a la información presentada. Asimismo, los colores pueden utilizarse para indicar

estados o condiciones específicas, lo que ayuda a comprender la situación de manera más rápida y efectiva.

Por otro lado, es esencial utilizar los colores de forma coherente y con sentido, evitando la sobrecarga de información y asegurándonos de que sean accesibles y legibles para todos los usuarios. Al seleccionar colores, debemos considerar el contexto de la visualización y la audiencia a la que va dirigida, para lograr una experiencia visual agradable y comprensible.

El color es un elemento crucial que nos facilita la interpretación rápida de la información. Una adecuada elección de colores tiene un impacto significativo en la correcta comprensión de los datos. Pero debemos tener en cuenta que el uso excesivo de colores diferentes puede complicar la interpretación y dificultar la identificación de patrones y tendencias en los datos. Es importante evitar sobrecargar la visualización con una paleta de colores abrumadora.

Cuando se trata de mostrar evoluciones temporales no es necesario utilizar una gran variedad de colores, es posible utilizar un degradado de colores según la intensidad, para resaltar la progresión de una variable a lo largo del tiempo o en un rango específico. En cambio, al utilizar una escala cualitativa, asociamos cada variable única a un color diferente, lo que facilita la identificación y comparación de categorías diferentes.

Hay que tener en cuenta que los pasos pueden ser recursivos y que es importante no solo informar sobre los datos, sino también analizarlos para saber las razones tras los hechos. Esto hace posible cuestionarnos ciertas preguntas: ¿Por qué se dio esto? ¿A qué se debió esto otro? Es por eso que estas cuestiones ayudarán a estructurar las estrategias de diseño y a establecer nuevos objetivos en el marco del problema a resolver y permitiendo el rediseño de los gráficos que mostrarán la información para adaptarse a la problemática a resolver.

Microsoft Power BI

En el contexto del análisis de datos a través de visualizaciones y diseño de dashboards, existen herramientas muy diversas en el mercado, algunas gratuitas y otras bajo esquemas de pago. Sin embargo, todas coinciden en algo, es necesario aprender a utilizarlas para aprovechar su potencial. No basta con tener acceso a los datos, hay que analizarlos, compartirlos y representarlos de manera que resulten atractivos y expliquen las áreas de oportunidad, de forma comprensible y lógica.

En nuestro caso en particular usaremos como herramienta para el diseño de nuestro dashboard la aplicación **Microsoft Power BI,** esta brinda soluciones permitiendo unificar datos de diferentes fuentes, para elaborar un modelado y un tratamiento de los mismos y terminar ofreciendo una serie de informes, principalmente empresariales, de gran valor para el usuario final. Esta capa de informes, además de ofrecer unos elementos muy dinámicos y visuales, pueden ser compartidos tanto con otros usuarios de Power BI como con el público general de la web.

La aportación de Microsoft Power BI para las empresas es muy notable. Destacando una gran variedad de aspectos:

- Múltiples fuentes de información: Ideal para sistemas que tienen su información en bases de datos legacy y parte en bases más modernas.
- Es ideal para empresas con diferentes canales de ventas
- Mantenimiento por personas de negocio sin conocimientos técnicos muy altos.
- Permite compartir información entre departamentos.
- No hay necesidad de una infraestructura especial al ser un servicio cloud que se licencia mensualmente.
- Tiene muchas opciones de integración y publicación de informes tanto con otras herramientas de Microsoft como publicando en páginas web públicas.
- Posibilidad de consultar datos usando lenguaje natural sin necesidad de programar
- Muchos elementos gráficos y templates para la capa visual.

Como se mencionó anteriormente el diseño de esquema de estrella es sumamente efectivo para desarrollar modelos de Power BI optimizando el rendimiento y la facilidad de uso.

Hay que tener en cuenta que cada objeto visual de informe de Power BI genera una consulta que se envía al modelo de Power BI (lo que el servicio Power BI denomina un conjunto de datos). Estas consultas se usan para filtrar, agrupar y resumir los datos del modelo. Por tanto, un modelo bien diseñado es aquel que proporciona tablas para filtrar y agrupar y tablas para resumir. Este diseño se ajusta bien a los principios de los esquemas de estrella donde tenemos:

- Las tablas de dimensiones, que admiten el filtrado y la agrupación.
- Las tablas de hechos, permitiendo el resumen de datos.

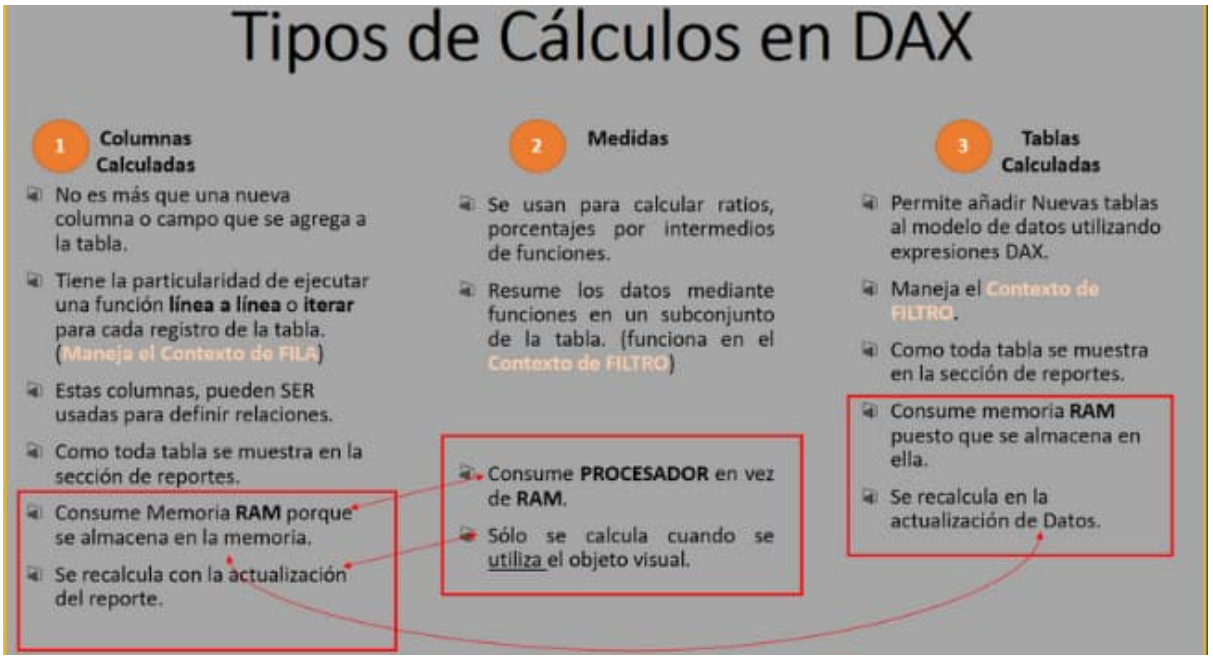
Medidas

Las medidas (también conocidas como KPI - Key Performance Indicators) son métricas específicas utilizadas para medir el rendimiento o el progreso hacia los objetivos en un

negocio o proyecto. Estas medidas cuantifican el desempeño de diferentes aspectos del negocio y son fundamentales para tomar decisiones informadas y evaluar el éxito de una estrategia.

Hay que tener en cuenta que en el diseño de esquemas de estrella, una medida es una columna de tabla de hechos que almacena valores que se van a resumir, en Power BI, algunas medidas son calculadas automáticamente por la herramienta cuando se crea un gráfico o visualización utilizando los datos en el modelo de datos. Estas medidas automáticas se basan en el tipo de datos y la jerarquía de las columnas utilizadas en el gráfico.

Por otro lado también podemos realizar una fórmula escrita en Expresiones de análisis de datos (DAX) que nos permite resumir los datos que seleccionemos. Las expresiones de medida suelen aprovechar funciones de agregación de DAX como SUM, MIN, MAX, AVERAGE, etc. para generar un resultado de valor escalar en tiempo de consulta (estos valores nunca se almacenan en el modelo). La expresión de medida puede abarcar desde agregaciones de columnas simples hasta fórmulas más sofisticadas que invalidan las propagaciones de contexto o de relación de filtrado.



Power BI utiliza estas características de modelado de datos, incluyendo relaciones y jerarquías, para analizar y visualizar datos de manera efectiva. También implementa DAX para crear cálculos y medidas personalizadas basadas en relaciones y jerarquías.

Al utilizar estas características y técnicas, es posible obtener información valiosa de los datos y tomar mejores decisiones empresariales. Las características de modelado de datos de Power BI, como las relaciones y jerarquías, permiten crear ideas significativas a partir de los datos.

Resultados obtenidos

Menú Principal

La primera página de nuestro dashboard posee un menú interactivo con enlaces que conducen a los diferentes informes, es una excelente manera de facilitar la navegación y mejorar la experiencia del usuario al permitirle navegar directamente a la información que está solicitando. Para navegar a cualquiera de estos informes, simplemente hay que seleccionar el enlace correspondiente y será dirigido al análisis detallado, por otra parte cada uno de estos informes contiene un enlace para regresar al menú principal.

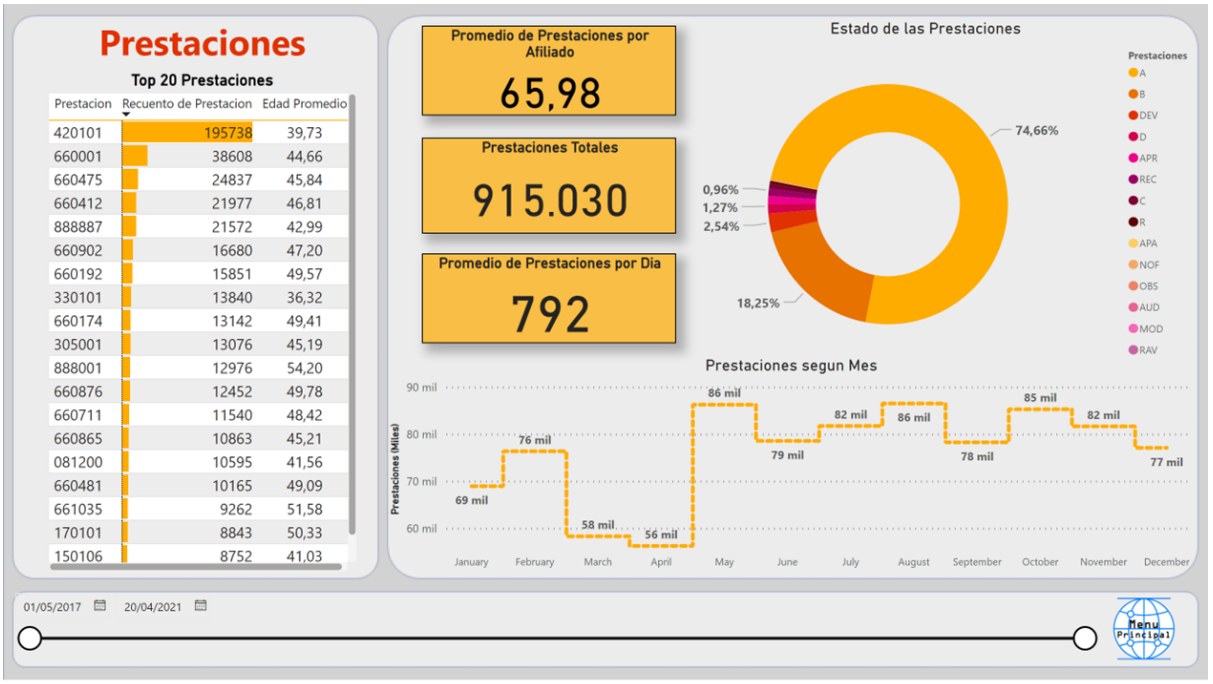
Segmentación de datos

Una funcionalidad que poseen los informes en común es la de segmentar los datos(Slice) por fecha mediante una barra deslizante. Esta es una forma efectiva de filtrar y visualizar datos dentro de un rango de fechas específicas. La barra deslizante permite a los usuarios ajustar fácilmente el período de tiempo para analizar datos en diferentes intervalos sin tener que seleccionar manualmente fechas individuales.

Con esta segmentación de datos por fecha mediante una barra deslizante, los usuarios pueden explorar rápidamente datos a lo largo del tiempo, centrarse en intervalos específicos y obtener una comprensión más profunda de los patrones y tendencias en sus datos. Es una herramienta poderosa para análisis temporales en Power BI y mejora la experiencia del usuario al proporcionar una forma intuitiva de interactuar con los datos.

Prestaciones

En el contexto de nuestra problemática, una "prestación" se refiere a un servicio o beneficio específico que se proporciona a los afiliados o beneficiarios de la mutual. Estas prestaciones pueden incluir una amplia variedad de servicios de salud, asistencia médica, cobertura de medicamentos, atención especializada, entre otros.



En el informe de prestaciones, se pueden visualizar diferentes aspectos importantes relacionados con las prestaciones solicitadas a través de diversos tipos de gráficos y KPIs.

El gráfico de barras (Top de 20 Prestaciones) nos muestra las prestaciones más solicitadas, ordenadas por la cantidad de veces que han sido requeridas y la edad promedio de los afiliados que la solicitaron. Cada barra representa una prestación específica, y la altura de la barra indica la cantidad de veces que se ha solicitado esa prestación en particular.

Al utilizar KPIs (Key Performance Indicators) podemos analizar el rendimiento de las prestaciones. Algunos medidas relevantes son:

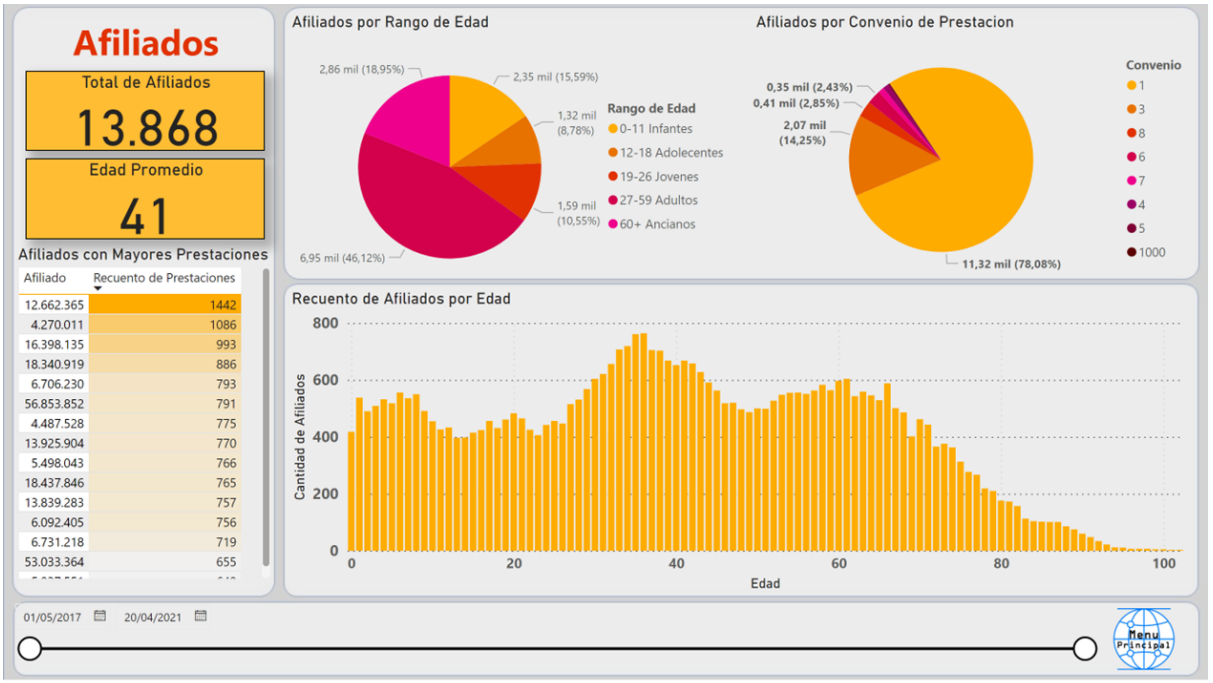
- Promedio de Prestaciones por Afiliado: Calcula el número promedio de prestaciones solicitadas por cada afiliado durante un período específico.
- Totalidad de Prestaciones: Muestra el número total de prestaciones realizadas en un período determinado.
- Promedio de Prestaciones por Día: Indica la cantidad promedio de prestaciones realizadas por día en un período dado.

En cuanto al estado de las prestaciones el gráfico circular (también conocido como gráfico de torta) muestra el estado actual de las prestaciones solicitadas. Cada porción del círculo representa un estado específico y el tamaño de cada porción está relacionado con el porcentaje de prestaciones en ese estado en comparación con el total.

El Gráfico de Líneas Escalonadas que muestra la Cantidad de Prestaciones por Mes (también llamado gráfico de líneas de paso) muestra la cantidad de prestaciones solicitadas durante cada mes a lo largo del tiempo. Cada meseta en el gráfico representa el número de prestaciones solicitadas en un mes específico, y las líneas escalonadas conectan los puntos, lo que permite visualizar cualquier tendencia o variación en la cantidad de prestaciones a lo largo del tiempo.

Afiliados

El informe de los afiliados en Power BI presenta una visión general de los datos relacionados con los afiliados proporcionando información detallada sobre los afiliados, permitiendo identificar patrones, tendencias y segmentaciones importantes para la toma de decisiones y el análisis en el contexto de los datos de los afiliados.



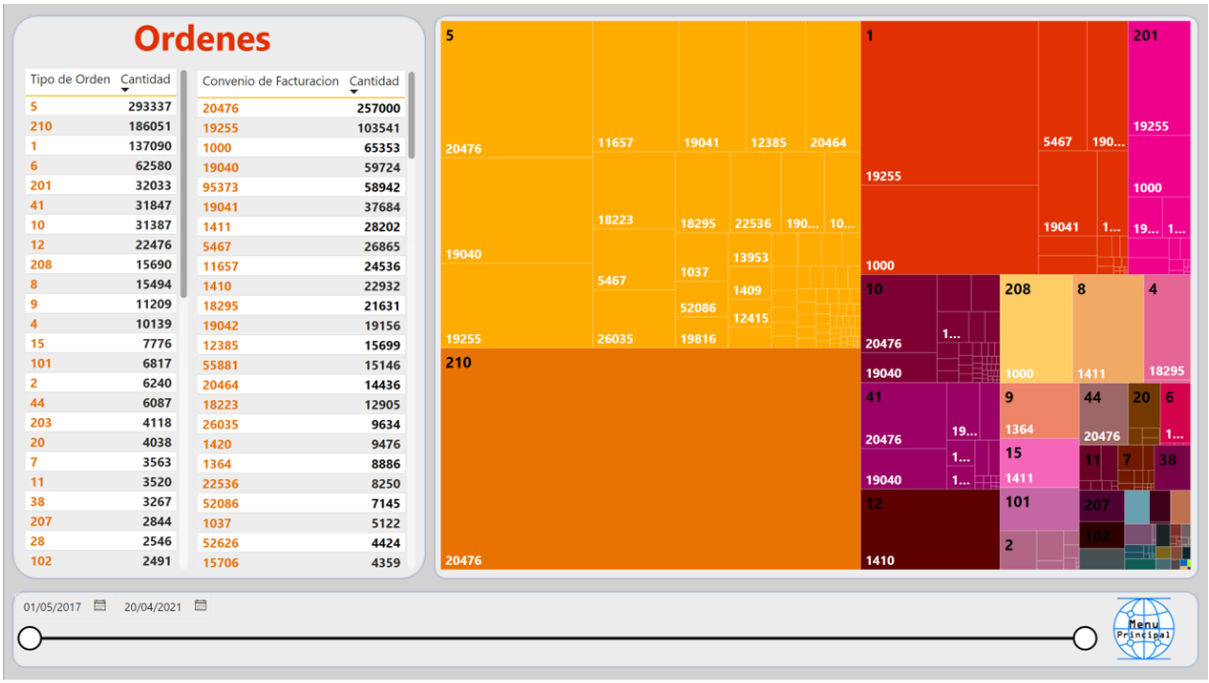
Tenemos por un lado información en forma de tarjetas mostrando la cantidad total de afiliados, como también la edad promedio de los afiliados.

Por otro lado, el informe presenta un gráfico circular mostrando la proporción de afiliados en diferentes rangos de edad, lo que permite visualizar la distribución de edades en forma de porcentaje. Otro gráfico circular muestra la cantidad de afiliados agrupados según el convenio de prestaciones que poseen, lo que brinda una visualización rápida de la cantidad de afiliados por cada tipo de convenio.

Una tabla nos muestra los afiliados que tienen la mayor cantidad de prestaciones registradas, lo que permite identificar a aquellos que utilizan más servicios y un gráfico de barras nos presenta el recuento de afiliados en los diferentes años de edad, lo que ayuda a comprender mejor la distribución de edades y permite identificar las franjas etarias con mayor presencia.

Órdenes

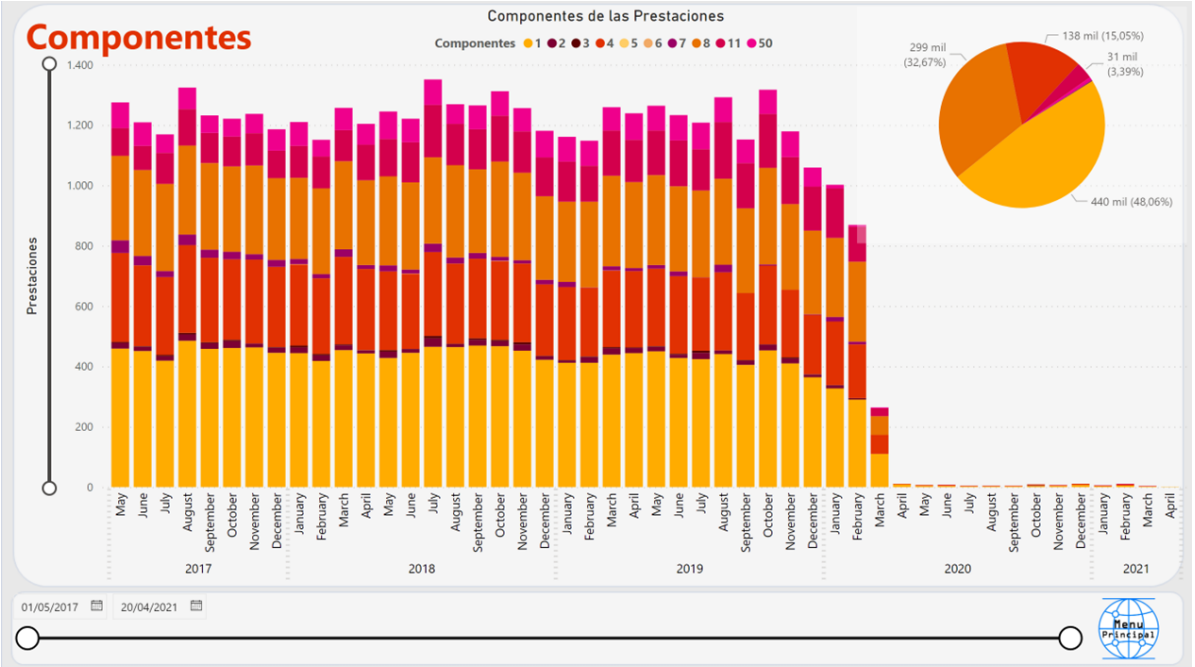
El informe sobre las órdenes, presenta información sobre los tipos de órdenes y la cantidad de cada tipo de orden realizada, así como los convenios de facturación y la cantidad de órdenes solicitadas para cada convenio.



Esta información también se puede visualizar de manera efectiva mediante un gráfico Treemap(mapeo de árboles), esta visualización muestra jerarquías y proporciones de datos en forma de rectángulos anidados. Cada rectángulo representa una categoría de orden y como subcategoría los convenios de facturación, el tamaño se corresponde con la cantidad o el valor asociado. Con esta visualización, se muestra clara y concisamente la distribución de órdenes por tipo y convenio de facturación, esta es una excelente opción para presentar datos jerárquicos y proporcionar una vista global y visualmente atractiva de la información relacionada con las órdenes.

Componentes

En el informe sobre los componentes podemos observar la cantidad de componentes solicitados para las prestaciones, diferenciándolos por color según la categoría a la que correspondan.



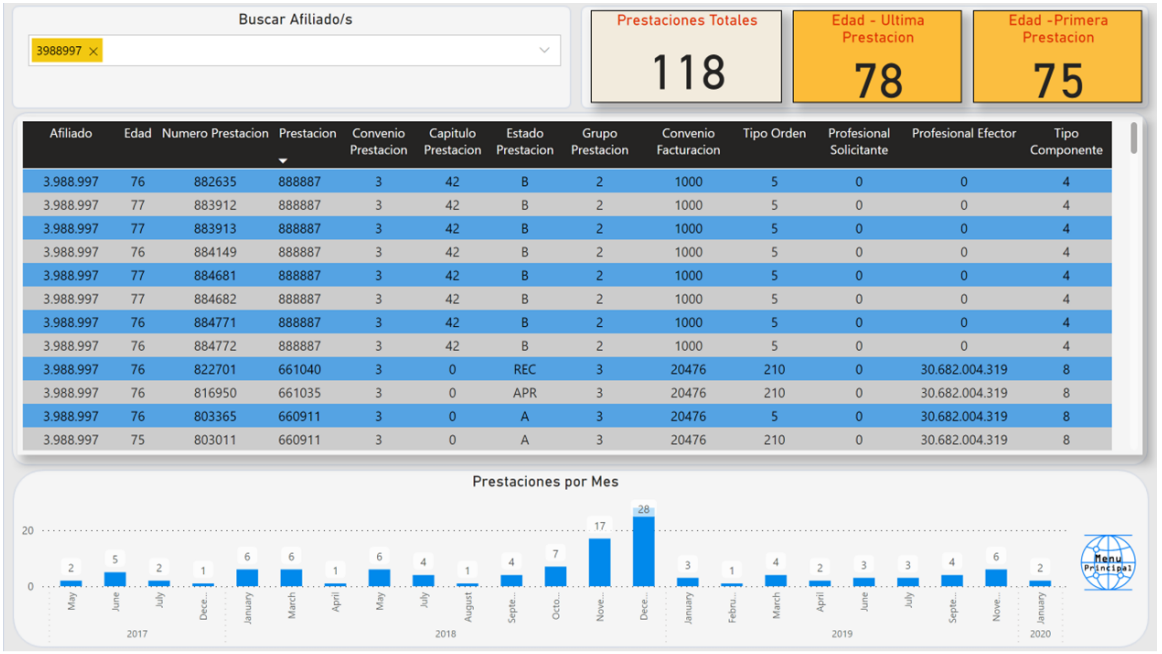
Esta información se visualiza mediante un gráfico de columnas apiladas, donde las columnas representan la cantidad de componentes dividida en segmentos de diferentes colores que representan los diferentes componentes requeridos por mes durante el periodo seleccionado. Además, este gráfico cuenta con un control deslizante que permite controlar la escala del eje Y permitiendo que los usuarios puedan ajustar la escala del eje vertical para enfocarse en una sección específica del gráfico y ver con mejor detalle la cantidad de componentes solicitados.

Junto con el gráfico de columnas apiladas, el informe también incluye un gráfico circular que muestra la cantidad de componentes por categoría, mostrando la cantidad de componentes y sus porcentajes.

Búsquedas

Por último se han implementado dos paneles de búsquedas, aplicando un filtro que permite a los usuarios realizar búsquedas específicas de afiliados o profesionales efectores. Estas búsquedas se realizan a través de una barra de búsqueda, donde los usuarios pueden ingresar el nombre o identificación del afiliado o profesional que desean buscar.

Una vez que se ingresa el término de búsqueda en la barra de búsqueda, el filtro se aplica y se muestra información relevante de las personas buscadas en los paneles correspondientes.

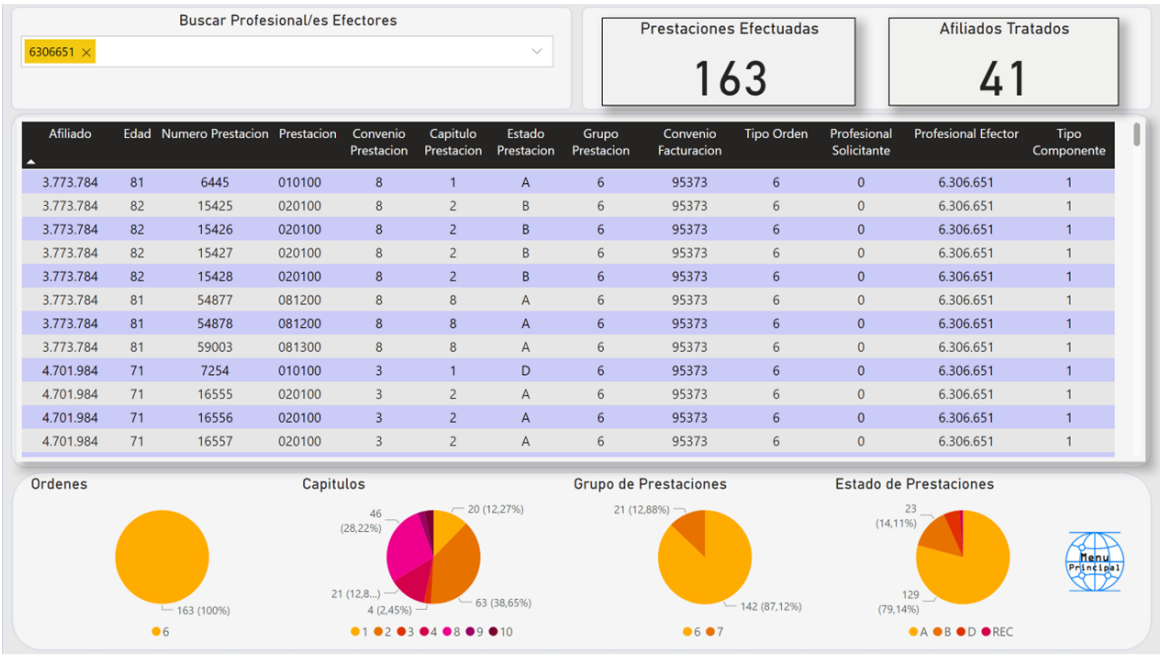


Afiliado	Edad	Numero Prestacion	Prestacion	Convenio Prestacion	Capitulo Prestacion	Estado Prestacion	Grupo Prestacion	Convenio Facturacion	Tipo Orden	Profesional Solicitante	Profesional Efector	Tipo Componente
3.988.997	76	882635	888887	3	42	B	2	1000	5	0	0	4
3.988.997	77	883912	888887	3	42	B	2	1000	5	0	0	4
3.988.997	77	883913	888887	3	42	B	2	1000	5	0	0	4
3.988.997	76	884149	888887	3	42	B	2	1000	5	0	0	4
3.988.997	77	884681	888887	3	42	B	2	1000	5	0	0	4
3.988.997	77	884682	888887	3	42	B	2	1000	5	0	0	4
3.988.997	76	884771	888887	3	42	B	2	1000	5	0	0	4
3.988.997	76	884772	888887	3	42	B	2	1000	5	0	0	4
3.988.997	76	822701	661040	3	0	REC	3	20476	210	0	30.682.004.319	8
3.988.997	76	816950	661035	3	0	APR	3	20476	210	0	30.682.004.319	8
3.988.997	76	803365	660911	3	0	A	3	20476	5	0	30.682.004.319	8
3.988.997	75	803011	660911	3	0	A	3	20476	210	0	30.682.004.319	8

En el caso del panel de búsqueda de afiliados, nos encontramos con KPIs que indican la cantidad total de prestaciones realizadas por el paciente seleccionado, así como la edad de la última y primera prestación registrada para ese paciente.

El panel también presenta una tabla con la información relevante del afiliado seleccionado, donde se pueden visualizar detalles como el nombre, número de afiliado, edad, convenio de prestaciones y otros datos importantes.

Además podemos ver un gráfico de barras que muestra las prestaciones realizadas por el paciente a lo largo del tiempo, segmentadas por mes. Este gráfico proporciona una visión clara de la frecuencia de prestaciones a lo largo del tiempo, lo que permite identificar tendencias o patrones en la utilización de servicios de salud por parte del afiliado.



Afiliado	Edad	Numero Prestacion	Prestacion	Convenio Prestacion	Capitulo Prestacion	Estado Prestacion	Grupo Prestacion	Convenio Facturacion	Tipo Orden	Profesional Solicitante	Profesional Efector	Tipo Componente
3.773.784	81	6445	010100	8	1	A	6	95373	6	0	6.306.651	1
3.773.784	82	15425	020100	8	2	B	6	95373	6	0	6.306.651	1
3.773.784	82	15426	020100	8	2	B	6	95373	6	0	6.306.651	1
3.773.784	82	15427	020100	8	2	B	6	95373	6	0	6.306.651	1
3.773.784	82	15428	020100	8	2	B	6	95373	6	0	6.306.651	1
3.773.784	81	54877	081200	8	8	A	6	95373	6	0	6.306.651	1
3.773.784	81	54878	081200	8	8	A	6	95373	6	0	6.306.651	1
3.773.784	81	59003	081300	8	8	A	6	95373	6	0	6.306.651	1
4.701.984	71	7254	010100	3	1	D	6	95373	6	0	6.306.651	1
4.701.984	71	16555	020100	3	2	A	6	95373	6	0	6.306.651	1
4.701.984	71	16556	020100	3	2	A	6	95373	6	0	6.306.651	1
4.701.984	71	16557	020100	3	2	A	6	95373	6	0	6.306.651	1

El panel de búsqueda de profesionales efectores ofrece una visión detallada de la información relacionada con cada profesional. Aquí encontramos los KPIs que muestran la cantidad total de prestaciones que ha realizado cada profesional, lo que proporciona una medida de la productividad y la carga de trabajo de cada profesional. Además, se visualiza la cantidad de pacientes tratados por cada profesional, lo que permite evaluar su nivel de actividad y su alcance en el tratamiento de pacientes.

En cuanto a la tabla con información extra sobre el profesional, se pueden encontrar detalles adicionales como el nombre del profesional, su especialidad, su número de registro o identificación, entre otros datos relevantes.

Para tener una comprensión más completa de la actividad del profesional, se presentan gráficos circulares que muestran información detallada sobre las órdenes, los capítulos, los grupos de prestaciones y el estado de las prestaciones asociadas a dicho profesional. Estos gráficos permiten identificar rápidamente la distribución de las prestaciones por categorías y su estado, lo que puede ser útil para evaluar el enfoque de trabajo del profesional y detectar posibles áreas de mejora.

Dashboard interactivo

Dashboard Web Interactivo

Un dashboard interactivo de Power BI publicado en la web es una herramienta poderosa para visualizar y analizar datos de manera dinámica y accesible para usuarios en línea. Una vez que el dashboard está listo, se publica en Power BI Service, que es la plataforma en línea de Microsoft donde los dashboards y reportes de Power BI pueden ser alojados y compartidos. Esto se hace a través de una cuenta de Power BI, ya sea una cuenta personal o una cuenta de organización. Hay que tener en cuenta que es necesaria una cuenta de Power BI para acceder al mismo.

Una vez que el dashboard está publicado, los usuarios pueden acceder a él a través de un enlace web. Al acceder al dashboard, los usuarios pueden interactuar con los datos y visualizaciones de varias formas. Pueden aplicar filtros para enfocarse en un segmento específico de datos, hacer clic en elementos visuales para obtener más detalles o realizar análisis exploratorios para obtener insights en tiempo real.

Los creadores del dashboard pueden compartir el enlace del dashboard con otras personas o equipos para que puedan acceder al mismo y ver los datos en tiempo real. Además, Power BI Service permite la colaboración en tiempo real, lo que significa que múltiples usuarios pueden ver y trabajar en el dashboard simultáneamente.

También en esta manera de presentación del informe es posible que reciba actualizaciones automáticas, si los datos subyacentes cambian, Power BI puede configurarse para actualizar automáticamente el dashboard con los nuevos datos, lo que asegura que los usuarios siempre accedan a información actualizada.

Proyecto como Microservicio

Una vez completados todos los pasos del proyecto, se llevó a cabo la implementación los procesos realizados como un microservicio en contenedores utilizando la herramienta Docker, buscando proporcionar portabilidad y aislamiento como así la también la ejecución en diferentes entornos, agilizando el proceso de implementación y eliminando problemas de dependencias y configuraciones que a menudo afectan a las implementaciones tradicionales.

Los microservicios son un enfoque arquitectónico y organizativo para el desarrollo de software donde el software está compuesto por pequeños servicios independientes que se comunican a través de API bien definidas. Los propietarios de estos servicios son equipos pequeños independientes.

Las arquitecturas de microservicios hacen que las aplicaciones sean más fáciles de escalar y más rápidas de desarrollar. Esto permite la innovación y acelera el tiempo de comercialización de las nuevas características.

Con una arquitectura de microservicios, una aplicación se crea con componentes independientes que ejecutan cada proceso de la aplicación como un servicio. Los servicios se crean para las capacidades empresariales y cada servicio desempeña una sola función. Debido a que se ejecutan de forma independiente, cada servicio se puede actualizar, implementar y escalar para satisfacer la demanda de funciones específicas de una aplicación.

La adopción de contenedores proporciona un entorno de ejecución estandarizado y aislado para los servicios realizados. Los contenedores permiten empaquetar todas las dependencias y bibliotecas necesarias para que cada servicio funcione de manera coherente en cualquier entorno. Esto garantiza que el software se ejecutara de manera predecible y consistente en todos los entornos, eliminando posibles problemas de compatibilidad y facilitando la implementación rápida y repetible.

Implementación con Docker

Docker es una plataforma de contenedores que permite a los desarrolladores empaquetar aplicaciones y servicios en un contenedor portátil y ligero. Cada contenedor es una unidad de software que contiene todo lo necesario para que una aplicación se ejecute, incluyendo el código, las bibliotecas y las dependencias.

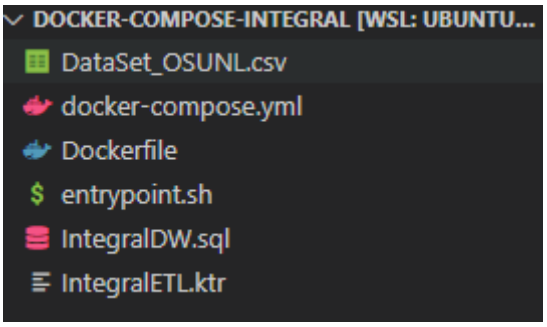
En Docker, los contenedores se ejecutan en un host, que puede ser una máquina virtual o un servidor físico. Cada contenedor se ejecuta en su propio espacio aislado y tiene acceso a sus propias dependencias y recursos.

Algunos de los conceptos básicos de Docker:

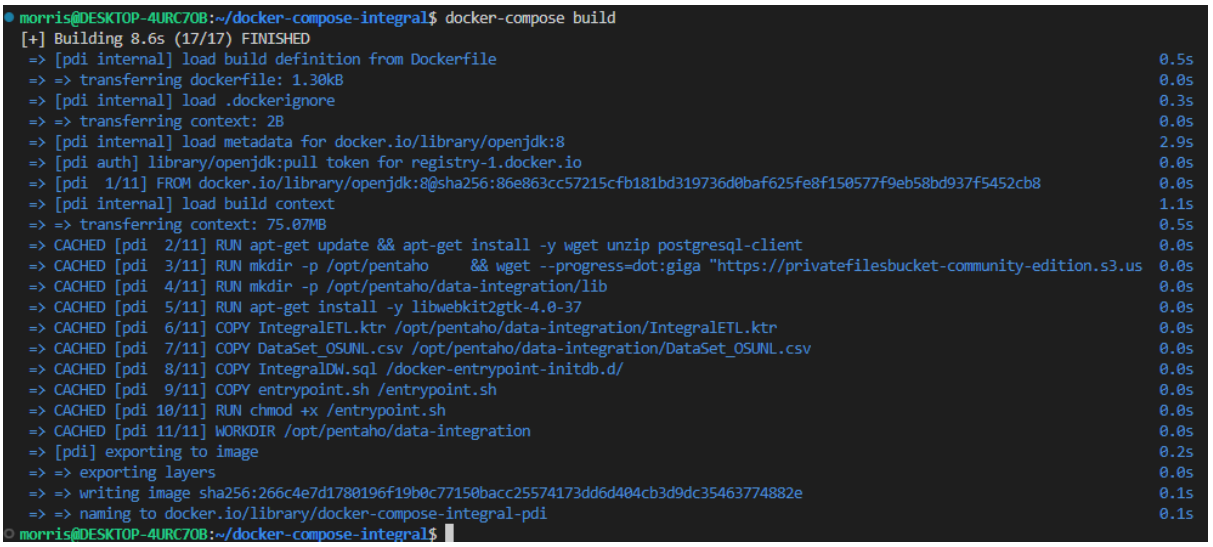
Imagen

Una imagen de Docker es un paquete que incluye todo lo necesario para ejecutar una aplicación, incluyendo el código, las bibliotecas y las dependencias.

Archivos para crear nuestra imagen con Docker-compose



Construcción de la Imagen del proyecto



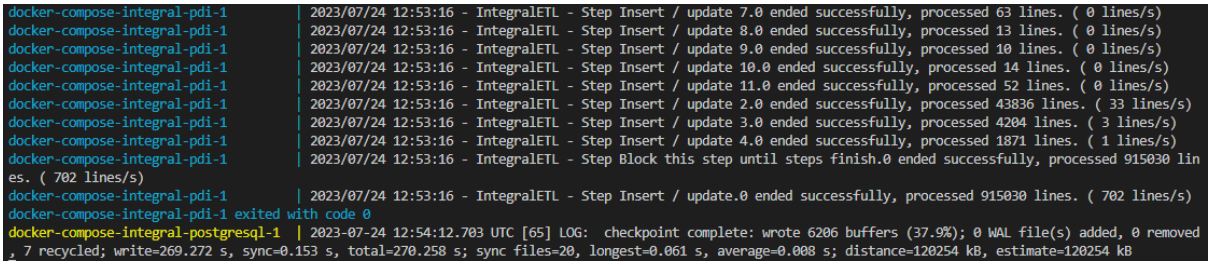
Contenedor

Un contenedor de Docker es una instancia en tiempo de ejecución de una imagen. Cada contenedor es una unidad aislada que contiene todo lo necesario para que una aplicación se ejecute de manera independiente y segura.

Los contenedores se ejecutan en un sistema operativo subyacente compartido, pero están aislados de los demás contenedores y del sistema operativo host. Esto significa que los contenedores pueden ejecutarse en cualquier plataforma que admita Docker, sin preocuparse por las diferencias de configuración y las dependencias de la infraestructura subyacente.

Además, un contenedor Docker es órdenes de magnitud más ligero que una máquina virtual, por lo que supone un paso hacia delante en comparación con lo que se venía realizando hasta ahora.

Ejecución del contenedor



Dockerfile

Un archivo Dockerfile es un archivo de texto que contiene una serie de instrucciones que Docker utiliza para crear una imagen. El archivo Dockerfile incluye información sobre el entorno de la aplicación, las dependencias y las instrucciones para construir y configurar la imagen.

Archivo dockerfile del proyecto

```
FROM openjdk:8

# Install additional dependencies
RUN apt-get update && apt-get install -y wget unzip postgresql-client

# Download and Install PDI
ENV PDI_VERSION=9.2
ENV PDI_BUILD=9.2.0.0-290
ENV PDI_HOME=/opt/pentaho
ENV PATH=$PDI_HOME/data-integration:$PATH

RUN mkdir -p $PDI_HOME \
    && wget --progress=dot:giga "https://privatefilesbucket-community-edition.s3
    && unzip -q "pdi-ce-$PDI_BUILD.zip" -d $PDI_HOME \
    && rm "pdi-ce-$PDI_BUILD.zip"

RUN mkdir -p /opt/pentaho/data-integration/lib

# Install additional dependencies for PDI
RUN apt-get install -y libwebkit2gtk-4.0-37

# Copy Transformation Files
COPY IntegralETL.ktr $PDI_HOME/data-integration/IntegralETL.ktr
COPY DataSet_OSUNL.csv $PDI_HOME/data-integration/DataSet_OSUNL.csv
```

Docker Compose

La herramienta Docker Compose permite a los desarrolladores definir y ejecutar aplicaciones multicontenedor. Docker Compose utiliza un archivo YAML para definir los contenedores que forman una aplicación y las relaciones entre ellos.

Archivo docker-compose del proyecto

```yaml
services:
  postgresql:
    image: postgres:latest
    environment:
      POSTGRES_USER: postgres
      POSTGRES_PASSWORD: 1988
      POSTGRES_DB: IntegralDW
    volumes:
      - pgdata:/var/lib/postgresql/data
      - ./IntegralDW.sql:/docker-entrypoint-initdb.d/IntegralDW.sql  # Mount the Integ
    ports:
      - "5433:5432"

  pdi:
    build:
      context: .
      dockerfile: Dockerfile
    depends_on:
      - postgresql
    volumes:
      - /opt/pentaho/data-integration/
    command: /entrypoint.sh
    ports:
      - "8080:8080"
    mem_limit: "4g"  # Change "2g" to the desired memory limit (e.g., "1g" for 1GB)
```

Archivo entrypoint

```bash
# Database connection parameters
db_host="postgresql"          # Replace with the hostname of the PostgreSQL container
db_port="5432"                # Replace with the port number of the PostgreSQL container
db_name="IntegralDW"          # Replace with the name of your PostgreSQL database
db_user="postgres"            # Replace with the PostgreSQL username
db_password="1988"            # Replace with the PostgreSQL password

# Wait for PostgreSQL to be fully ready (you can customize this loop if needed)
until psql "postgresql://${db_user}:${db_password}@${db_host}:${db_port}/${db_name}" -c '\l'; do
  >&2 echo "PostgreSQL is unavailable - sleeping"
  sleep 2
done

# Execute the SQL script to create tables in the PostgreSQL database
echo "Creating tables in the database..."
psql "postgresql://${db_user}:${db_password}@${db_host}:${db_port}/${db_name}" -f /docker-entrypoint-initdb.d/IntegralDW.sql


# Execute the Kettle Transformation (KTR)
echo "Running KTR transformation..."
/opt/pentaho/data-integration/pan.sh -file=/opt/pentaho/data-integration/IntegralETL.ktr
```

Conclusiones

El proyecto de análisis de datos sobre la empresa Integral Software SRL muestra la aplicación de diferentes herramientas y metodologías aprendidas a lo largo de la carrera de Procesamiento y Explotación de datos para para gestionar y mejorar la forma en que se brinda la información generada a partir de una base de datos de dicha empresa. La utilización de minería de datos, modelos predictivos y descriptivos, así como la implementación de un Datawarehouse nos permite aprovechar al máximo la información disponible y facilitar la toma de decisiones fundamentadas en información histórica.

El proceso de minería sobre estos datos nos facilitó la extracción de conocimiento significativo a partir de ellos, permitiéndonos identificar patrones, relaciones y reglas previamente desconocidas. Asimismo, los modelos predictivos y descriptivos serán valiosos para explorar y explicar las propiedades de los datos, así como para estimar valores desconocidos de las variables de interés.

Un punto importante a destacar en este proyecto es el uso de la metodología Kimball en el diseño del Datawarehouse, garantizando una implementación rápida y efectiva, permitiendo generar informes y análisis complejos de manera rápida y eficiente. La organización de los datos en tablas de hechos y tablas dimensionales, con sus respectivas relaciones, facilita el análisis multidimensional y la agregación de medidas.

En cuanto al proceso de limpieza y transformación de datos, debemos considerar que es una parte crucial para garantizar la calidad de la información utilizada en el proyecto. La identificación y corrección de valores faltantes, outliers y datos erróneos nos aseguran que los resultados obtenidos son confiables y precisos, permitiéndonos realizar el proceso de ETL basándonos en lo descubierto durante esta fase.

Llegado el proceso ETL (Extract, Transform, Load) realizamos la gestión de los datos para la preparación e integración de la fuente de datos suministrada, con el objetivo de cargarlos en el Datawarehouse creado anteriormente. Estas operaciones fueron realizadas con la herramienta de código abierto Pentaho Data Integration (PDI). Esta herramienta nos permitió que los datos se someten a una serie de transformaciones, permitiendo crear las dimensiones, la tabla de hechos, realizar la correspondiente limpieza de datos, el filtrado de valores, la normalización de los datos ayudando a reducir la redundancia y optimizar la carga y el almacenamiento de los datos.

Una vez finalizado el proceso ETL, los datos están listos para ser utilizados en el proceso de visualización. Al utilizar la herramienta Microsoft Power BI para diseñar y desarrollar el dashboard, nos permite unificar los datos de las dimensiones y la tabla de hechos de nuestro Datawarehouse. Una vez cargadas las tablas la herramienta nos ofrece diversas opciones de visualización para crear nuestros informes empresariales.

El diseño del dashboard implementado facilita la navegación a través de los informes, segmentación de datos, y nos muestran información relevante sobre nuestra problemática en cuestión. Al utilizar diversos tipos de gráficos, como barras, líneas, circulares y treemaps nos permite presentar la información de manera clara y visualmente atractiva incluso ofreciendo la opción de una difusión de la información de manera interactiva a través de un sitio web.

La implementación de herramientas como Pentaho Data Integration y Python, junto con el uso de Power BI para el análisis y visualización de datos, demuestran la aplicación de tecnologías y metodologías avanzadas aprendidas para lograr los objetivos establecidos por la empresa.

Bibliografía

Business Intelligence

Metodologías para la implementación de proyectos de inteligencia de negocios: un mapeo sistemático de la literatura
(S/f). Researchgate.net. Recuperado el 3 de julio de 2023, de https://www.researchgate.net/publication/349108102_Metodologias_para_la_implementacion_de_proyectos_de_inteligencia_de_negocios_un_mapeo_sistematico_de_la_literatura

Introducción A La Minería de Datos
Introducción A La Minería de Datos: José Hernández Orallo M José Ramírez Quintana Cèsar Ferri Ramírez. (s/f). Scribd. Recuperado el 3 de julio de 2023, de https://es.scribd.com/document/496554106/Jose-Hernandez-Orallo-Cesar-Ferri-Ramirez-Maria-Jose-Ramirez-Quintana-Introduccion-a-La-Mineria-de-Datos-2004-Pearson-Educacion-Libgen-li

KPI's ¿Qué son, para qué sirven y por qué y cómo utilizarlos?
What is a data warehouse? (s/f). Oracle.com. Recuperado el 26 de julio de 2023, de https://www.oracle.com/database/what-is-a-data-warehouse/

Datawarehouse

Data Warehouse Systems Design and Implementation - Alejandro Vaisman Esteban Zimányi
(S/f-b). Researchgate.net. Recuperado el 3 de julio de 2023, de https://www.researchgate.net/publication/265600713_Data_Warehouse_Systems_Design_and_Implementation

The application of volume-outcome contouring in data warehousing
(S/f-c). Researchgate.net. Recuperado el 3 de julio de 2023, de https://www.researchgate.net/publication/51368385_The_application_of_volume-outcome_contouring_in_data_warehousing

Conceptos de Data Warehouse: enfoque de Kimball vs. Inmon
Naeem, T. (2020, febrero 3). Conceptos de data warehouse: Enfoque de Kimball vs. Inmon. Astera. https://www.astera.com/es/type/blog/data-warehouse-concepts/

Difference between Kimball and Inmon
Follow, M. (2020, julio 27). Difference between Kimball and inmon. GeeksforGeeks. https://www.geeksforgeeks.org/difference-between-kimball-and-inmon/

What Is a Data Warehouse?
What is a data warehouse? (s/f). Oracle.com. Recuperado el 26 de julio de 2023, de https://www.oracle.com/database/what-is-a-data-warehouse/

Dashboard

¿Qué es un dashboard y para qué se usa?
Ortiz, D., & Cyberclick. (s/f). ¿Qué es un dashboard y para qué se usa? (2023). Cyberclick.es. Recuperado el 26 de julio de 2023, de https://www.cyberclick.es/numerical-blog/que-es-un-dashboard

Best Chart to Show Trends Over Time
Best chart to show trends over time. (s/f). Ppcexpo.com; PPCexpo. Recuperado el 3 de julio de 2023, de https://ppcexpo.com/blog/best-chart-to-show-trends-over-time

Essential Chart Types for Data Visualization
Sapountzis, M. Y. (2019, septiembre 30). Essential chart types for data visualization. Chartio. https://chartio.com/learn/charts/essential-chart-types-for-data-visualization/

Descripción de un esquema de estrella e importancia para Power BI
Descripción de un esquema de estrella e importancia para Power BI - Power BI. (s/f). Microsoft.com. Recuperado el 3 de julio de 2023, de https://learn.microsoft.com/es-es/power-bi/guidance/star-schema

Power BI: How to Use Data Modeling Features, Relationships, and Hierarchies for Effective Analysis and Visualization
Gabe, A., & Sc., M. (2023, marzo 24). Power BI: How to use data modeling features, relationships, and hierarchies for effective analysis and visualization. Microsoft Power BI. https://medium.com/microsoft-power-bi/power-bi-how-to-use-data-modeling-features-relationships-and-hierarchies-for-effective-8d75d08f4c04