UNIVERSIDAD DE EL SALVADOR
FACULTAD DE INGENIERÍA Y ARQUITECTURA
ESCUELA DE INGENIERÍA DE SISTEMAS INFORMÁTICOS

CURSO DE ESPECIALIZACIÓN EN INGENIERÍA DE DATOS

PROTOTIPO DE DISEÑO E IMPLEMENTACIÓN DE UN MODELO DIMENSIONAL PARA LOS PROCESOS DE NEGOCIO DE VENTAS E INVENTARIOS DE LA TIENDA “ALMACENES EL REY”

PRESENTADO POR:

BONILLA RAMOS, ESPERANZA ARELÍ

GANUZA RAMÍREZ, GLORIA MARÍA

MARTÍNEZ GALDÁMEZ, LILIAN PATRICIA

PARA OPTAR AL TITULO DE:

INGENIERO(A) DE SISTEMAS INFORMÁTICOS

CIUDAD UNIVERSITARIA, ENERO 2023

UNIVERSIDAD DE EL SALVADOR

RECTOR:

MSC. ROGER ARMANDO ARIAS ALVARADO

SECRETARIO GENERAL:

MSC. FRANCISCO ANTONIO ALARCÓN SANDOVAL

FACULTAD DE INGENIERÍA Y ARQUITECTURA

DECANO:

PHD. EDGAR ARMANDO PEÑA FIGUEROA

SECRETARIO:

ING. JULIO ALBERTO PORTILLO

ESCUELA DE INGENIERÍA DE SISTEMAS INFORMÁTICOS

DIRECTOR:

ING. RUDY WILFREDO CHICAS VILLEGAS

UNIVERSIDAD DE EL SALVADOR

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA DE INGENIERÍA DE SISTEMAS INFORMÁTICOS

Trabajo de Graduación previo a la opción al Grado de:

INGENIERO(A) DE SISTEMAS INFORMÁTICOS

Título:

PROTOTIPO DE DISEÑO E IMPLEMENTACIÓN DE UN MODELO DIMENSIONAL PARA LOS PROCESOS DE NEGOCIO DE VENTAS E INVENTARIOS DE LA TIENDA "ALMACENES EL REY"

Presentado por:

BONILLA RAMOS, ESPERANZA ARELÍ

GANUZA RAMÍREZ, GLORIA MARÍA

MARTÍNEZ GALDÁMEZ, LILIAN PATRICIA

Trabajo de Graduación Aprobado por:

Docente Asesor:

ING. RENE FABRICIO QUINTANILLA GOMEZ

SAN SALVADOR, ENERO 2023

Trabajo de Graduación Aprobado por:

Docente Asesor:

ING. RENE FABRICIO QUINTANILLA GOMEZ

Índice

Índice de Tablas

Índice de Figuras

1. Introducción

Después de implementar sistemas para automatizar sus procesos, las empresas ahora cuentan con una gran cantidad de datos, cantidad que día con día aumentan. Estos datos se convierten en un pilar fundamental en la toma de decisiones gerenciales.

ALMACENES EL REY, es una empresa que vende sus productos en línea, por lo que genera una gran cantidad de información diaria, para analizar la data, que, en los meses de Julio del 2022 hasta el mes de noviembre del 2022, se generó una cantidad de 13.33MB, se implementará una solución de Big Data, que permitirá que la empresa tome decisiones a partir de sus datos, lo que ayudará en el proceso de planificación de inventario, por ejemplo.

En el precedente documento se puede observar la planificación del proyecto, también la definición oficial de los principales requerimientos de negocio y principalmente se pretende mostrar el ciclo de vida de desarrollo de Data Warehouse para apoyar a los procesos de negocio de la empresa ALMACENES EL REY.

Para el almacén masivo de datos se usará un Data Lake en AWS, el cual dividirá y organizará la información en tres capas principales. Una vez establecidos los requerimientos se prosigue con un perfilado de datos para utilizarlos en el modelo dimensional.

Se muestra un diseño del modelo dimensional en el cual se aplica el proceso de negocio, se definen de granularidades, identifican dimensiones y determinan métricas. A continuación, se presenta el mapeo de datos para cada uno de los componentes del modelo dimensional.

Definido el modelo se realiza la construcción de los correspondientes ETL’s creados en la herramienta Talend. Teniendo los datos ya trasformados y limpios se podrán consumir para la presentación de los resultados al negocio, para esto se hará uso de la herramienta Power BI, la cual permite construir reportes que satisfacen las preguntas del negocio.

2. Objetivos

a. General

Diseñar un modelo dimensional mediante el análisis de la base datos transaccional y el software de operaciones de la tienda en línea Almacenes el Rey, con ello, brindar una solución a las métricas establecidas para los procesos de negocio de venta e inventario.

b. Específicos

- Analizar el software y la base de datos transaccional utilizada por el negocio y la lógica empleada en el manejo de los procesos venta e inventario en el aplicativo, identificando las entidades relevantes para construir la solución.
- Crear un modelo dimensional que se adecúe a las métricas definidas por el usuario y al nivel de granularidad acordado para visualizar los resultados finales.
- Desarrollar los procesos ETL en la herramienta Open Talend Studio para llevar a cabo la extracción, limpieza y procesamiento de información transaccional.
- Hacer uso de componentes S3 de Amazon Web Services para albergar los datos procesados en sus distintas fases.
- Implementar el modelo dimensional creado en Amazon Redshift.
- Realizar la carga de datos para las dimensiones y tablas de hechos en Amazon Redshift.
- Elaborar los reportes y dashboards requeridos para la presentación de resultados finales al usuario con la herramienta Power BI, utilizando como insumo los datos alojados en Redshift.

3. Marco Teórico

3.1 Negocio

Es decir, un negocio es una actividad económica que busca tener utilidades principalmente a través de la venta o intercambio de productos o servicios que satisfagan las necesidades de los clientes. Puede incluir una o varias etapas de la cadena de producción tales como: extracción de recursos naturales, fabricación, distribución, almacenamiento, venta o reventa.

Algunas veces se utiliza el término negocio para designar el local comercial donde se vende algún bien o servicio como un restaurante, una tienda de ropa, una farmacia, etc.

Objetivo del negocio

El principal objetivo del negocio es el lucro, esto es, obtener ganancias. Cuando una organización no busca el lucro no se puede hablar de que sus actividades son un negocio aun cuando presente características similares. Así, por ejemplo, los servicios gubernamentales relacionados a trámites burocráticos (como obtención de documentos de conducir, solicitud de residencia, solicitud de convalidación de títulos, etc.) no corresponden a un negocio aun cuando se entrega un servicio y este implica un cobro[1].

3.2 Inteligencia de negocios

La inteligencia de negocios se define como la habilidad corporativa para tomar decisiones. Esto se logra mediante el uso de metodologías, aplicaciones y tecnologías que permiten reunir, depurar, transformar datos, y aplicar en ellos técnicas analíticas de extracción de conocimiento, los datos pueden ser estructurados para que indiquen las características de un área de interés, generando el conocimiento sobre los problemas y oportunidades del negocio para que puedan ser corregidos y aprovechados respectivamente.

Implementar herramientas de BI dentro de la organización permite soportar las decisiones que se toman; al nivel interno ayuda en la gestión del personal y del lado externo produce ventajas sobre sus competidores[2]

Durante el transcurso de la historia ha existido el concepto de inteligencia de negocio, pero es hasta hoy en día que ha comenzado el auge de este tema. La inteligencia de negocio

[1] Paula Nicole Roldán, 31 de julio, 2017. Negocio. Economipedia.com

[2] Rosado Gómez, A. A., & Rico Bautista, D. W. (2010). Inteligencia de negocios: Estado del arte. *Scientia Et Technica*, *1*(44), 321–326. https://doi.org/10.22517/23447214.1803

brinda apoyo a los diferentes niveles organizacionales, siendo estos niveles los mostrados en la *figura 1*.

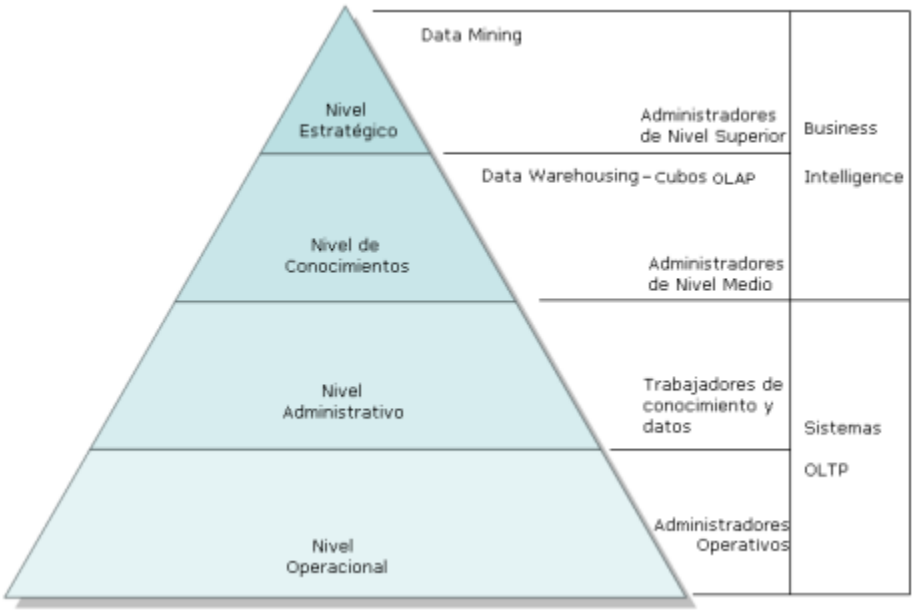


Figura 1: Niveles organizacionales

• Nivel operacional: Se utilizan sistemas de información que monitorean las actividades y transacciones elementales de la organización. Son sistemas que han cobrado un auge

• Nivel de conocimiento: En este nivel se encuentran los trabajadores de conocimiento y datos, cubriendo el núcleo de operaciones tradicionales de captura masiva de datos y servicios básicos de tratamiento de datos con tareas predefinidas

• Nivel de administración: Se realizan tareas de administradores de nivel intermedio apoyando las actividades de análisis, de seguimiento, de control y toma de decisiones, realizando consultas sobre información almacenada en el sistema, proporcionando informes y facilitando la gestión de la información por parte de los niveles intermedios.

• Nivel estratégico: Tiene como objetivo realizar las actividades de planificación a largo plazo, tanto del nivel de administración como de los objetivos que la empresa posee.

Bajo el nombre de Business Intelligence (Inteligencia de Negocios) se agrupan diferentes acrónimos, herramientas y disciplinas que apuntan a dar soporte a la tarea de toma de decisiones. Fundamentalmente podemos mencionar:

• Data warehousing: Su traducción en castellano “almacenes de datos”. Se basan en estructuras multidimensionales en las que se almacena la información calculando

previamente todas las combinaciones de todos los niveles de todas las aperturas de análisis.

• Data mart: Constituye una parte de un data warehouse. Si un data warehouse está formado por todos los procesos de la organización, un Data mart constituye un determinado proceso. Por ejemplo, podríamos tener un Data mart para finanzas, otro para logística. Pueden ser elaborados independiente o no de un data warehouse.

• Data mining: Asociado al nivel organizacional estratégico y tiene por objetivo eliminar los errores cometidos por las personas al analizar los datos debido a prejuicios y dejar que sean los datos que demuestren hechos reales.

• Tecnología OLAP (online analytical process): Es la tecnología que permite aprovechar como está estructurada la información de un Data mart o un Data warehouse. Fundamentalmente es una tecnología que permite analizar información dinámicamente a los niveles tácticos y estratégicos basados en Cubos que contienen las medidas y las dimensiones[3]

3.3 Tecnología OLAP

OLAP es la repuesta más próxima al análisis de consultas de naturaleza dimensional. Es una parte de la amplia categoría de inteligencia de negocios, que también incluye ETL, reportes relacionales y minería de datos. El concepto de OLAP también puede ser descrito como el análisis rápido de información multidimensional compartida (FASMI). Posee aspectos de bases de datos relacionales en navegación y jerarquía, optimizan el procesamiento más que un sistema de tipo relacional. Las Bases de datos OLAP emplean modelos de datos multidimensionales, brindando la realización de análisis complejos con una ejecución rápida en tiempo. Características por resaltar de un sistema OLAP:

- Rapidez: Significa que los sistemas están enfocados a brindar respuesta a usuarios entre aproximadamente 5 segundos, con análisis simples no más de 1 segundo y muy pocos más de 20 segundos.
- Análisis: El sistema puede hacer frente a cualquier lógica de negocios y análisis estático que es relevante para la aplicación y los usuarios, manteniendo la simpleza suficiente para los usuarios finales.
- Compartido: Se debe implementar toda la seguridad requerida para proveer confidencialidad y si se necesita múltiple acceso para escritura y actualización concurrente es necesario establecer la seguridad a un apropiado nivel.
- Multidimensional: El sistema debe proveer una vista conceptual multidimensional de los datos, incluyendo soporte completo para jerarquización y múltiple

[3] René Cuchillas, Oscar Hernández, Yuri Mejía, Héctor Silva (2010). Tesis DESARROLLO DE UN "DATA WAREHOUSE" PARA EL PROCESO DE DENUNCIAS DE LA DEFENSORÍA DEL CONSUMIDOR

jerarquización, ya que es la manera más lógica de analizar el negocio y la organización.

- Información: Incluye todos los datos e información derivada, de donde sea y cuan relevante sea para la aplicación.

Tipos de OLAP

Existen 3 tipos de OLAP. Cada tipo posee ciertos benéficos y así mismo desventajas.

- OLAP Multidimensional: MOLAP es la forma clásica de OLAP y es a veces referido solo como OLAP. MOLAP utiliza bases de datos estructuradas que son generalmente optimizadas para recuperación de datos. A diferencia de la base de datos relacional, esta forma de almacenamiento esta optimizada para la aceleración de cálculos. Frecuentemente son optimizadas para la recuperación de patrones a lo largo del acceso jerarquizado. La dimensión de cada cubo son típicamente atributos como periodos de tiempo, ubicación y productos o código de cuenta. MOLAP trabaja mucho mejor en sistemas de datos pequeños, pues es más fácil calcular las agregaciones y retornar respuesta utilizando menos espacio de almacenamiento.
- OLAP Relacional: ROLAP trabaja directamente con bases de datos relacionales. Sin embargo, en organizaciones con procesamiento de un alto volumen de datos es difícil de implementar, frecuentemente este tipo de OLAP es ignorado.
- OLAP Hibrido: No existe un acuerdo formal sobre lo que constituye un OLAP hibrido, exceptuando que la base de datos divide los datos entre el relacional y almacenamiento especializado. HOLAP se encuentra entre los otros dos tipos de OLAP, más sin embargo puede realizar procesamientos rápidos y escalables.

3.4 Data Warehouse

“Un Data Warehouse es un almacén único, completo y consistente de datos obtenidos de una variedad de fuentes y puestos a disposición de los usuarios finales de una manera que puedan entender y utilizar en un contexto empresarial” (Cuellar, 2011).

“El Data Warehounsing es un proceso, no un producto, para ensamblar y administrar datos de diversas fuentes con el fin de obtener una visión única y detallada de una parte o de toda una empresa” (Devlin, 2011).

De lo mencionado por los autores citados anteriormente se puede concluir que el Data Warehousing es el proceso mediante el cual los datos son transformados y agrupados dentro de un almacén de datos, permitiendo una fusión de los mismos los cuales provienen de diversas fuentes hacia un solo contexto, lo cual permite un acceso rápido a la información y por consiguiente a la toma de decisiones.

En el contexto informático, Data Warehouse en español almacén de datos es una colección de información corporativa, derivada directamente del sistema operativo y algunas fuentes

externas de datos, su propósito específico es apoyar las decisiones empresariales. Se trata de un expediente completo de una empresa, más allá de la información transaccional y operacional, almacenada en una base de datos, diseñada para favorecer el análisis y la divulgación eficiente de datos (Peña J & Súarez, 2008)[4]. En la *figura 2* se puede observar un diagrama de la arquitectura de data Warehouse.

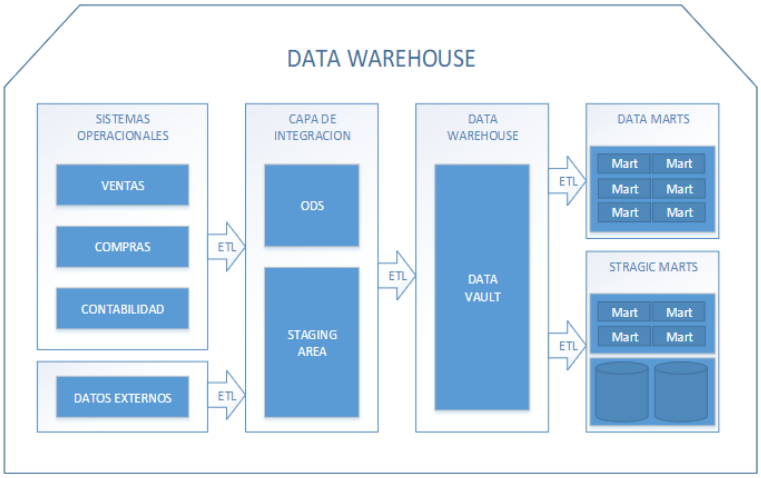


Figura 2. Arquitectura de un Data Warehouse

3.4.1 Modelo de Inmon[5]

La arquitectura de Inmon consiste en todos los sistemas de información y sus respectivas bases de datos de una determinada organización. Él llama a este gigante "Fabrica de Información Organizacional", GIF por sus siglas en inglés (Corporate Information Factory).

Inmon divide el ambiente de bases de datos organizacionales en cuatro niveles:

- Operacional.
- Data warehouse atómico.
- Departamental.
- Individual.

Los últimos tres niveles conforman el data warehouse. El primer nivel contiene datos de sistemas heredados y otros sistemas transaccionales. Este nivel es el que da soporte a las operaciones diarias de la organización, en otras palabras, el primer nivel soporta todo el procesamiento transaccional. En los sistemas transaccionales, la información es manipulada considerablemente y luego es movida la data warehouse atómico.

Inmon utiliza un ejemplo para mostrar la diferencia entre los datos operacionales y los datos almacenados en el data warehouse atómico. En el ejemplo, la entidad es un cliente y sus

[4] Peñafiel, G. E. S., Yánez, V. M. Z., Guamán, K. P. M., & Padilla, L. M. T. (2019). Análisis de metodologías para desarrollar Data Warehouse aplicado a la toma de decisiones. Ciencia digital, 3(3.4.), 397-418.

[5] René Cuchillas, Oscar Hernández, Yuri Mejía, Héctor Silva (2010). Tesis DESARROLLO DE UN "DATA WAREHOUSE" PARA EL PROCESO DE DENUNCIAS DE LA DEFENSORÍA DEL CONSUMIDOR

atributos de más interés de un cliente es su record crediticio. La base de datos operacional contiene el actual record crediticio e información relacionada a este (como balances de préstamos, direcciones, etc) en un solo registro. En contraste, el data warehouse atómico, contiene el histórico del record crediticio para este cliente, totalizado por año, con un registro por año.

Los usuarios individuales crean el cuarto y último nivel de la arquitectura cuando crean sets de datos heurísticos y ad hoc como parte de soporte a la toma de decisiones. Este nivel tiende a ser temporal y a ser almacenado en la computadora personal del usuario. Por ejemplo, un usuario que trabaja en el departamento de créditos puede pedir un reporte para visualizar todas las cuentas que han sido robadas en los últimos tres años.

Si la base de datos del departamento no cuenta con la información detallada histórica, es posible realizar la consulta al data warehouse atómico. Las consultas al data warehouse atómico generalmente se realzan atreves del departamento de TI. Inmon argumenta que es compensado el esfuerzo inicial requerido para construir el data warehouse atómico, ya que permite la creación de múltiples bases de datos departamentales sin correr el riesgo de generar datos inconsistentes entre ellos. Esto se logra utilizando un modelo de datos de tres niveles.

Modelo de Datos de Tres Niveles

Inmon propone el modelado de datos de tres niveles. El primero lo conforman los DER (Diagrame Entidad Relación). Al igual que en el desarrollo de bases de datos operacionales, el DER es utilizado para determinar y refinar las entidades, sus atributos y relaciones entre ellas. El equipo de desarrollo crea un conjunto de DERs por cada departamento que se pretende cubrir con el Data Warehouse. El DER corporativo es la suma de todos los DER departamentales.

El segundo nivel, establece una serie de datos (DIS, data ítem set) para cada departamento. Al igual que el primer nivel, la suma de los DIS departamentales conforma el DIS empresarial. Este nivel cuenta con cuatro construcciones:

- Una agrupación de datos primaria.
- Una agrupación de datos secundaria.
- Un conector, que representa las relaciones entre los principales datos de cada área grande.
- El tipo de datos

Un aspecto crítico de este nivel es que la agrupación de datos primaria existe una sola vez para cada área grande. Esto significa que un DER creado en el modelo de datos del primer nivel es la base para un DIS en el segundo nivel. También muestra como las diferentes vistas de los usuarios son combinadas en un DER y DIS organizacional. En el DIS cada rectángulo representa una tabla lógica ya sea en el DIS departamental o en el organizacional. Las

relaciones entre estas tablas son las mismas que relacionan las entidades en los DER. Los rectángulos a la derecha de un determinado DIS representan la agrupación de datos secundaria.

Para aclarar esto Inmon utiliza un ejemplo. En un banco, la entidad "cliente" genera una agrupación de datos primaria con la cuenta. La "cuenta" puede tener diferentes formas, como préstamos y ahorros (grupo secundario). Las relaciones muestran que un cliente puede tener diferentes cuentas. Por último, cada cuenta puede tener datos generados por actividades similares, como depósitos en ATM, depósitos bancarios, retiros en ATM y agencias bancarias. Estos son ejemplos del tipo de datos.

El último nivel del modelo de Inmon es el físico. "El modelo físico es creado desde el segundo nivel del modelo de datos simplemente extendiéndolo, incluyendo las llaves y características físicas del modelo. En este punto, el nivel físico del modelo de datos se asemeja a una serie de tablas, muchas veces llamadas tablas relacionadas"6 . Inmon explica varias técnicas para optimizar el rendimiento del data warehouse en los niveles departamentales y atómicos.

Aunque las técnicas no son muy familiares su propósito es optimizar el rendimiento de E/S al igual que en los sistemas de base de datos transaccionales. La mayoría de las técnicas involucran la desnormalización de las tablas.

Existen muchas razones por las cuales se deben de desnormalizar las tablas en el nivel físico. Por ejemplo, registros en el data warehouse atómico son rara vez actualizados puesto que son datos históricos. Esto hace posible colocar los datos físicamente en formas que no serían funcionales para la base de datos transaccional ya que su actualización es constante. Una vez se ha completado el modelo de datos de tres niveles, el desarrollo del data warehouse comienza.

3.4.2 Metodología de Kimball

La metodología se basa en lo que Kimball denomina Ciclo de Vida Dimensional del Negocio (Business Dimensional Lifecycle) (Kimball et al 98, 08, Mundy & Thornthwaite 06). Este ciclo de vida del proyecto de DW, está basado en cuatro principios básicos:

- Centrarse en el negocio: Hay que concentrarse en la identificación de los requerimientos del negocio y su valor asociado, y usar estos esfuerzos para desarrollar relaciones sólidas con el negocio, agudizando el análisis del mismo y la competencia consultiva de los implementadores.
- Construir una infraestructura de información adecuada: Diseñar una base de información única, integrada, fácil de usar, de alto rendimiento donde se reflejará la amplia gama de requerimientos de negocio identificados en la empresa.

- Realizar entregas en incrementos significativos: crear el almacén de datos (DW) en incrementos entregables en plazos de 6 a 12 meses. Hay que usa el valor de negocio de cada elemento identificado para determinar el orden de aplicación de los incrementos. En esto la metodología se parece a las metodologías ágiles de construcción de software.
- Ofrecer la solución completa: proporcionar todos los elementos necesarios para entregar valor a los usuarios de negocios. Para comenzar, esto significa tener un almacén de datos sólido, bien diseñado, con calidad probada, y accesible. También se deberá entregar herramientas de consulta ad hoc, aplicaciones para informes y análisis avanzado, capacitación, soporte, sitio web y documentación.

La construcción de una solución de DW/BI es sumamente compleja, y Kimball nos propone una metodología que nos ayuda a simplificar esa complejidad. Las tareas de esta metodología (ciclo de vida) las podemos observar en la *figura 3.* De las tareas podemos observar que primero, hay que resaltar el rol central de la tarea de definición de requerimientos. Los requerimientos del negocio son el soporte inicial de las tareas subsiguientes. También tiene influencia en el plan de proyecto. En segundo lugar, podemos ver tres rutas o caminos que se enfocan en tres diferentes áreas:

1. **Tecnología** (Camino Superior). Implica tareas relacionadas con software específico, por ejemplo, Microsoft SQL Analysis Services.
2. **Datos** (Camino del medio). En la misma diseñaremos e implementaremos el modelo dimensional, y desarrollaremos el subsistema de Extracción, Transformación y Carga (Extract, Transformation, and Load - ETL) para cargar el DW.
3. **Aplicaciones de Inteligencia de Negocios** (Camino Inferior). En esta ruta se encuentran tareas en las que diseñamos y desarrollamos las aplicaciones de negocios para los usuarios finales[6].

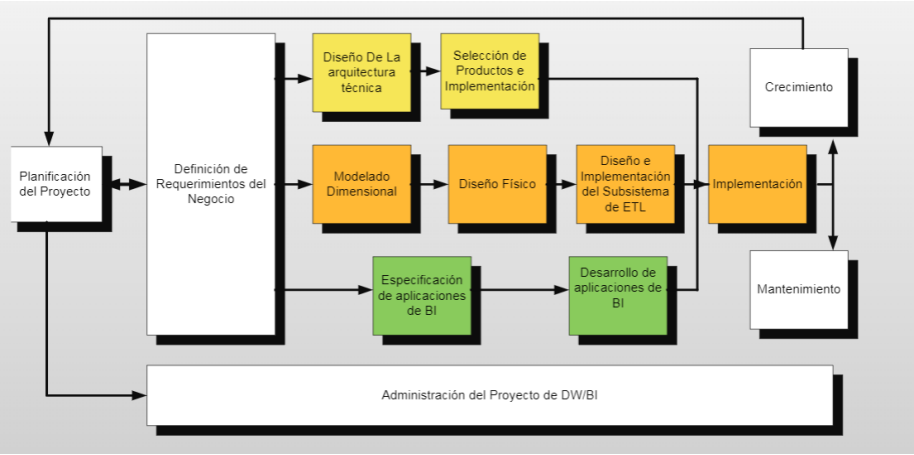


Figura 3: Ciclo de vida Kimball

[6] Rivadera, G. R. (2010). La metodología de Kimball para el diseño de almacenes de datos (Data warehouses).

Las tareas de esta metodología (ciclo de vida) se describen a continuación:

Planificación del Proyecto

En este proceso se determina el propósito del proyecto de DW/BI, sus objetivos específicos y el alcance del mismo, los principales riesgos y una aproximación inicial a las necesidades de información.

Esta tarea incluye las siguientes acciones típicas de un plan de proyecto:

- Definir el alcance (entender los requerimientos del negocio).
- Identificar las tareas
- Programar las tareas
- Planificar el uso de los recursos.
- Asignar la carga de trabajo a los recursos
- Elaboración de un documento final que representa un plan del proyecto.

Además, en esta parte definimos cómo realizar la administración o gestión de esta subfase que es todo un proyecto en si mismo, con las siguientes actividades:

- Monitoreo del estado de los procesos y actividades.
- Rastreo de problemas
- Desarrollo de un plan de comunicación comprensiva que direccione la empresa y las áreas de TI

Definición de Requerimientos del Negocio

La definición de requerimientos, es un proceso de entrevistar al personal de negocio y técnico, aunque siempre conviene, tener un poco de preparación previa. En esta tarea, se debe aprender sobre el negocio, los competidores, la industria y los clientes del mismo. Se debe dar una revisión a todos los informes posibles de la organización; rastrear los documentos de estrategia interna; entrevistar a los empleados, analizar lo que se dice en la prensa acerca de la organización, la competencia y la industria y se deben conocer los términos y la terminología del negocio.

Se sugiere entrevistar al personal que se encuentra en los cuatro grupos que se mencionan a continuación:

- El directivo responsable de tomar las decisiones estratégicas.
- Los administradores intermedios y de negocio responsables de explorar alternativas estratégicas y aplicar decisiones

- El personal de sistemas, si existe (estas son las personas que realmente saben qué tipos de problemas informáticos y de datos existen en la organización)
- El personal que se entrevista por razones políticas.

Modelado Dimensional

Es un proceso dinámico y altamente iterativo. Comienza con un modelo dimensional de alto nivel obtenido a partir de los procesos priorizados y descritos en la tarea anterior, y El proceso iterativo consiste en cuatro pasos:

a. Elegir el proceso de negocio: que consiste en, elegir el área a modelizar. Esta es una decisión de la dirección, y depende fundamentalmente del análisis de requerimientos y de los temas analíticos anotados en la etapa anterior.
b. Establecer el nivel de granularidad: La granularidad significa especificar el nivel de detalle. La elección de la granularidad depende de los requerimientos del negocio y lo que es posible a partir de los datos actuales. La sugerencia general es comenzar a diseñar el DW al mayor nivel de detalle posible, ya que se podrían realizar agrupamientos posteriores, al nivel deseado.
c. Elegir las dimensiones: Las dimensiones surgen naturalmente de las discusiones del equipo, y facilitadas por la elección del nivel de granularidad y de la matriz de procesos/dimensiones. Las tablas de dimensiones tienen un conjunto de atributos (generalmente textuales) que brindan una perspectiva o forma de análisis sobre una medida en una tabla hechos. Una forma de identificar las tablas de dimensiones es que sus atributos son posibles candidatos para ser encabezado en los informes, tablas pivot, cubos, o cualquier forma de visualización, unidimensional o multidimensional.
d. Identificar medidas y las tablas de hechos: Este paso, consiste en identificar las medidas que surgen de los procesos de negocios. Una medida es un atributo (campo) de una tabla que se desea analizar, sumando o agrupando sus datos y usando los criterios de corte conocidos como dimensiones. Las medidas habitualmente se vinculan con el nivel de granularidad del punto 2, y se encuentran en tablas que denominamos tablas de hechos (fact en inglés). Cada tabla de hechos tiene como atributos una o más medidas de un proceso organizacional, de acuerdo a los requerimientos. Un registro contiene una medida expresada en números, como ser cantidad, tiempo, dinero, etc., sobre la cual se desea realizar una operación de agregación (promedio, conteo, suma, etc.) en función de una o más dimensiones. La granularidad, en este punto, es el nivel de detalle que posee cada registro de una tabla de hechos.

En la *figura 4* podemos observar un modelo dimensional básico.

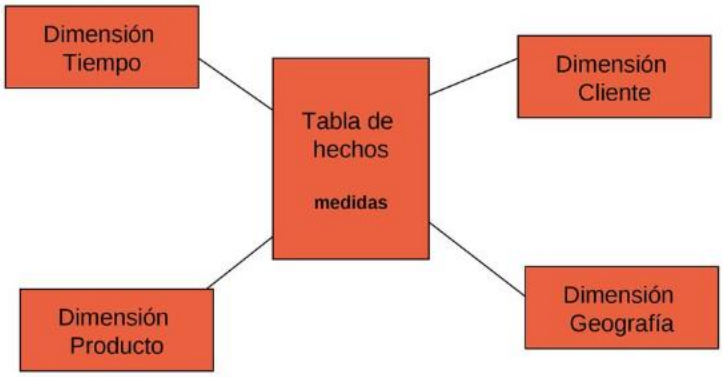


Figura 4: Modelamiento dimensional

Diseño Físico

En esta tarea, se contestan las siguientes preguntas:
¿Cómo puede determinar cuán grande será el sistema de DW/BI?
¿Cuáles son los factores de uso que llevarán a una configuración más grande y más compleja?
¿Cómo se debe configurar el sistema?
¿Cuánta memoria y servidores se necesitan? ¿Qué tipo de almacenamiento y procesadores?
¿Cómo instalar el software en los servidores de desarrollo, prueba y producción?
¿Qué necesitan instalar los diferentes miembros del equipo de DW/BI en sus estaciones de trabajo?
¿Cómo convertir el modelo de datos lógico en un modelo de datos físicos en la base de datos relacional?
¿Cómo conseguir un plan de indexación inicial?
¿Debe usarse la partición en las tablas relacionales?

Diseño e Implementación del subsistema de Extracción, Transformación y Carga (ETL

El subsistema de Extracción, Transformación y Carga (ETL) es la base sobre la cual se alimenta el Data warehouse. Si se diseña adecuadamente, puede extraer los datos de los sistemas de origen de datos, aplicar diferentes reglas para aumentar la calidad y consistencia de los mismos, consolidar la información proveniente de distintos sistemas, y finalmente cargar (grabar) la información en el DW en un formato acorde para la utilización por parte de las herramientas de análisis.

Carga Incremental.

La carga incremental toma como base una carga full, luego se define un periodo de tiempo entre cada carga incremental. En cada ejecución de una carga incremental se toman los

datos que han sido creados o modificados desde la última carga incremental y los actualiza en sus respectivas dimensiones o fact tables

Carga Full.

La carga full difiere significativamente de la carga incremental, esta se ejecuta una única vez al inicio de la vida del modelo, incluye datos desde una fecha establecida en un SLA hasta T – 1, donde T puede ser una hora, día, semana, etc. La mayor preocupación durante la carga full es el volumen de datos, a veces miles de veces mayor a una carga incremental diaria. En ocasiones la carga full puede durar varios días, lo cual es usualmente tolerable

Implementación

La implementación representa la convergencia de la tecnología, los datos y las aplicaciones de usuarios finales accesible desde el escritorio del usuario del negocio. Existen varios factores extras que aseguran el correcto funcionamiento de todas estas piezas, entre ellos se encuentran la capacitación, el soporte técnico, la comunicación y las estrategias de feedback.

Mantenimiento y Crecimiento del Data Warehouse

Para administrar el entorno del Data Warehouse existente es importante enfocarse en los usuarios de negocio, los cuales son el motivo de su existencia, además de gestionar adecuadamente las operaciones del Data Warehouse, medir y proyectar su éxito y comunicarse constantemente con los usuarios para establecer un flujo de retroalimentación, En esto consiste el Mantenimiento. Finalmente, es importante sentar las bases para el crecimiento y evolución del Data Warehouse en donde el aspecto clave es manejar el crecimiento y evolución de forma iterativa utilizando el Ciclo de Vida propuesto, y establecer las oportunidades de crecimiento y evolución en orden por nivel prioridad.

Especificación de aplicaciones de BI

En esta tarea se proporciona, a una gran comunidad de usuarios una forma más estructurada y por lo tanto, más fácil, de acceder al almacén de datos. Se proporciona este acceso estructurado a través de lo que llamamos, aplicaciones de inteligencia de negocios (Business Intelligence Aplications). Las aplicaciones de BI son la cara visible de la inteligencia de negocios: los informes y aplicaciones de análisis proporcionan información útil a los usuarios. Las aplicaciones de BI incluyen un amplio espectro de tipos de informes y herramientas de análisis, que van desde informes simples de formato fijo, a sofisticadas

aplicaciones analíticas que usan complejos algoritmos e información del dominio. Kimball divide a estas aplicaciones en dos categorías basadas en el nivel de sofisticación, y les llama:

a. Informes estándar: son informes relativamente simples, de formato predefinido, y parámetros de consulta fijos, proporcionan a los usuarios un conjunto básico de información acerca de lo que está sucediendo en un área determinada de la empresa y se utilizan día a día.
b. Aplicaciones analíticas: son más complejas que los informes estándar. Estas aplicaciones pueden incluir algoritmos y modelos de minería de datos, que ayudan a identificar oportunidades o cuestiones subyacentes en los datos, y el usuario puede pedir cambios en los sistemas transaccionales basándose en los conocimientos obtenidos del uso de la aplicación de BI. Algunas aplicaciones analíticas comunes incluyen:
    - Análisis de la eficacia de las promociones
    - Análisis de rutas de acceso en un sitio Web
    - Análisis de afinidad de programas
    - Planificación del espacio en espacios comerciales
    - Detección de fraudes
    - Administración y manejo de categorías de productos

Diseño de la Arquitectura Técnica

El área de arquitectura técnica cubre los procesos y herramientas que se aplican a los datos. En el área técnica existen dos conjuntos que tienen distintos requerimientos, brindan sus propios servicios y componentes de almacenaje de datos, por lo que se consideran cada uno aparte: El back room (habitación trasera) y el front room (habitación frontal). El back room es el responsable de la obtención y preparación de los datos, por lo que también se conoce como adquisición de datos y el front room es responsable de entregar los datos a la comunidad de usuario y también se le conoce como acceso de datos[7].

[7] Rivadera, G. R. (2010). La metodología de Kimball para el diseño de almacenes de datos (Data warehouses).

4. Desarrollo

a. Capítulo I: Especificación de proyecto

1. Situación actual

1.1 Antecedentes

1.1.1 Aclaraciones iniciales

En las organizaciones, los datos y la información que estas generan son los activos más valiosos con lo que se pueda contar, ya que más allá de simplemente llevar un registro de las actividades y transacciones realizadas en la misma, son grandes fuentes de datos útiles para realizar análisis y mediciones de su situación actual, y en base a ello, tomar decisiones estratégicas que ayuden a las empresas a crecer y mejorar.

La ingeniería de datos es una disciplina que se encarga de tratar esta información; recolectando, trasladando, explorando y validando los datos de las empresas, y con ello, construir modelos de datos, Data Warehouses, que garanticen disponibilidad, consistencia, mantenimiento, seguridad y limpieza de datos para ser utilizados en la generación de documentos y reportes óptimos que apoyen a la gerencia y los altos mandos.

En el curso de especialización en ingeniería de datos que se realizó a lo largo del año 2022, se estudiaron varios factores repartidos en 4 módulos, abarcando desde Datawarehousing, Big Data y Cloud Computing, procesamiento masivo de datos y visualización de datos, las cuales se ponen en práctica en este proyecto.

Dentro del curso de especialización se nos establecieron e informaron las diferentes formas de evaluación que se llevaron a cabo a lo largo del año, entre ellas, la realización de un trabajo de aplicación donde pueda reflejarse los conocimientos adquiridos a lo largo del curso. Para efectos de estudio y aprendizaje, este trabajo de aplicación es didáctico, asignado por la catedra.

En otras palabras, el actual proyecto es un prototipo de solución para la empresa ficticia “Almacenes El Rey” donde se han respetado tablas, estructuras y relaciones de la base de datos alojada en MySQL, así como también el funcionamiento y la lógica de la aplicación Magento, siendo tratada y manejada como una empresa real, lo cual no desmerita el análisis y el proceso realizado para la construcción e implementación de una solución real para las problemáticas planteadas a dicha empresa.

1.1.2 Introducción a la lógica del negocio

Para el desarrollo de este proyecto se analizan los datos brindados por la tienda Almacenes El Rey, la cual se dedica a la venta de ropa para dama, caballeros y niños, así como también

calzados, accesorios y productos digitales como libros, audiolibros, cursos en línea, entre otros. La venta de estos productos se hace a través de su tienda de e-commerce alojada en el software Magento, plataforma de comercio electrónico especializado para compra y venta de productos en línea, y para el almacenamiento y gestión de sus datos hacen uso del gestor de base de datos MySQL.

La tienda Almacenes El Rey cuenta con dos tipos de usuario los cuales tienen diferentes vistas y transacciones. En forma general se dividen en: los clientes y el administrador del sistema.

Para realizar una compra, los clientes deben crear una cuenta con un correo electrónico, además de ingresar otros datos personales que son importantes para su identificación. Creada la cuenta, el cliente puede guardar productos en el carrito de compra, dependiendo si estos se encuentran en existencia y la cantidad mínima y máxima que se puede llevar de un producto. Finalizada su selección, se procede a detallar información referente al total de la compra, el impuesto y el envío del pedido. Es en este punto donde el cliente puede hacer uso de cupones de descuento, si se cuenta con el código correspondiente y se cumple con los requerimientos necesarios para aplicar a ese tipo de descuento. Con todo esto listo, la orden de compra es solicitada.

El cliente también posee con un “perfil de usuario” en el cual puede configurar su cuenta, ver el estado de sus pedidos, los productos digitales que ha adquirido, su lista de deseos, entre otras opciones.

El administrador del sistema es el responsable de manejar y configurar todo lo relacionado a la página web, que abarca desde la administración de los elementos del contenido y el diseño de la tienda, la configuración de las operaciones del sistema y servicios web, configuración de impuestos, moneda, atributos de productos y grupos de clientes, el establecimiento de reglas de precio, cupones y promociones, la creación y definición de categorías y productos, y la administración de las cuentas de los clientes, hasta el manejo de las operaciones relacionadas a las órdenes de venta y manejo de inventario y sus fuentes. Así como también puede generar informes que brindan información sobre los aspectos de la tienda.

Las pedidos u órdenes de venta son recibidas por el administrador quien se encarga de gestionar, facturar, seleccionar la fuente y enviar la mercancía a los clientes o, en otros casos, de cancelar el pedido o crear notas de crédito para proceder a la devolución de estos. El ciclo de vida de las órdenes de compra puede pasar desde pendiente, procesando, en espera, completo, cerrado o cancelado, dependiendo del tipo de transacción que se esté realizando en ese momento al pedido.

El inventario que se maneja en la tienda es de tipo PEPS, (primero en entrar, primero en salir). Para comenzar con su configuración, primero se deben crear las fuentes, estas representan las ubicaciones físicas donde se administra y envía el inventario de productos. Luego se configuran el stock o las existencias, las cuales se les puede asignar desde una o muchas fuentes. El inventario de la tienda Almacenes El Rey solo esta asociar a un solo stock que tiene múltiples fuentes debido a las restricciones del software Magento que establecen que: "*Los Canales de Venta solo se pueden asociar a un Stock. Cada canal de ventas solo puede tener asignado un único stock, y un único stock puede asignarse a varios sitios web*.[8]"

Una vez realizada la configuración anterior, se procede a asignar los productos a las fuentes, esto se realiza al momento de configurar un producto, cuando se crea o se modifica un producto se tiene la opción de incluir su fuente, que pueden ser de 1 a muchas fuentes, se establece la cantidad de artículos que cada fuente le dará y también el costo al que fue comprado de forma individual, este último es independiente a las fuentes.

La tienda maneja dos tipos de descuentos, aquellos dados por promociones, los cuales consisten en la reducción de precio de ciertos productos para un periodo de tiempo determinado, y los cupones, los cuales son limitados y enviados a algunos clientes, y dependiendo del tipo de cupón, se puede crear un código el cual el cliente utiliza para la reducción del precio total de su compra.

La tienda permite la devolución de mercancía a los clientes que soliciten devolver un artículo para reemplazo o reembolsarlo. Dependiendo del motivo del cliente para devolver ciertos artículos, se evalúa si estos deben ir devuelva al inventario o por el contrario se retira del mismo a causa de desperfectos de fábrica o daños. Estas devoluciones no aplican para productos virtuales o descargables. Al realizar la devolución el cliente puede recibir la misma cantidad de dinero que se ocupó en la compra de los artículos o una cantidad diferente a causa de las tarifas de ajuste.

1.2 Descripción del problema

La tienda almacenes el Rey ha solicitado que se realice un análisis de datos a través de un Data Warehouse para los procesos de ventas e inventario. Para los cuales ha establecido los siguientes requerimientos:

Para el proceso de negocio de venta:

- Conocer cuanto representa en términos monetarios la cantidad de producto que se vende en un periodo de tiempo específico para evaluar su rentabilidad en el mercado.

[8] Guía de usuario de Magento, apartado Stocks: https://docs.magento.com/user-guide/v2.3/catalog/inventory-about-sources-stocks.html

- Conocer que productos son más solicitados por sus clientes.
- Conocer cuanto representa una promoción o un cupón de descuento en venta. Así como también, se quiere conocer que descuentos se utilizan más por los clientes, y en base a esto, mejorar sus estrategias de marketing.
- Conocer cuáles son los ingresos reales que obtiene la tienda luego de impuestos, descuentos y devoluciones, y como estas afectan su ganancia final.

Para el proceso de negocio de inventario:

- Conocer la cantidad de producto disponible y como este va cambiando a medida se realizan las diferentes transacciones de compras, ventas y devoluciones.
- Conocer el costo promedio de un producto que se vendió, lo cual servirá para apoyar en el monitoreo del crecimiento o disminución del valor de ese producto en el mercado y evaluar si sigue siendo rentable o no para la tienda.
- Manejo de devoluciones.

Después de establecido lo que la tienda solicita y luego de analizar la base de datos transaccional brindado se establecieron las siguientes métricas.

Métricas establecidas para el proceso de negocio de venta.

Venta.

- Determinar el monto total de ventas por productos en una fecha determinada
- Conocer la cantidad de producto vendido en una fecha determinada
- Determinar el total de las ventas por producto incluyendo descuentos en una fecha determinada
- Determinar el total ganado por un producto luego de incluir impuestos, descuentos y devoluciones

Métricas establecidas para el proceso de negocio de Inventario.

Inventario.

- Determinar la cantidad de producto que se tiene disponible en inventario para una fecha determinada.
- Determinar el costo promedio de un producto para un periodo de tiempo determinado
- Manejar devoluciones de mercadería en inventario.

Para resolver la problemática correctamente y al tratarse de dos áreas diferentes en la tienda, se debe crear dos modelos dimensionales diferentes, uno por cada proceso de negocio.

1.3 Planteamiento del problema

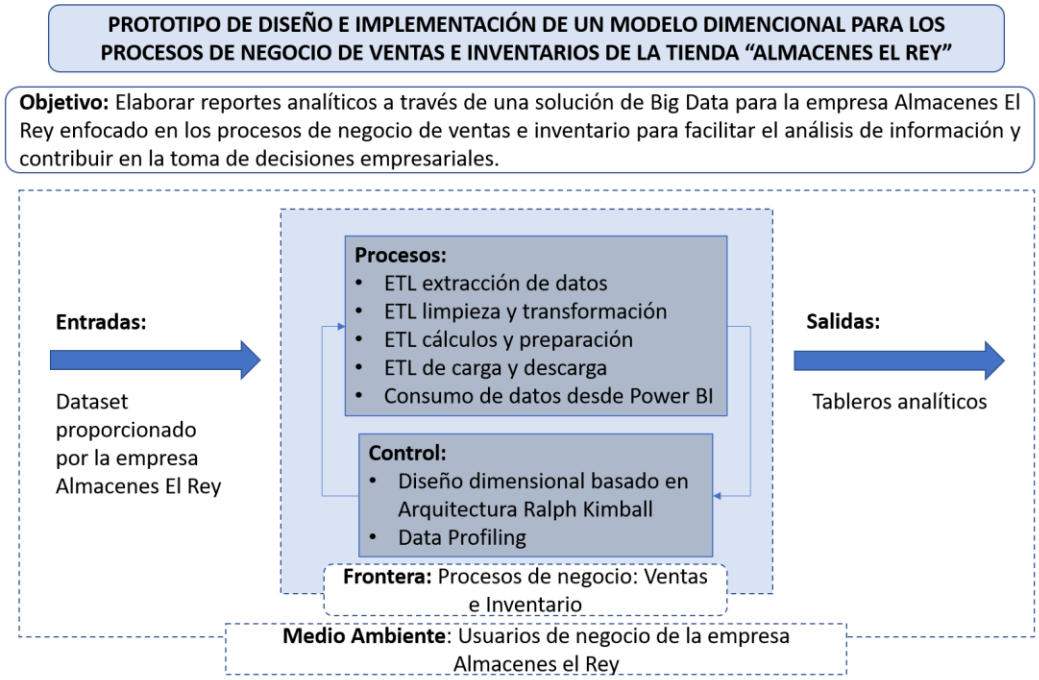


Figura 5: Diagrama de enfoque de sistemas

La *figura 5*, muestra el planteamiento del problema haciendo uso del enfoque de sistemas.

OBJETIVO:
Elaborar reportes analíticos a través de una solución de Big Data para la tienda Almacenes El Rey enfocado en los procesos de negocio de ventas e inventario para facilitar el análisis de información y contribuir en la toma de decisiones empresariales.

ENTRADAS:
Dataset proporcionado por la tienda Almacenes El Rey: Entrada que se refiere al Dataset seleccionado de la base de datos origen proporcionada por la tienda Almacenes El Rey que abarca los procesos de negocio de ventas e inventario, de las cuales se cuenta con su diseño, estructura, relaciones y datos crudos.

PROCESOS:
ETL extracción de datos: Este proceso hace referencia a las ETL utilizadas para la extracción de información desde la base datos origen hasta su almacenamiento en la respectiva zona de datos crudos.

ETL limpieza y transformación: Este proceso hace referencias a las ETL utilizadas para la limpieza de los datos crudos y la transformación de los mismos, siendo extraídos desde la zona de datos crudos y almacenados en la zona de datos en proceso.
ETL cálculos y preparación: Este proceso hace referencia a las ETL utilizadas para realizar cálculos en los datos (donde sea requerido), el filtrado y realización de algunas configuraciones de los datos antes de ser consultados.
ETL de carga y descarga: Este proceso hace referencia a las ETL utilizadas para realizar la carga de archivos .CSV al repositorio de AWS S3 y la descarga desde ese mismo repositorio para las demás fases del análisis. Así como también las ETL para cargar los datos listos a subirse a Redshift.
Consumo de datos desde Power BI: Este proceso hace referencia al desarrollo de las respectivas métricas que se realizarán en Power BI para generar los respectivos reportes. Así como también abarca el diseño de las gráficas y la lógica de los tableros a presentar.

SALIDAS:
Tableros analíticos: Se refiere a los Dasboard o tableros que se generarán al finalizar todos los procesos antes mencionados los cuales están formados por gráficos y reportes.

CONTROL:
Diseño dimensional basado en Arquitectura Ralph Kimball: Ralph Kimball desarrollo una metodología para el diseño de modelos dimensionales, del cual nos hemos basado para realizar el modelo dimensional adecuado para la problemática presentada. La solución gira en torno a ese modelo dimensional.
Data Profiling: Son los datos proporcionados por la tienda, a través de su base de datos origen, que han sido examinados para determinar incongruencias, relaciones, tipos de datos y determinar qué información es útil para poder implementar la solución solicitada.

FRONTERA:
Proceso de negocio: ventas e inventario: La tienda Almacenes El Rey cuanta con una variedad procesos de negocio, pero el límite hasta donde abarca la solución presentada solo se establece a las actividades relacionadas con los procesos de venta e inventario.

MEDIO AMBIENTE:
Usuarios de negocio de la tienda Almacenes el Rey: Son todos aquellos usuarios que hacen uso de la aplicación con la que cuenta la tienda Almacenes el Rey, así como los usuarios que utilizarán la solución presentada.

2. Alcances

La solución construida tiene los siguientes elementos:

- Diseño de un modelo dimensional, mediante el cual sea posible responder a las métricas requeridas por el negocio, y adicionalmente brinde insumos para un mejor análisis del rendimiento de los procesos venta e inventario.
- Creación del espacio de almacenamiento de datos en sus diferentes fases, utilizando los componentes S3 y Redshift de Amazon Web Services.
- Implementación de procesos ETL para el procesamiento y almacenamiento en cada zona definida en S3 para los datos transaccionales que los convierte a información medible. Los ETL realizados se clasifican en:
    1. ETL para extracción de datos y almacenamiento en la zona Raw en S3.
    2. ETL para limpieza y transformación y almacenamiento en la zona Stage en S3.
    3. ETL para cálculos y preparación final de la información y almacenamiento en la zona Presentation en S3.
- Presentación de los resultados finales, haciendo uso de reportes y dashboards en Power BI. Cada reporte corresponde a una metrica establecida, específicamente se obtendrán los siguientes informes para los procesos venta e inventario:
    1. Monto total de ventas por productos en una fecha determinada
    2. Cantidad de producto vendido en una fecha determinada
    3. Total, de las ventas por producto incluyendo descuentos en una fecha determinada
    4. Total, ganado por un producto luego de incluir impuestos, descuentos y devoluciones
    5. Cantidad de producto que se tiene disponible en inventario para una fecha determinada.
    6. Costo promedio de un producto para un periodo de tiempo determinado
    7. Devoluciones de mercadería en inventario.

    Cada uno de los reportes contendrá elementos visuales como gráficas, listas, tablas, etc. según mejor representen los resultados obtenidos.

3. Justificación

Anteriormente, gestionar los procesos operativos y almacenar/extraer información en una base de datos transaccional era lo conveniente para los usuarios, en la actualidad esto no llega a ser suficiente. En los últimos años se ha dado realce al valor que poseen los datos, se ha determinado que, por medio de estos, después de un adecuado análisis y procesamiento se pueden obtener respuestas a muchas preguntas, ya que mediante técnicas y herramientas es posible: realizar mediciones estadísticas y probabilísticas, encontrar patrones o trayectorias de comportamiento de un fenómeno en estudio, crear predicciones y proyecciones, por mencionar algunos. Debido a esto ha surgido la necesidad de ahondar más en la información y obtener respuestas de ella.

Cualquier institución independientemente al rubro que se dedique necesita medir el rendimiento de los procesos que posee, con la finalidad de tener resultados precisos de los acontecimientos dados, ya que estos serán la base y respaldo para las decisiones ya fuesen a nivel operativo, administrativo o gerencial que se tomen.

La actual solución atiende a la necesidad de medir el comportamiento de los procesos venta e inventario de un comercio en línea, el cual se compone principalmente de un software donde se ejecutan dichos procesos y una base de datos donde se almacena la información generada. Para originar las mediciones que requiere el usuario, se tendrían que procesar los datos cada vez que se desee obtener cierto resultado, además agregar nuevas características al software actual (de ser eso posible) lo cual aumentaría la complejidad de la actividad y no fuera la solución óptima.

Para otorgar una solución eficiente a este tipo de tareas se emplea el concepto de modelo dimensional, cuya idea central es construir una base de datos que contenga solamente la información útil para obtener las métricas que se requieran. Dicho modelo se divide en tablas de dimensión, las cuáles contienen las entidades que intervienen en los procesos, por mencionar: Productos, Proveedores, etc. y por otro lado se sitúan las tablas de hechos, las cuales almacenan información de las operaciones que se llevaron a cabo, al nivel de detalle que el usuario decida y que la base transaccional permita.

Implementando esta solución, la información se representa de manera entendible y sumarizada, lista para calcular las mediciones necesarias. Adicionalmente provee alto rendimiento y flexibilidad para realizar consultas de información.

4. Métodos de obtención de información

La solución brindada en este proyecto engloba dos grandes procesos, el primero es el proceso de negocio de venta y el segundo, es el proceso de negocio de inventario. Para obtener los datos origen, que se ocuparán para el análisis y diseño de los modelos dimensionales correspondientes, se obtuvo información mediante dos diferentes formas:

Información obtenida de la tienda

La tienda Almacenes El Rey hace uso del Open Source Magento, el cual puede encontrarse en internet e instalarlo en cualquier computador, y luego de realizada cierta configuración se obtiene una aplicación funcional con su respectiva base de datos, la cual cuenta con una estructura definida, sus correspondientes relaciones, lógica ya establecida para los diferentes procesos de negocio y transacciones, así como también ciertos datos para su correcto funcionamiento.

Es de estos datos que ya contiene la base de datos y la lógica de negocio con la que cuenta el software, los que utilizamos para obtener parte de la información que se utiliza para el desarrollo de la solución. Entre algunos de esos datos tenemos los países y sus correspondientes códigos, tablas catalogo que contienen diferentes atributos como: colores, tallas, climas, material, etc., datos por default, entre otras tablas que no se consideraron necesarias para este proyecto.

A pesar de que la aplicación y la base de datos ya trae ciertos datos, estos son insuficientes y no abarcan las áreas que necesitamos para el análisis, por lo que, mayormente, se utilizó la siguiente forma de obtención de datos.

Información generada por el equipo

Así como se explicó anteriormente, este proyecto da un prototipo de solución para una tienda ficticia seleccionada y asignada por la catedra para poner en práctica lo aprendido durante la especialización, razón por la cual fue necesario realizar ciertos procesos para obtener la información pertinente y dar una correcta solución.

Al utilizar el Open Source Magento con su base datos en MySQL, esta base solo cuenta con información que ya trae por defecto para que funcione tanto la aplicación como para mostrar cierta información de forma general, pero referente a procesos de ventas e inventario no cuenta con esos datos, por lo cual se tuvo que realizar el ingreso de esa información por parte del equipo de trabajo.

Para ello, el equipo de trabajo tuvo que utilizar la aplicación simulando diferentes transacciones para poder generar datos con ella, entre los procesos que se tuvieron que realizar están: la creación de clientes y el login de los mismos para la realización de compra de productos, la realización de ventas de producto, así como su respectiva facturación, la administración de inventario, la realización de devoluciones, la administración y creación de cupones; la activación, creación y configuración de las promociones; el ingreso, creación y configuración de productos, el ingreso y la administración de categorías, configuración de los impuestos correspondiente para cada país, entre otras acciones necesarias para la generar datos origen. Estos procesos se realizaron intentando seguir siempre los lineamientos y la lógica de negocio establecida por el software.

Y por último, para la información referente a las fechas, se utilizó un script proporcionado por la cátedra que contiene un catálogo de fechas abarcando desde el 1° de enero de 2020 hasta la el 31 de diciembre de 2025, este script se tiene en formato .CSV.

La información adquirida con esta forma de obtención de datos corresponde, aproximadamente, al 85% de toda la información origen con la que se trabaja en este proyecto.

5. Descripción de data set

A continuación, se describe de forma general el Dataset utilizado para el análisis y creación de los modelos dimensionales.

Formato: El Dataset consta de 20 tablas cada una con su respectivo encabezado, tipo de dato y tamaño.

Estructura: Cada una de las tablas que conforman el Dataset poseen sus respectivas llaves primarias y foráneas (en las que sea necesario), así como también la metadata correspondiente y los tipos de dato para cada columna. En algunas columnas se aceptan nulos y en otros no.

Periodo de tiempo del dataset: El Dataset contiene información desde el mes de Julio del 2022 hasta el mes de noviembre del 2022.

Modelo relacional del Dataset de la tienda Almacenes El Rey:

En la siguiente *figura 6* se presenta el modelo relacional del dataset de la tienda Almacenes El Rey.

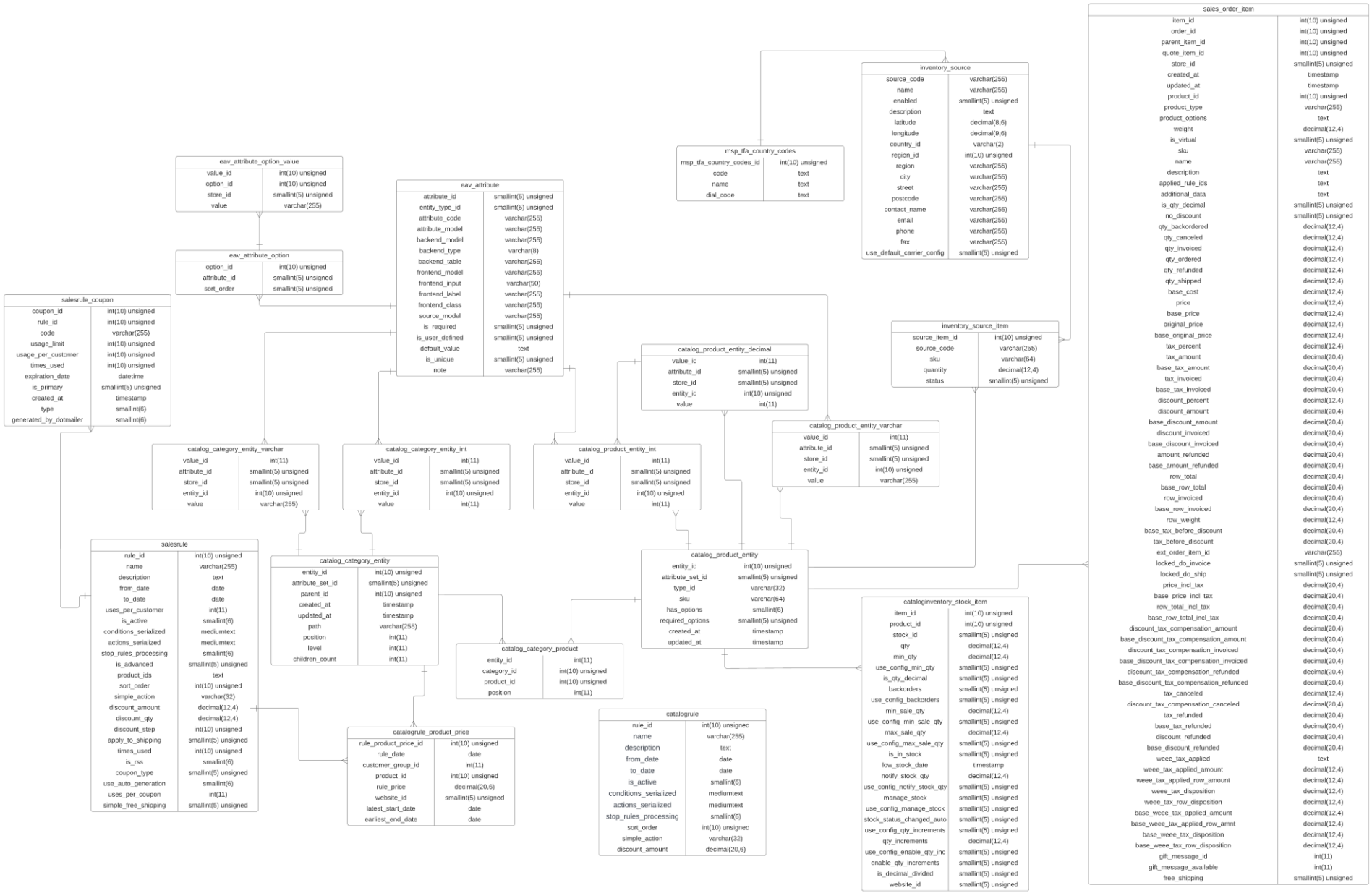


Figura 6: Modelo relacional del Dataset de la tienda Almacenes El Rey

6. Diccionario de datos del data set

Nombre de la tabla: catalog_category_entity

Descripción: Es la tabla catalogo donde se almacenan las categorías.

A continuación, se presenta la *Tabla 1* con la descripción de datos de catalog_category_entity

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
entity_id	Llave primaria	Int	10	Unsigned	No		Autoincremental		Número secuencial que representa el Id de la categoría
attribute_set_id		Smallint	5	Unsigned	No	0			Contiene el Id del conjunto de atributos
parent_id		Int	10	Unsigned	No	0			Id de la categoría principal
created_at		Timestamp			No	current_timestamp()			Fecha de creación
updated_at		Timestamp			No	current_timestamp()			Fecha de modificación
path		Varchar	255		No				Orden de posiciones de las categorías con las que está relacionado
position		Int	11		No				Posición
level		Int	11		No	0			Nivel

children_count		Int	11		No				Recuento de hijos

Tabla 1: Descripción de datos del Dataset. Tabla: catalog_category_entity

Nombre de la tabla: catalog_category_entity_int

Descripción: Categoría de catálogo de atributos tipo enteros.

A continuación, se presenta la *Tabla 2* con la descripción de datos de catalog_category_entity

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
value_id	Llave primaria	int	11		No		Autoincremental		Número secuencial que representa el Id
attribute_id	Llave foránea	smallint	5	Unsigned	No		Unique	eav_attribute	Id del atributo al que pertenece el valor
store_id	Llave foránea	smallint	5	Unsigned	No			store	Id de la tienda
entity_id	Llave foránea	int	10	Unsigned	No			catalog_category_entity	Id de catálogo de categoría
value		int	11			Null			Valor

Tabla 2: Descripción de datos del Dataset. Tabla: catalog_category_entity

Nombre de la tabla: catalog_category_entity_varchar

Descripción: Categoría de catálogo de atributos tipo Varchar.

A continuación, se presenta la *Tabla 3* con la descripción de datos de catalog_category_entity_varchar

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
value_id	Llave primaria	int	11		No		Autoincremental		Número secuencial que representa el Id
attribute_id	Llave foranea	smallint	5	Unsigned	No		Unique	eav_attribute	Id del atributo al que pertenece el valor
store_id	Llave foranea	smallint	5	Unsigned	No			store	Id de la tienda
entity_id	Llave foranea	int	10	Unsigned	No			catalog_category_entity	Id de catálogo de categoría
value		varchar	255			Null			Valor

Tabla 3: Descripción de datos del Dataset. Tabla: catalog_category_entity_varchar

Nombre de la tabla: catalog_product_entity

Descripción: Contiene el catálogo de los productos.

A continuación, se presenta la *Tabla 4* con la descripción de datos de catalog_product_entity

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
entity_id	Llave primaria	Int	10	Unsigned	No		Autoincremental		Número secuencial que representa el Id
attribute_set_id		smallint	5	Unsigned	No	0			Id del conjunto de atributos
type_id		Varchar	32		No	'simple'			Tipo de identificación
sku		Varchar	64						sku
has_options		Smallint	6		No				Tiene opciones
required_options		Smallint	5	Unsigned	No				Opciones requeridas
created_at		Timestamp			No	current_timestamp()			Fecha de creación
updated_a		Timestamp			No	current_timestamp()			Fecha de modificación

Tabla 4: Descripción de datos del Dataset. Tabla: catalog_product_entity

Nombre de la tabla: catalog_product_entity_decimal

Descripción: Contiene los datos de los atributos decimales del catálogo de productos.

A continuación, se presenta la *Tabla 5* con la descripción de datos de catalog_product_entity_decimal

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
value_id	Llave primaria	int	11		No		Autoincremental		Número secuencial que representa el Id
attribute_id	Llave foránea	smallint	5	Unsigned	No	0	Unique key	eav_attribute	Id del atributo al que pertenece el valor
store_id	Llave foránea	smallint	5	Unsigned	No	0		store	Id de la tienda
entity_id	Llave foránea	int	10	Unsigned	No	0		catalog_product_entity	Id de catálogo de categoría
value		decimal	(20,6)			null			Valor decimal

Tabla 5: Descripción de datos del Dataset. Tabla: catalog_product_entity_decimal

Nombre de la tabla: catalog_product_entity_int

Descripción: Contiene los datos de los atributos enteros del catálogo de productos.

A continuación, se presenta la *Tabla 6* con la descripción de datos de catalog_product_entity_int

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
value_id	Llave primaria	int	11		No		Autoincremental		Número secuencial que representa el Id
attribute_id	Llave foránea	smallint	5	Unsigned	No	0	Unique key	eav_attribute	Id del atributo al que pertenece el valor
store_id	Llave foránea	smallint	5	Unsigned	No	0		store	Id de la tienda
entity_id	Llave foránea	int	10	Unsigned	No	0		catalog_product_entity	Id de catálogo de categoría
value		int	11			null			Valor decimal

Tabla 6: Descripción de datos del Dataset. Tabla: catalog_product_entity_int

Nombre de la tabla: catalog_product_entity_varchar

Descripción: Contiene los datos de los atributos de caracteres del catálogo de productos.

A continuación, se presenta la *Tabla 7* con la descripción de datos de catalog_product_entity_varchar

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
value_id	Llave primaria	int	11		No		Autoincremental		Número secuencial que representa el Id
attribute_id	Llave foránea	smallint	5	Unsigned	No	0	Unique key	eav_attribute	Id del atributo al que pertenece el valor
store_id	Llave foránea	smallint	5	Unsigned	No	0		store	Id de la tienda
entity_id	Llave foránea	int	10	Unsigned	No	0		catalog_product_entity	Id de catálogo de categoría
value		varchar	255			null			Valor de caracteres

Tabla 7: Descripción de datos del Dataset. Tabla: catalog_product_entity_varchar

Nombre de la tabla: cataloginventory_stock_item

Descripción: Contiene el catálogo de los artículos en stock del inventario

A continuación, se presenta la *Tabla 8* con la descripción de datos de cataloginventory_stock_item

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
item_id	Llave primaria	int	10	Unsigned	No		Autoincremental		Número secuencial que representa el Id
product_id	Llave foranea	int	10	Unsigned	No	0	Unique key	catalog_product_entity	Id del producto al que pertenece
stock_id	Llave foranea	smallint	5	Unsigned	No	0		store	Id de la tienda
Qty	Llave foranea	decimal	12,4			Null			Cantidad
min_qty		decimal	12,4		No	0.0000			Cantidad mínima
use_config_min_qty		smallint	5	Unsigned	No	1			Usar cantidad mínima de

									configuración
is_qty_decimal		smallint	5	Unsigned	No	0			Es decimal la cantidad
Backorders		smallint	5	Unsigned	No	0			Pedidos atrasados
use_config_backorders		smallint	5	Unsigned	No	1			Usar pedidos pendientes de configuración
min_sale_qty		decimal	12,4		No	1.0000			Cantidad mínima de venta
use_config_min_sale_qty		smallint	5	Unsigned	No	1			Usar configuracion minima
max_sale_qty		decimal	12,4		No	0.0000			Cantidad máxima de venta
use_config_max_sale_qty		smallint	5	Unsigned	No	1			Usar cantidad maxima de venta de configuración
is_in_stock		smallint	5	Unsigned	No	0			Esta en stock

low_stock_date		Timestamp				Null			Fecha de existencias bajas
notify_stock_qty		decimal	12,4			Null			Notificar cantidad de existencias
use_config_notify_stock_qty		smallint	5	Unsigned	No	1			Usar configuracion notificar cantidad de existencias
manage_stock		smallint	5	Unsigned	No	0			Administrar existencias
use_config_manage_stock		smallint	5	Unsigned	No	1			Usar configuracion administrar stock
stock_status_changed_auto		smallint	5	Unsigned	No	0			El estado del stock cambio automaticamente
use_config_qty_increments		smallint	5	Unsigned	No	1			Usar incrementos de cantidad de configuración

qty_increments		decimal	12,4		No	0.0000			Incrementos de cantidad
use_config_enable_qty_inc		smallint	5	Unsigned	No	1			Usar configuracion habilitar incrementos de cantidad
enable_qty_increments		smallint	5	Unsigned	No	0			Habilitar incremnetos de cantidad
is_decimal_divided		smallint	5	Unsigned	No	0			Se divide en varias cajas para el envio
website_id		smallint	5	Unsigned	No	0			Id del sitio web

Tabla 8: Descripción de datos del Dataset. Tabla: cataloginventory_stock_item

Nombre de la tabla: catalogrule

Descripción: Catalogo de promociones

A continuación, se presenta la *Tabla 9* con la descripción de datos de catalogrule

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
rule_id	Llave primaria	Int	10	Unsigned	No	-	Autoincremental		Número secuencial que

									representa el Id
name		Varchar	255			NULL			Nombre
Description		Text				NULL			Descripción
from_date		date				NULL			Desde que fecha
to_date		date				NULL			Hasta que fecha
is_active		smallint	6		No	0			Es activo
conditions_serialized		mediumtext				NULL			Condiciones serializadas
actions__serialized		mediumtext				NULL			Acciones realizadas
stop_rules_processing		smallint	6		No	1			Detener procesamiento de reglas
sort_order		int	10	Unsigned	No	0			Orden de clasificación
simple_action		varchar	32			NULL			Acción sencilla
discount_amount		decimal	20,6		No	0.00000			Importe de descuento

Tabla 9: Descripción de datos del Dataset. Tabla: catalogrule

Nombre de la tabla: catalogrule_product_price

Descripción: Catalogo regla del precio de producto.

A continuación, se presenta la *Tabla 10* con la descripción de datos de catalogrule_product_price

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
rule_product_price_id	Llave primaria	int	10	Unsigned	No		Autoincremental		Número secuencial que representa el Id
rule_date		date			No				Fecha de la regla
customer_group_id		int	11			Null			Id del grupo de clientes a quienes aplica
product_id		Int	10	Unsigned	No	0			Id del producto
rule_price		Decimal	20,6		No	0.000000			Regla del precio
website_id		Smallint	5	Unsigned	No				Id del sitioweb
latest_start_date		date				Null			Ultima fecha de inicio
earliest_end_date		date				Null			Fecha de finalización mas temprana

Tabla 10: Descripción de datos del Dataset. Tabla: catalogrule_product_price

Nombre de la tabla: eav_attribute

Descripción: Tabla de atributos eav

A continuación, se presenta la *Tabla 11* con la descripción de datos de eav_attribute

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
attribute_id	Llave primaria	smallint	5	Unsigned	No		Autoincremental		Número secuencial que representa el Id
entity_type_id	Llave foranea	smallint	5		No		Unique	eav_entity_type	Id del tipo de entidad
attribute_code		varchar	255		No				Código de atributo
attribute_model		varchar	255						Modelo de atributo
backend_model		varchar	255						Modelo de fondo
backend_type		varchar	8		No				Tipo de servidor
backend_table		varchar	255						Tabla de fondo
frontend_model		varchar	255						Modelo de interfaz
frontend_input		varchar	50						Entrada de interfaz
frontend_label		varchar	255						Etiqueta frontal
frontend_class		varchar	255						Clase de interfaz
source_model		varchar	255						Modelo fuente
is_required		smallint	5		No				Se requiere definir

is_user_defined		smallint	5		No				Es definido por el usuario
default_value		text							Valor por defecto
is_unique		smallint	5		No				Define es único
note		varchar	255						Nota

Tabla 11: Descripción de datos del Dataset. Tabla: eav_attribute

Nombre de la tabla: eav_attribute_option

Descripción: Opción del atributo eav.

A continuación, se presenta la *Tabla 12* con la descripción de datos de eav_attribute_option

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
option_id	Llave primaria	int	10	Unsigned	No		Autoincremental		Número secuencial que representa el Id
attribute_id	Llave foránea	smallint	5		No			eav_attribute	Id del atributo
sort_order		smallint	5		No				Orden de clasificación

Tabla 12: Descripción de datos del Dataset. Tabla: eav_attribute_option

Nombre de la tabla: eav_attribute_option_value

Descripción: Valor de la opción del atributo eav.

A continuación, se presenta la *Tabla 13* con la descripción de datos de eav_attribute_option_value

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
value_id	Llave primaria	int	10	Unsigned	No		Autoincre Mental		Número secuencial que representa el Id
option_id	Llave foránea	Int	10	Unsigned	No				Id de la opción
store_id	Llave foránea	Smallint	5	Unsigned	No				Id de la tienda
value		varchar	255						Valor

Tabla 13: Descripción de datos del Dataset. Tabla: eav_attribute_option_value

Nombre de la tabla: inventory_source

Descripción: Catalogo de la fuente de inventario.

A continuación, se presenta la *Tabla 14* con la descripción de datos de inventory_source

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
source_code	Llave primaria	varchar	255		No		Autoincremental		Número secuencial que

									representa el Id
Name		varchar	255		No				Nombre
Enabled		smallint	5	Unsigned	No	1			Habilitado
Description		Text				Null			Descripción
Latitude		decimal	8,6			Null			Latitud
Longitude		decimal	9,6			Null			Longitud
country_id		varchar	2		No				Id del país
region_id		Int	10	Unsigned		Null			Id de la región
Región		varchar	255			Null			Región
City		varchar	255			Null			Ciudad
Street		varchar	255			Null			Calle
Postcode		varchar	255		No				Código postal
contact_name		varchar	255			Null			Nombre de contacto
Email		varchar	255			Null			Email
Pone		varchar	255			Null			Teléfono

Fax		varchar	255			Null			fax
use_default_carrier_config		smallint	5	Unsigned	No	1			Usar la configuración de operador predeterminada

Tabla 14: Descripción de datos del Dataset. Tabla: inventory_source

Nombre de la tabla: inventory_source_item

Descripción: Contiene los datos del inventario en items.

A continuación, se presenta la *Tabla 15* con la descripción de datos de inventory_source_item

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
source_item_id	Llave primaria	Int	10	Unsigned	No		Autoincremental		Número secuencial que representa el Id
source_code	Llave foránea	Varchar	255		No			inventory_source	Código fuente
sku		Varchar	64		No				Sku
quantity		decimal	12,4		No				Cantidad
status		smallint	5	Unsigned	No				Estado

Tabla 15: Descripción de datos del Dataset. Tabla: inventory_source_item

Nombre de la tabla: msp_tfa_country_codes_id

Descripción: Contiene los datos del código de país.

A continuación, se presenta la *Tabla 16* con la descripción de datos de msp_tfa_country_codes_id

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
msp_tfa_country_codes_id	Llave primaria	int	10	Unsigned	No		Autoincremental		Número secuencial que representa el Id
Code		text			No				Código del país
Name		text			No				Nombre del país
dial_code		text			No				Prefijo

Tabla 16: Descripción de datos del Dataset. Tabla: msp_tfa_country_codes_id

Nombre de la tabla: sales_order_item

Descripción: Contiene los datos de las ventas ordenadas por item.

A continuación, se presenta la *Tabla 17* con la descripción de datos de sales_order_item

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo

item_id	Llave primaria	Int	10	Unsigned	No		Autoincremental		Número secuencial que representa el Id del cliente
Order_id		Int	10	Unsigned	No	0			Id de la orden
parent_item_id		Int	10	Unsigned		NULL			Identificación del articulo principal
quote_item_id		Int	10	Unsigned		NULL			Id del articulo de cotización
store_id		Smallint	5	Unsigned		NULL			Identificación de la tienda
create_at		Timestamp			No	Current_timestamp()			Creado en
update_at		Timestamp			No	Current_timestamp()			Actualizado en
product_id		Int	10	Unsigned		NULL			Identificación de producto
product_type		Varchar	255			NULL			Tipo de producto
producto_options		Text				NULL			Opciones de productos
Weight		Decimal	12,4			0.0000			Altura
is_virtual		Smallint	5	Unsigned		NULL			Es virtual
Sku		Varchar	255			NULL			SKU
Name		Varchar	255			NULL			Nombre
Description		Text				NULL			Descripción

applied_rule_ids		Text				NULL			Id de regla aplicada
additional_data		Text				NULL			Datos adicionales
is_qty_decimal		Smallint	5	Unsigned		NULL			Es la cantidad decimal
no_discount		Smallint	5	Unsigned	No	0			Sin descuento
qty_backordered		decimal	12,4			0.0000			Cantidad en espera
qty_canceled		decimal	12,4			0.0000			Cantidad cancelada
qty_invoiced		decimal	12,4			0.0000			Cantidad facturada
qty_ordered		decimal	12,4			0.0000			Cantidad ordenada
qty_refunded		decimal	12,4			0.0000			Cantidad reembolsada
qty_shipped		decimal	12,4			0.0000			Cantidad enviada
base_cost		decimal	12,4			0.0000			Costo base
Price		decimal	12,4		No	0.0000			Precio
base_price		decimal	12,4		No	0.0000			Precio base
original_price		decimal	12,4			NULL			Precio original
base_original_price		decimal	12,4			NULL			Precio base original
tax_percent		decimal	12,4			0.0000			Porcentaje de impuestos

tax_amount		decimal	20,4			0.0000			Importe del impuesto
base_tax_amount		decimal	20,4			0.0000			Importe base del impuesto
tax_invoiced		decimal	20,4			0.0000			Impuesto facturado
base_tax_invoiced		decimal	20,4			0.0000			Base imponible facturada
discount_percent		decimal	12,4			0.0000			Porcentaje de descuento
discount_amount		decimal	20,4			0.0000			Importe de descuento
base_discount_amount		decimal	20,4			0.0000			Cantidad de descuento base
discount_invoiced		decimal	20,4			0.0000			Descuento facturado
base_discount_invoiced		decimal	20,4			0.0000			Descuento base facturado
amount_refunded		decimal	20,4			0.0000			Cantidad reembolsada
base_amount_refunded		decimal	20,4			0.0000			Importe base reembolsado
row_total		decimal	20,4		No	0.0000			Total de filas
base_row_total		decimal	20,4		No	0.0000			Total de la fila base
row_invoiced		decimal	20,4		No	0.0000			Fila facturada
base_row_invoiced		decimal	20,4		No	0.0000			Fila base facturada
row_weight		decimal	12,4			0.0000			Peso de fila

base_tax_before_discount		decimal	20,4			NULL			Impuesto base antes del descuento
tax_before_discount		decimal	20,4			NULL			Impuesto antes del descuento
ext_order_item_id		Varchar	255			NULL			Id de articulo de pedido externo
locked_do_invoice		Smallint	5	Unsigned		NULL			Factura bloqueada
locked_do_ship		Smallint	5	Unsigned		NULL			Envio bloqueado
price_incl_tax		decimal	20,4			NULL			Precio incl.. impuestos
base_price_incl_tax		decimal	20,4			NULL			Precio base con impuestos incluidos
row_total_incl_tax		decimal	20,4			NULL			Fila total impuestos incluidos
base_row_total_incl_tax		decimal	20,4			NULL			Fila base total impuestos incluidos
discount_tax_compensation_amount		decimal	20,4			NULL			Importe de compensacion de impuestos de descuento
base_discount_tax_compensation_amount		decimal	20,4			NULL			Importe de compensación de impuestos

									de descuento base
discount_tax_compensation_invoiced		decimal	20,4			NULL			descuento impuesto compensacion facturada
base_discount_tax_compensation_invoiced		decimal	20,4			NULL			Reembolso de compensación de impuestos de descuento
discount_tax_compensation_refunnded		decimal	20,4			NULL			Reembolso de compensación de impuestos de descuento base
base_discount_tax_compensation_refunded		decimal	20,4			NULL			Impuesto cancelado
tax_canceled		decimal	12,4			NULL			Cancelación de compensación de impuestos de descuento
discount_tax_compensation_canceled		decimal	20,4			NULL			Impuesto reembolsado
tax_refunded		decimal	20,4			NULL			Impuesto base reembolsado
base_tax_refunded		decimal	20,4			NULL			Descuento reembolsado
discount_refunded		decimal	20,4			NULL			Descuento base reembolsado

base_discount_refunded		decimal	20,4			NULL			Impuesto Weee aplicado
weee_tax_applied		text				NULL			Importe aplicado del impuesto Weee
weee_tax_applied_amount		decimal	12,4			NULL			Importe aplicado del impuesto Weee
weee_tax_applied_row_amount		decimal	12,4			NULL			Monto de la fila del impuesto Weee aplicado
weee_tax_disposition		decimal	12,4			NULL			Disposición de impuestos Weee
weee_tax_row_disposition		decimal	12,4			NULL			Disposición de la fila de impuestos Weee
base_weee_tax_applied_amount		decimal	12,4			NULL			Importe aplicado del impuesto base Weee
base_weee_tax_applied_row_amnt		decimal	12,4			NULL			Monto de fila de impuesto de Weee base aplicado
base_weee_tax_disposition		decimal	12,4			NULL			Disposición fiscal base Weee

base_weee_tax_row_disposition		decimal	12,4			NULL			Disposición de la fila de impuestos Weee base
gift_message_id		Int	11			NULL			Id de mensaje de regalo
gift_message_available		Int	11			NULL			Mensaje de regalo disponible
free_shipping		Smallint	5	Unsigned	No	0			Envio gratis

Tabla 17: Descripción de datos del Dataset. Tabla: sales_orden_item

Nombre de la tabla: salesrule

Descripción: Contiene los datos de los cupones.

A continuación, se presenta la *Tabla 18* con la descripción de datos de salesrule

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
rule_id	Llave primaria	Int	10	Unsigned	No		Autoincremental		Número secuencial que representa el Id del cliente
Name		Varchar	250		No				Nombre
Description		text							Descripcion
from_date		date							De
to_date		date							A

uses_per_customer		int	11		No				Usos por cliente
is_active		smallint	6		No				Estado activo
conditions_serialized		Mediumtext							Condiciones serializadas
actions_serialized		Mediumtext							Acciones serializadas
stop_rules_processing		smallint	6		No				Detener procesamiento de reglas
is_advanced		smallint	5	Unsigned	No				Es avanzado
product_ids		text							Identificadores de productos
sort_order		int	10	Unsigned	No				Orden de clasificación
simple_action		varchar	32						Acción sencilla
discount_amount		decimal	12,4		No				Importe de descuento
discount_qty		decimal	12,4						Cantidad de descuento
discount_step		int	10	Unsigned					Paso de descuento
apply_to_shipping		smallint	5	Unsigned					Aplicar al envío
times_used		int	10	Unsigned	No				Veces utilizado
is_rss		smallint	6		No				Es rss
coupon_type		smallint	5	Unsigned	No				Tipo de cupón

use_aouto_generation		smallint	6		No				Usar generación automática
uses_per_coupon		int	11		No				Usuario por cupón
simple_free_shipping		smallint	5	Unsigned	No				Envío sencillo gratuito

Tabla 18: Descripción de datos del Dataset. Tabla: salesrule

Nombre de la tabla: salesrule_coupon

Descripción: Tabla de las reglas de ventas de cupones

A continuación, se presenta la *Tabla 19* con la descripción de datos de salesrule_coupon

Nombre del campo	Clave	Tipo de dato	Tamaño	Atributos	Acepta nulo	Valor predeterminado	Detalle	Enlace a	Descripción del campo
coupon_id	Llave primaria	int	10	Unsigned	No		Autoincremental		Número secuencial que representa el Id
rule_id		int	10	Unsigned	No		Unique	salesrule	Id de regla
Code		varchar	255			Null	Unique		Código
usage_limit		int	10	Unsigned		Null			Límite de uso
usage_per_customer		int	10	Unsigned		Null			Uso por cliente
times_used		int	10	Unsigned	No	0			Veces utilizado

expiration_date		datetime				Null			Fecha de expiración
is_primary		smallint	5	Unsigned		Null	Unique		Es Primario
created_at		Timestamp				Null			Fecha de creación del código de cupón
Type		smallint	6			0			Tipo de código de cupón
generated_by_dotmailer		smallint	6			Null			

Tabla 19: Descripción de datos del Dataset. Tabla: salesrule_coupon

7. Data profiling

El data profiling o perfilado de datos permite descubrir, comprender y organizar datos mediante la identificación de sus características y la evaluación de su calidad. Este proceso puede revelar si los datos están completos o si son únicos, si se detectan errores y patrones inusuales y si se determina la facilidad de uso.

1. Herramientas utilizadas para realizar el Data Profiling

Para realizar el perfilado de datos se utilizaron las siguientes herramientas:

DataCleaner

Es una herramienta que sirve para analizar la calidad de los datos obtenidos, con capacidad para encontrar patrones y supervisar los valores de los datos. En la *figura 7* podrá observar el logo de la herramienta. DataCleaner fue útil para determinar el tipo de dato de la información base y realizar diferentes análisis en base a ello. Determinaba el número de filas, el número de nulos, en el caso de encontrar valores enteros daba que valor era el más alto, cuál era el más bajo, la suma de todos esos valores, la desviación estándar, la varianza, entre otros valores relacionados a operaciones numéricas; al encontrar valores de tipo de dato string realizaba un conteo de espacios en blanco, conteo de mayúsculas y minúsculas, conteo total de caracteres así como los máximos y mínimos, contador de palabras, sus máximos y mínimos, entre otro tipo de información relacionada a textos y por último, dio información relevante a tipo de dato de fecha o tiempo, dando que fecha era la más alta, cuál era la más baja, el tiempo total más alto y el más bajo, entre otro tipo de información.

Figura 7: DataCleaner

DBeaver.

Es una aplicación de software cliente de SQL y una herramienta de administración de bases de datos. En la *figura 8* podrá observar el logo de la herramienta. Esta herramienta fue útil para continuar con el análisis debido a que, por el tamaño de la base de datos original, que constaba aproximadamente con más de 200 tablas, era muy difícil lograr ver todas las relaciones y las dependencias entre ellas, DBeaver fue de apoyo para tener una visualización más clara tanto en las relaciones entre tablas como en la evaluación de los datos, haciendo uso de filtros y sentencias SQL para determinar errores en los datos, valores nulos o vacíos, duplicidad en los datos y utilidad de los mismos.

Figura 8: DBeaver

2. Resultados del Data Profiling

Utilizando las herramientas antes mencionadas y filtrando solo aquellas columnas de las tablas del Dataset establecido anteriormente, se obtuvieron los siguientes resultados:

Nombre de la tabla: catalog_category_entity

- De los 9 campos de los que está conformada la tabla, se seleccionaron 5 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador entity_id
- Ninguno de los campos posee valores nulos

A continuación, se presenta la *Tabla 20* con el análisis de la tabla catalog_category_entity

Nombre del campo	Resultado	Útil para el análisis
entity_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
attribute_set_id	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No
parent_id	Datos 100% validos, no se requiere transformación y modificación. Útil para identificar relación recursiva.	**Si**
created_at	Datos 100% validos, solo requiere modificación en el formato de la fecha.	**Si**
updated_at	Datos 100% validos, solo requiere modificación en el formato de la fecha.	**Si**
path	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No
position	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No
level	Datos 100% validos, no se requiere transformación y modificación	**Si**
children_count	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No

Tabla 20: Data Profiling – Tabla: catalog_category_entity

Nombre de la tabla: catalog_category_entity_int

- De los 5 campos de los que está conformada la tabla, se seleccionaron 4 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador value_id
- Solo se encontró una columna con datos nulos, todos los demás no poseen datos nulos

A continuación, se presenta la *Tabla 21* con el análisis de la tabla catalog_category_entity_int

Nombre del campo	Resultado	Útil para el análisis
value_id	Datos 100% validos, no se requiere transformación y modificación	**Si**
attribute_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla eav_attribute	**Si**
store_id	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No
entity_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla catalog_category_entity	**Si**
value	Datos 98% validos, Se encontraron valores nulos y necesario aplicar filtros. Guarda información requerida para el análisis.	**Si**

Tabla 21: Data Profiling – Tabla: catalog_category_entity_int

Nombre de la tabla: catalog_category_entity_varchar

- De los 5 campos de los que está conformada la tabla, se seleccionaron 4 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador value_id
- Solo se encontró una columna con datos nulos, todos los demás no poseen datos nulos

A continuación, se presenta la *Tabla 22* con el análisis de la tabla catalog_category_entity_varchar

Nombre del campo	Resultado	Útil para el análisis
value_id	Datos 100% validos, no se requiere transformación y modificación	**Si**

attribute_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla eav_attribute	**Si**
store_id	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No
entity_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla catalog_category_entity	**Si**
value	Datos 89% validos, Se encontraron valores nulos y necesario aplicar filtros. Guarda información requerida para el análisis.	**Si**

Tabla 22: Data Profiling – Tabla: catalog_category_entity_varchar

Nombre de la tabla: catalog_category_product

- De los 4 campos de los que está conformada la tabla, se seleccionaron 3 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador entity_id
- Ninguno de los campos posee valores nulos

A continuación, se presenta la *Tabla 23* con el análisis de la tabla catalog_category_product

Nombre del campo	Resultado	Útil para el análisis
entity_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
category_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla catalog_category_entity	**Si**
producto_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla catalog_product_entity	**Si**
position	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No

Tabla 23: Data Profiling – Tabla: catalog_category_product

Nombre de la tabla: catalog_product_entity

- De los 8 campos de los que está conformada la tabla, se seleccionaron 5 que se consideran útiles para realizar el análisis correspondiente a este proyecto.

- No se encontraron valores duplicados en el identificador entity_id
- Ninguno de los campos posee valores nulos

A continuación, se presenta la *Tabla 24* con el análisis de la tabla catalog_product_entity

Nombre del campo	Resultado	Útil para el análisis
entity_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
attribute_set_id	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No
type_id	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No
sku	Datos 100% validos, no se requiere transformación y modificación.	**Si**
has_options	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No
required_options	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No
created_at	Datos 100% validos, solo requiere modificación en el formato de la fecha.	**Si**
updated_at	Datos 100% validos, solo requiere modificación en el formato de la fecha.	**Si**

Tabla 24: Data Profiling – Tabla: catalog_product_entity

Nombre de la tabla: catalog_product_entity_decimal

- De los 5 campos de los que está conformada la tabla, se seleccionaron 4 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador value_id
- Solo se encontró una columna con datos nulos, todos los demás no poseen datos nulos

A continuación, se presenta la *Tabla 25* con el análisis de la tabla catalog_product_entity_decimal

Nombre del campo	Resultado	Útil para el análisis
value_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
attribute_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla eav_attribute	**Si**
store_id	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No

entity_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla catalog_product_entity	**Si**
Value	Datos 99% validos, Se encontraron valores nulos y necesario aplicar filtros. Guarda información requerida para el análisis.	**Si**

Tabla 25: Data Profiling – Tabla: catalog_product_entity_decimal

Nombre de la tabla: catalog_product_entity_int

- De los 5 campos de los que está conformada la tabla, se seleccionaron 4 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador value_id
- Ninguno de los campos posee valores nulos

A continuación, se presenta la *Tabla 26* con el análisis de la tabla catalog_product_entity_int

Nombre del campo	Resultado	Útil para el análisis
value_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
attribute_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla eav_attribute	**Si**
store_id	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No
entity_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla catalog_product_entity	**Si**
Value	Datos 100% validos, no se requiere transformación y modificación pero si es necesario aplicar filtros. Guarda información requerida para el análisis.	**Si**

Tabla 26: Data Profiling – Tabla: catalog_product_entity_int

Nombre de la tabla: catalog_product_entity_varchar

- De los 5 campos de los que está conformada la tabla, se seleccionaron 4 que se consideran útiles para realizar el análisis correspondiente a este proyecto.

- No se encontraron valores duplicados en el identificador value_id
- Ninguno de los campos posee valores nulos

A continuación, se presenta la *Tabla 27* con el análisis de la tabla catalog_product_entity_varchar

Nombre del campo	Resultado	Útil para el análisis
value_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
attribute_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla eav_attribute	**Si**
store_id	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No
entity_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla catalog_product_entity	**Si**
Value	Datos 100% validos, no se requiere transformación y modificación pero si es necesario aplicar filtros. Guarda información requerida para el análisis.	**Si**

Tabla 27: Data Profiling – Tabla: catalog_product_entity_varchar

Nombre de la tabla: cataloginventory_stock_item

- De los 25 campos de los que está conformada la tabla, se seleccionaron 7 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador item_id
- Solo se encontraron dos columnas con datos nulos, todos los demás no poseen datos nulos

A continuación, se presenta la *Tabla 28* con el análisis de la tabla cataloginventory_stock_item

Nombre del campo	Resultado	Útil para el análisis
item_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
product_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla catalog_product_entity	**Si**

stock_id	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No
Qty	Todos los datos son 0 y hay valores nulos. Información sin valor	No
min_qty	Datos 100% validos, se requiere transformar el tipo de dato de decimal a entero	**Si**
use_config_min_qty	Sin inconsistencia en los datos, pero no se considera relevante para al análisis	No
is_qty_decimal	Todos los datos son 0. Información sin valor para este análisis	No
Backorders	Todos los datos son 0 significando que no existen pedidos atrasados	**Si**
use_config_backorders	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
min_sale_qty	Datos 100% validos, se requiere transformar el tipo de dato de decimal a entero	**Si**
use_config_min_sale_qty	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
max_sale_qty	Datos 100% validos, se requiere transformar el tipo de dato de decimal a entero	**Si**
use_config_max_sale_qty	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
is_in_stock	Datos 100% validos, se requiere especificar cuando es 0 existe en stock, cuando es 1 no existe en stock	**Si**
low_stock_date	Existen valores nulos, además que no se considera relevante para al análisis.	No
notify_stock_qty	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No

use_config_notify_stock_qty	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
manage_stock	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
use_config_manage_stock	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
stock_status_changed_auto	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
use_config_qty_increments	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
qty_increments	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
use_config_enable_qty_inc	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
enable_qty_increments	Todos los datos son 0. Información sin valor para este análisis	No
is_decimal_divided	Todos los datos son 0. Información sin valor para este análisis	No
website_id	Todos los datos son 0. Información sin valor para este análisis	No

Tabla 28: Data Profiling – Tabla: cataloginventory_stock_item

Nombre de la tabla: catalogrule

- De los 12 campos de los que está conformada la tabla, se seleccionaron 7 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador rule_id
- Ninguno de los campos posee valores nulos

A continuación, se presenta la *Tabla 29* con el análisis de la tabla catalogrule

Nombre del campo	Resultado	Útil para el análisis
rule_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
name	Datos 100% validos, no se requiere transformación y modificación.	**Si**
Description	Datos 100% validos, no se requiere transformación y modificación.	**Si**
from_date	Datos 100% validos, solo requiere modificación en el formato de la fecha.	**Si**
to_date	Datos 100% validos, solo requiere modificación en el formato de la fecha.	**Si**
is_active	Datos 100% validos, se requiere especificar que 1 es activo y 0 inactivo.	**Si**
conditions_serialized	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
actions__serialized	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
stop_rules_processing	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
sort_order	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
simple_action	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
discount_amount	Datos 100% validos, no se requiere transformación y modificación.	**Si**

Tabla 29: Data Profiling – Tabla: catalogrule

Nombre de la tabla: catalogrule_product_price

- De los 8 campos de los que está conformada la tabla, se seleccionaron 5 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador rule_product_price_id
- Ninguno de los campos posee valores nulos
- Esta tabla no guarda persistencia de datos

A continuación, se presenta la *Tabla 30* con el análisis de la tabla catalogrule_product_price

Nombre del campo	Resultado	Útil para el análisis
rule_product_price_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
rule_date	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No

customer_group_id	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
product_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla catalog_product_entity	**Si**
rule_price	Datos 100% validos, no se requiere transformación y modificación. Contiene el precio del producto aplicada la promoción	**Si**
website_id	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
latest_start_date	Datos 100% validos, solo requiere modificación en el formato de la fecha.	**Si**
earliest_end_date	Datos 100% validos, solo requiere modificación en el formato de la fecha.	**Si**

Tabla 30: Data Profiling – Tabla: catalogrule_product_price

Nombre de la tabla: eav_attribute

- La mayoría de los datos que se encuentran en esta tabla van dirigidos al frontend de la aplicación en Magento, incluyendo los campos seleccionados, pero por motivos de la organización de información que se maneja dentro de la base de datos es requerido consultarlos para entender la demás información que está relacionada a esta tabla y es requerida para dar una solución a la problemática planteada.
- De los 17 campos de los que está conformada la tabla, se seleccionaron 3 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador attribute_id
- Se encontraron 10 campos con valores nulos

A continuación, se presenta la *Tabla 31* con el análisis de la tabla eav_attribute

Nombre del campo	Resultado	Útil para el análisis
attribute_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
entity_type_id	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
attribute_code	Datos 100% validos, no se requiere transformación y modificación.	**Si**

attribute_model	Solo el 4% de los datos son válidos, no es relevante para el análisis.	No
backend_model	Solo el 31% de los datos son válidos, no es relevante para el análisis.	No
backend_type	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
backend_table	No se encontraron datos válidos. No es relevante para el análisis.	No
frontend_model	Solo el 4% de los datos son válidos, no es relevante para el análisis.	No
frontend_input	Datos 98% validos, pero no es requerido para el análisis	No
frontend_label	Datos 92% validos, se encontraron valores nulos.	**Si**
frontend_class	Solo el 1% de los datos son válidos, no es relevante para el análisis.	No
source_model	Solo el 25% de los datos son válidos, no es relevante para el análisis.	No
is_required	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
is_user_defined	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
default_value	Solo el 21% de los datos son válidos, no es relevante para el análisis.	No
is_unique	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
note	Menos del 1% de los datos son válidos, no es relevante para el análisis.	No

Tabla 31: Data Profiling – Tabla: eav_attribute

Nombre de la tabla: eav_attribute_option

- Esta tabla solo contiene información para relacionar atributos con su identificador, solo utilizado para consultar identificados específicos de la columna attribute_id que se relaciona con eav_attribute.
- De los 3 campos de los que está conformada la tabla, se seleccionaron 2 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador option_id
- Ninguno de los campos posee valores nulos

A continuación, se presenta la *Tabla 32* con el análisis de la tabla eav_attribute_option

Nombre del campo	Resultado	Útil para el análisis
option_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
attribute_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla eav_attribute	**Si**
sort_order	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No

Tabla 32: Data Profiling – Tabla eav_attribute_option:

Nombre de la tabla: eav_attribute_option_value

- De los 4 campos de los que está conformada la tabla, se seleccionaron 3 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador value_id
- Ninguno de los campos posee valores nulos

A continuación, se presenta la *Tabla 33* con el análisis de la tabla eav_attribute_option_value

Nombre del campo	Resultado	Útil para el análisis
value_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
option_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con eav_attribute_option	**Si**
store_id	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
value	Datos 100% validos, no se requiere transformación y modificación, pero si es necesario aplicar filtros. Guarda información requerida para el análisis.	**Si**

Tabla 33: Data Profiling – Tabla: eav_attribute_option_value

Nombre de la tabla: inventory_source

- De los 17 campos de los que está conformada la tabla, se seleccionaron 6 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador source_code
- Se encontraron 11 campos con valores nulos

A continuación, se presenta la *Tabla 34* con el análisis de la tabla inventory_source

Nombre del campo	Resultado	Útil para el análisis
source_code	Datos 100% validos, no se requiere transformación y modificación.	**Si**
name	Datos 100% validos, no se requiere transformación y modificación.	**Si**
enabled	Datos 100% validos, no se requiere transformación y modificación.	**Si**
description	Datos 53% validos, se encontraron valores nulos	**Si**
latitude	Solo el 6% de los datos son válidos, no es relevante para el análisis.	No
longitude	Solo el 6% de los datos son válidos, no es relevante para el análisis.	No
country_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla msp_tfa_country_codes_id	**Si**
region_id	Datos 53% validos, pero no se requiere para el análisis	No
region	Datos 94% validos, pero no se requiere para el análisis	No
City	Datos 94% validos, se encontraron datos nulos	**Si**
street	Solo el 6% de los datos son válidos, no es relevante para el análisis.	No
postcode	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
contact_name	Datos 94% validos, pero no se requiere para el análisis	No
Email	Datos 94% validos, pero no se requiere para el análisis	No
Pone	Datos 82% validos, pero no se requiere para el análisis	No
Fax	No se encontraron datos válidos. No es relevante para el análisis.	No
use_default_carrier_config	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No

Tabla 34: Data Profiling – Tabla: inventory_source

Nombre de la tabla: inventory_source_item

- Se utilizarán todos los campos de la tabla debido a que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador source_item_id
- No se encontraron valores nulos

A continuación, se presenta la *Tabla 35* con el análisis de la tabla inventory_source_item

Nombre del campo	Resultado	Útil para el análisis
source_item_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
source_code	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla inventory_source	**Si**
Sku	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla catalog_product_entity	**Si**
quantity	Datos 100% validos, se requiere transformar los valores de decimal a entero	**Si**
status	Datos 100% validos, no se requiere transformación y modificación. Útil para determinar el estado de los productos	**Si**

Tabla 35: Data Profiling – Tabla: inventory_source_item

Nombre de la tabla: msp_tfa_country_codes_id

- De los 4 campos de los que está conformada la tabla, se seleccionaron 3 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador msp_tfa_country_codes_id
- No se encontraron valores nulos

A continuación, se presenta la *Tabla 36* con el análisis de la tabla msp_tfa_country_codes_id

Nombre del campo	Resultado	Útil para el análisis
msp_tfa_country_codes_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
Code	Datos 100% validos, no se requiere transformación y modificación.	**Si**
Name	Datos 100% validos, no se requiere transformación y modificación.	**Si**

dial_code	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No

Tabla 36: Data Profiling – Tabla: msp_tfa_country_codes_id

Nombre de la tabla: sales_order_item

- De los 80 campos de los que está conformada la tabla, se seleccionaron 15 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador item_id
- Se encontraron varias columnas con nulos, así como también columnas cuyo valor es 0.

A continuación, se presenta la *Tabla 37* con el análisis de la tabla sales_order_item

Nombre del campo	Resultado	Útil para el análisis
item_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
Order_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
parent_item_id	Solo el 23% de los datos es válido. No se considera relevante para el análisis.	No
quote_item_id	El 99% de los datos es válido, pero no se considera relevante para el análisis.	No
store_id	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
create_at	Datos 100% validos, solo requiere modificación en el formato de la fecha.	**Si**
update_at	Datos 100% validos, solo requiere modificación en el formato de la fecha.	**Si**
product_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con la tabla catalog_product_entity	**Si**
product_type	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
producto_options	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
Weight	Solo el 1% de los datos es válido. No se considera relevante para el análisis.	No
is_virtual	El 99% de los datos es válido, pero no se considera relevante para el análisis.	No

sku	Datos 100% validos, no se requiere transformación y modificación.	**Si**
Name	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
Description	No se encontraron datos válidos. No es relevante para el análisis.	No
applied_rule_ids	Solo el 12% de los datos es válido, pero se considera necesario para el análisis debido a que hace relación con la tabla salesrule	**Si**
additional_data	No se encontraron datos válidos. No es relevante para el análisis.	No
is_qty_decimal	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
no_discount	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
qty_backordered	No se encontraron datos válidos. No es relevante para el análisis.	No
qty_canceled	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
qty_invoiced	Datos 100% validos, se requiere que se transformen los datos de decimal a enteros	**Si**
qty_ordered	Datos 100% validos, se requiere que se transformen los datos de decimal a enteros	**Si**
qty_refunded	Datos 100% validos, se requiere que se transformen los datos de decimal a enteros	**Si**
qty_shipped	Datos 100% validos, pero no se requiere para el análisis	No
base_cost	El 73% de los datos es válido, pero no se considera relevante para el análisis.	No
Price	Datos 100% validos, no se requiere transformación y modificación.	**Si**
base_price	Datos 100% validos, pero no se requiere para el análisis	No
original_price	Datos 100% validos, no se requiere transformación y modificación.	**Si**
base_original_price	El 73% de los datos es válido, pero no se considera relevante para el análisis.	No
tax_percent	Datos 100% validos, no se requiere transformación y modificación.	**Si**

tax_amount	Datos 100% validos, pero no se requiere para el análisis	No
base_tax_amount	Datos 100% validos, pero no se requiere para el análisis	No
tax_invoiced	Datos 100% validos, pero no se requiere para el análisis	No
base_tax_invoiced	Datos 100% validos, pero no se requiere para el análisis	No
discount_percent	Datos 100% validos, pero no se requiere para el análisis	No
discount_amount	Datos 100% validos, no se requiere transformación y modificación.	**Si**
base_discount_amount	Datos 100% validos, pero no se requiere para el análisis	No
discount_invoiced	Datos 100% validos, pero no se requiere para el análisis	No
base_discount_invoiced	Datos 100% validos, pero no se requiere para el análisis	No
amount_refunded	Datos 100% validos, no se requiere transformación y modificación.	**Si**
base_amount_refunded	Datos 100% validos, pero no se requiere para el análisis	No
row_total	Datos 100% validos, pero no se requiere para el análisis	No
base_row_total	Datos 100% validos, pero no se requiere para el análisis	No
row_invoiced	Datos 100% validos, pero no se requiere para el análisis	No
base_row_invoiced	Datos 100% validos, pero no se requiere para el análisis	No
row_weight	Datos 100% validos, pero no se requiere para el análisis	No
base_tax_before_discount	No se encontraron datos válidos. No es relevante para el análisis.	No
tax_before_discount	No se encontraron datos válidos. No es relevante para el análisis.	No
ext_order_item_id	No se encontraron datos válidos. No es relevante para el análisis.	No
locked_do_invoice	No se encontraron datos válidos. No es relevante para el análisis.	No
locked_do_ship	No se encontraron datos válidos. No es relevante para el análisis.	No

price_incl_tax	El 77% de los datos es válido, pero no se considera relevante para el análisis.	No
base_price_incl_tax	El 77% de los datos es válido, pero no se considera relevante para el análisis.	No
row_total_incl_tax	El 77% de los datos es válido, pero no se considera relevante para el análisis.	No
base_row_total_incl_tax	El 77% de los datos es válido, pero no se considera relevante para el análisis.	No
discount_tax_compensation_ amount	El 77% de los datos es válido, pero no se considera relevante para el análisis.	No
base_discount_tax_ compensation_amount	El 96% de los datos es válido, pero no se considera relevante para el análisis.	No
discount_tax_ compensation_invoiced	El 96% de los datos es válido, pero no se considera relevante para el análisis.	No
base_discount_tax_ compensation_invoiced	El 96% de los datos es válido, pero no se considera relevante para el análisis.	No
discount_tax_ compensation_refunnded	Solo el 11% de los datos es válido. No se considera relevante para el análisis.	No
base_discount_tax_ compensation_refunded	Solo el 11% de los datos es válido. No se considera relevante para el análisis.	No
tax_canceled	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
discount_tax_ compensation_canceled	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
tax_refunded	Solo el 11% de los datos es válido. No se considera relevante para el análisis.	No
base_tax_refunded	Solo el 11% de los datos es válido. No se considera relevante para el análisis.	No
discount_refunded	Solo el 11% de los datos es válido. No se considera relevante para el análisis.	No
base_discount_refunded	Solo el 11% de los datos es válido. No se considera relevante para el análisis.	No
weee_tax_applied	El 74% de los datos es válido, pero no se considera relevante para el análisis.	No
weee_tax_applied_amount	No se encontraron datos válidos. No es relevante para el análisis.	No
weee_tax_applied_ row_amount	No se encontraron datos válidos. No es relevante para el análisis.	No
weee_tax_disposition	No se encontraron datos válidos. No es relevante para el análisis.	No
weee_tax_row_disposition	No se encontraron datos válidos. No es relevante para el análisis.	No

base_weee_tax_ applied_amount	No se encontraron datos válidos. No es relevante para el análisis.	No
base_weee_tax_applied_ row_amnt	No se encontraron datos válidos. No es relevante para el análisis.	No
base_weee_tax_disposition	No se encontraron datos válidos. No es relevante para el análisis.	No
base_weee_tax_ row_disposition	No se encontraron datos válidos. No es relevante para el análisis.	No
gift_message_id	No se encontraron datos válidos. No es relevante para el análisis.	No
gift_message_available	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
free_shipping	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No

Tabla 37: Data Profiling – Tabla: sales_order_item

Nombre de la tabla: salesrule

- De los 24 campos de los que está conformada la tabla, se seleccionaron 6 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
- No se encontraron valores duplicados en el identificador rule_id
- Solo se encontró dos columnas con datos nulos, todos los demás no poseen datos nulos

A continuación, se presenta la *Tabla 38* con el análisis de la tabla salesrule

Nombre del campo	Resultado	Útil para el análisis
rule_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
name	Datos 100% validos, no se requiere transformación y modificación.	**Si**
description	Datos 100% validos, no se requiere transformación y modificación.	**Si**
from_date	Datos 100% validos, solo requiere modificación en el formato de la fecha.	**Si**
to_date	Datos 100% validos, solo requiere modificación en el formato de la fecha.	**Si**
uses_per_customer	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
is_active	Datos 100% validos, se requiere especificar si es 1 activo, si es 0 inactivo	**Si**

conditions_serialized	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
actions_serialized	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
stop_rules_processing	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
is_advanced	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
product_ids	No se encontraron datos válidos. No es relevante para el análisis.	No
sort_order	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
simple_action	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
discount_amount	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
discount_qty	No se encontraron datos válidos. No es relevante para el análisis.	No
discount_step	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
apply_to_shipping	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
times_used	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
is_rss	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
coupon_type	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
use_aouto_generation	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No

uses_per_coupon	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
simple_free_shipping	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No

Tabla 38: Data Profiling – Tabla: salesrule

Nombre de la tabla: salesrule_coupon

- De los 11 campos de los que está conformada la tabla, se seleccionaron 5 que se consideran útiles para realizar el análisis correspondiente a este proyecto.
  No se encontraron valores duplicados en el identificador coupon_id
- Solo se encontró tres columnas con datos nulos, todos los demás no poseen datos nulos

A continuación, se presenta la *Tabla 39* con el análisis de la tabla salesrule_coupon

Nombre del campo	Resultado	Útil para el análisis
coupon_id	Datos 100% validos, no se requiere transformación y modificación.	**Si**
rule_id	Datos 100% validos, no se requiere transformación y modificación. Útil para relacionar con salesrule	**Si**
code	Datos 100% validos, no se requiere transformación y modificación.	**Si**
usage_limit	Datos 100% validos, no se requiere transformación y modificación.	**Si**
usage_per_customer	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
times_used	Datos 100% validos, no se requiere transformación y modificación.	**Si**
expiration_date	No se encontraron datos válidos. No es relevante para el análisis.	No
is_primary	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No
created_at	No se encontraron datos válidos. No es relevante para el análisis.	No
type	Sin inconsistencia en los datos, pero no se considera relevante para al análisis.	No

generated_by_dotmailer	No se encontraron datos válidos. No es relevante para el análisis.	No

Tabla 39: Data Profiling – Tabla: salesrule_coupon

b. Capitulo II: Análisis y diseño de la propuesta de solución

1. Metodología de trabajo

Actualmente muchas organizaciones que han incurrido dentro del mundo de la tecnología y la informática ven estas nuevas herramientas tecnológicas como soporte para muchas de las actividades que se realizaban a mano o eran parte de procedimientos largos e ineficientes. Los Data Warehouse son parte de estas nuevas tecnologías debido a que ayudan al análisis de los datos almacenada dentro de estas organizaciones y contribuyen a la toma de decisiones empresariales. Por ello, surge la necesidad de crear Data Warehouse confiables y que brinden solución a las dudas y problemas que estas organizaciones presentan; razón por la cual, nacen metodologías dirigidas exclusivamente para su correcto análisis y modelado, siento las más resonadas: la metodología de Ralph Kimball y la metodología de Bill Inmon.

Para el desarrollo de este proyecto se tomará como base la metodología de Ralph Kimball, el cual, brevemente, consiste en crear una base de datos (o data mart) por cada proceso de negocio y que tiene como objeto principal que su arquitectura general se integre a múltiples bases de datos para que sean interoperables.

En el modelado de datos, Ralph Kimball se centra más en el diseño de las tablas en lugar del análisis entidad-relación. Por ello establece el diseño de las tablas como:

- **Tablas de hecho:** son las que contienen las métricas
- **Tablas de dimensión:** son las que contienen los atributos de las métricas en las tablas de hecho.

Cuatro pasos del proceso para el diseño dimensional

Kimball recomienda una metodología de desarrollo que es única para el almacenamiento de datos. Se trata en un enfoque ascendente, que en el caso de almacenes de datos significa construir el DataMart a la vez. Las cuatro fases del diseño dimensional son:

- Seleccionar el proceso de negocio
- Establecer la granularidad o nivel de detalle
- Elegir las dimensiones
- Identificar los hechos

Tomando en cuenta la metodología de Ralph Kimball, para el desarrollo de este proyecto realizó el siguiente proceso:

- Se realizo una evaluación a la base de datos origen para seleccionar el Dataset correspondiente que abarque los procesos de negocio de ventas y los procesos de negocio de inventario, determinado la viabilidad de los datos.
- Se realizaron los cuatro pasos para el diseño de modelos dimensionales para cada uno de los procesos de negocio
- Se realizó una matriz de bus para relacionar cada una las dimensiones identificadas con cada uno de las tablas de hechos
- Se realizo el modelo dimensional de acuerdo con el análisis realizado en los pasos anteriores

2. Descripción de la propuesta de solución

2.1 Cuatro pasos para diseñar un modelo dimensional

Para crea el modelo dimensional se siguen los siguientes pasos, tanto para el proceso de ventas como el de inventario.

Proceso de ventas.

1. **Selección del proceso de negocio:** Transacciones de ventas
2. **Granularidad:** Producto individual en una venta.
   Las ventas se pueden ver por producto, promoción, cupón, departamento, tipo de producto y día
3. **Identificar dimensiones y tablas de hechos con sus atributos:**

- **DimProducto:** productoKey(SK), productoID(BK), categoriaKey, nombre, descripción, talla, color, material, clima, sku, precio, costo, estado, fechaColumnaEfectiva, fechaColumnaExpiración, estadoColumna.
- **DimPromocion:** promocionKey(SK), promocionID(BK), nombre, descripción, fechaInicio, fechaFin, estaActiva.
- **DimCupon:** cuponKey(SK), cuponID(BK), nombre, descripción, código, cantidadCupones, cantidad Usados, fechaInicio, fechaFin, estado.
- **DimTiempo:** fechaKey(SK), fecha(BK), fechaCompleta, díaDeSemana, numeroDíaDeMes, numeroDiaDelAnio, nombreDia, diaLaboralNoLaboral, numeroDeSemanaAlAnio, numeroDeSemana, fechaInicioDeLaSemana, fechaInicioDeLaSemanaKey, mes, numeroDelMes, nombreMes, anio, trimestre, numeroTrimestre, semestre, numeroSemetre.
- **DimCategoria:** categoriaKey(SK), categoriaID(BK), nombre, tipoProducto, departamento, estado.

- **FactlessProductoPromocion:** productoKey, promocionKey, fechaKey, promocionContador, precioPromocion.
- **FactVentas:** productoKey, cuponKey, promocionKey, fechaKey, numeroOrden, cantidadVendida, precioOriginal, precioVendido, subtotal, totalImpuestos, cantidadDescuentoxCupon, totalVendidoAntesImpuesto, totalVendidoConImpuesto, totalReembolso, cantidadReembolso, montoFinal.

4. **Identificar Hechos:**

   Métricas:
   - El total de ventas por productos en una fecha determinada.
   - La cantidad de unidades vendidas por producto, para una fecha determinada
   - El total de las ventas por producto incluyendo descuentos (promociones y cupones) en una fecha determinada.
   - El total ganado por un producto luego de incluir impuestos, descuentos y devoluciones.

Proceso de inventario.

1. **Selección del proceso de negocio:** Transacciones de Inventario
2. **Granularidad:** Tipo de transacción de un producto

   Las transacciones de inventario se pueden ver por producto, por fuente de inventario y por día
3. **Identificar dimensiones y tablas de hechos con sus atributos:**
   - **DimProducto:** productoKey(SK), productoID(BK), categoriaKey, nombre, descripción, talla, color, material, clima, sku, precio, costo, clasificación, estado, fechaColumnaEfectiva, fechaColumnaExpiración, estadoColumna.
   - **DimTiempo:** fechaKey(SK), fecha(BK), fechaCompleta, díaDeSemana, numeroDíaDeMes, numeroDiaDelAnio, nombreDia, diaLaboralNoLaboral, numeroDeSemanaAlAnio, numeroDeSemana, fechaInicioDeLaSemana, fechaInicioDeLaSemanaKey, mes, numeroDelMes, nombreMes, anio, trimestre, numeroTrimestre, semestre, numeroSemetre.
   - **DimFuenteInventario:** fuenteInventarioKey(SK), fuenteInventarioID(BK), nombre, descripción, país, cuidad.
   - **FactlessProductoFuenteInventario:** fuenteInventarioKey, productoKey, cantidadPorFuente, estado.
   - **InventarioTransaccionFact:** productoKey, fechaKey, fuenteInventarioKey, tipoTransacción, cantidad, umbralFueraStock, pedidosPendientes, cantidadMinimaVenta, cantidadMaximaVenta, cantidadVendida, cantidadComprada, cantidadDevuelta costoPromedio, estaEnStock.

4. **Identificar Hechos:**

Métricas:

- La cantidad de producto que se tiene disponible en inventario para una fecha en específico
- Determinar el costo promedio de un producto para un periodo de tiempo.
- Cantidad de devoluciones de mercadería en inventario.

2.2 Matriz de bus

Con la matriz de bus identificamos todos los procesos de negocio que se van a involucrar dentro del Data Warehouse asumiendo que cada proceso de negocio tendrá su propia tabla de hechos y que cada tabla de hechos tendrá relación con varias dimensiones. En la tabla 40 se puede observar la matrix de bus entre las tablas de hechos y dimensiones.

			Dimensiones comunes				
Procesos de negocio (Fact table)	**DimTiempo**	**DimProducto**	**DimCategoria**	**DimPromocion**	**DimCupon**	**DimFuenteInvetario**	**FactlessProductoFuenteInventario**
FactVentas	**X**	**X**	**X**	**X**	**X**		
FactlessProductoPromocion (Forma parte de la solución de ventas)	**X**	**X**		**X**			
InventarioTransaccionFact	**X**	**X**				**X**	**X**

Tabla 40: Matrix de bus entre las tablas de hechos y dimensiones

2.3 Modelo Dimensional propuesto

Proceso de ventas. El modelo de este proceso lo puede observar en la *figura 9 y* en la *figura 10* pude también observar el modelo dimensional para FactlessProductoPromocion

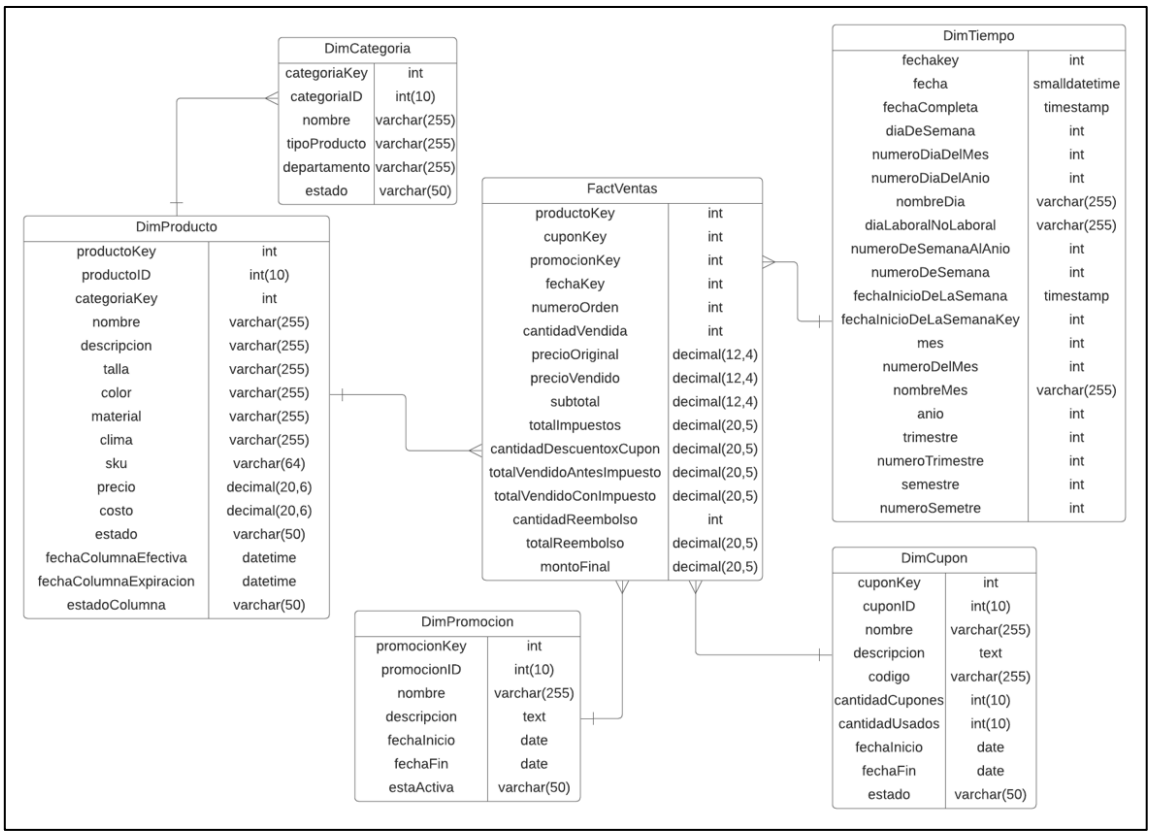


Figura 9: Modelo dimensional para el proceso de negocio ventas

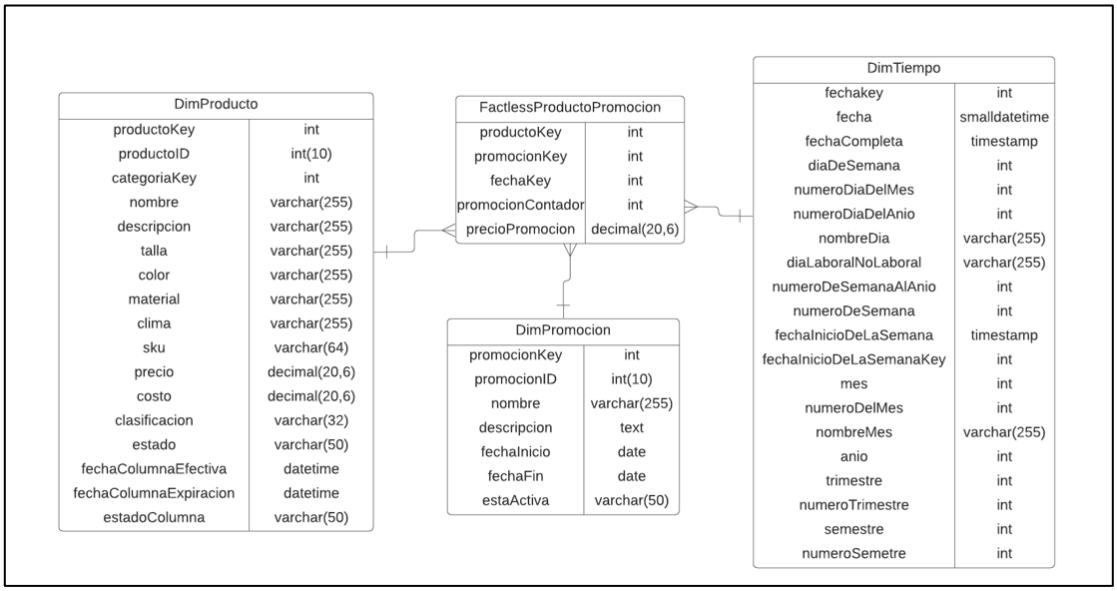


Figura 10: Modelo dimensional para FactlessProductoPromocion

Proceso de inventario.

El modelo de este proceso lo puede observar en la *figura 11*

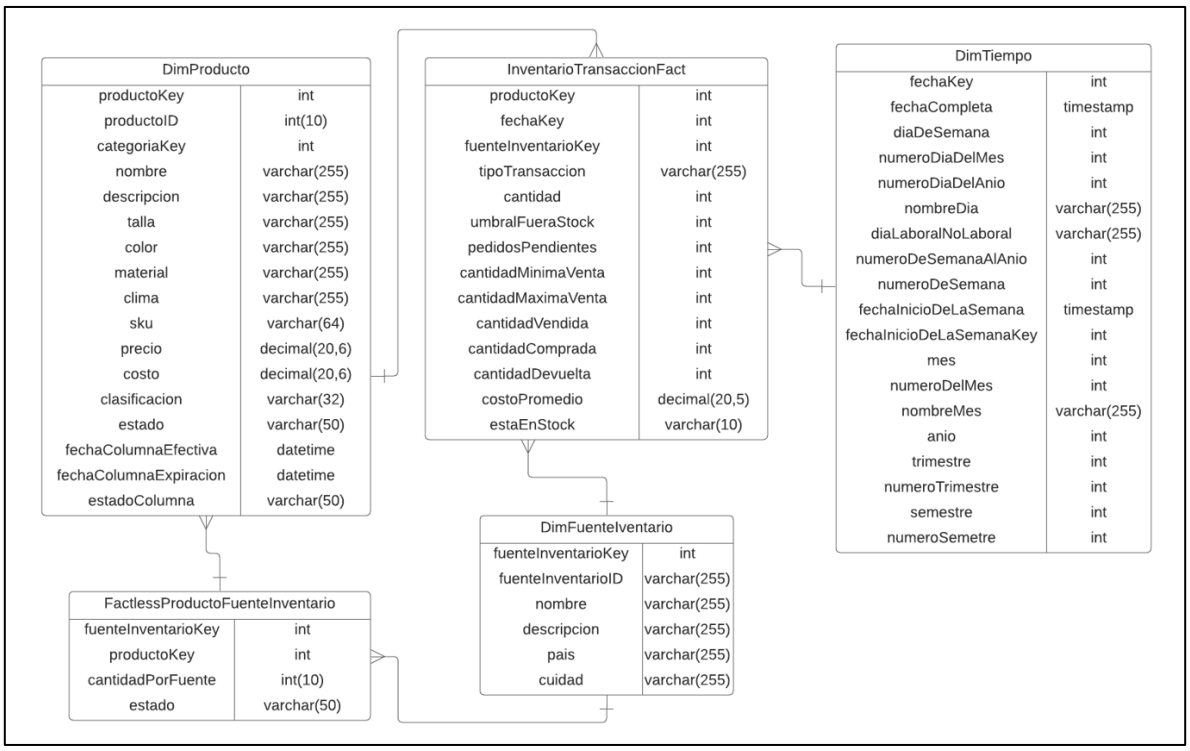


Figura 11: Modelo dimensional para el proceso de negocio de inventarios

3. Descripción de la tecnología a utilizar

DataCleaner

DataCleaner es una herramienta que sirve para analizar la calidad de los datos obtenidos, con capacidad para encontrar patrones y supervisar los valores de los datos.
Está construida para poder manejar pequeñas y grandes cantidades de datos. Es posible diseñar nuestras propias reglas de limpieza de datos y componerlas en múltiples escenarios distintos o bases de datos objetivo, dichas reglas pueden ser: reglas de búsqueda y/o reemplazo, expresiones regulares, coincidencia de patrones (pattern matching) o transformaciones totalmente personalizadas. Su logo lo podemos observar en la *figura 7*.

DBeaver.

DBeaver es un potente software para la gestión de bases de datos, libre y de código abierto para múltiples sistemas operativos. Aunque DBeaver nació en 2010, fruto de su popularidad en la comunidad de código abierto, ha experimentado una rápida expansión de sus características iniciales, incorporando las principales bases de datos SQL y NoSQL. Su logo lo podemos observar e la *figura 8*.

DBeaver permite la mayoría de funcionalidades básicas de cualquier gestor de bases de datos y mucho más. A través de su amigable interfaz podemos:

- Crear todos los componentes de una base de datos: esquemas, tablas, disparadores, funciones, usuarios, roles, etc.
- Realizar consultas SQL y NoSQL
- Crear/Modificar/Eliminar registros
- Exportar y migrar datos
- Generar backups
- Generar datos simulados para realizar pruebas
- Crear diagramas del modelo entidad-relación
- Visualización de información espacial

Talend Open Studio

Talend Open Studio (TOS) es una suite que aporta un conjunto muy complejo, variado y completo de herramientas para llevar a cabo la integración de datos que se ofrece en una versión de código libre. Precisamente por ello, esta es una de las herramientas de integración ETL (extract, transform, load) más utilizadas dentro del mundo Big Data; es más, es la cuarta en la lista después de Informática Powercenter, IBM InfoSphere Datastage y Oracle Data Integrator (ODI).

Por otra parte, esta suite cuenta con un Comunity Edition (CE) totalmente funcional. Además, se puede utilizar una gran cantidad de componentes (más o menos 900) para llevar a cabo una gestión de datos personalizada. De hecho, TOS permite hacer grandes cosas de manera sencilla gracias a esta variedad de servicios. En la *figura 12* se puede apreciar el logo de Talend Open Studio.

Figura 12: Logo de talend Open Studio

MySQL

MySQL es un sistema de gestión de bases de datos relacionales (RDBMS) de código abierto respaldado por Oracle y basado en el lenguaje de consulta estructurado (SQL). MySQL funciona prácticamente en todas las plataformas, incluyendo Linux, UNIX y Windows. Aunque puede utilizarse en una amplia gama de aplicaciones, MySQL se asocia más a menudo con las aplicaciones web y la publicación en línea.

MySQL permite almacenar y acceder a los datos a través de múltiples motores de almacenamiento, incluyendo InnoDB, CSV y NDB. MySQL también es capaz de replicar datos y particionar tablas para mejorar el rendimiento y la durabilidad. Los usuarios de MySQL no tienen que aprender nuevos comandos; pueden acceder a sus datos utilizando comandos SQL estándar.

Para la seguridad, MySQL utiliza un sistema de privilegios de acceso y contraseñas encriptadas que permite la verificación basada en el host. Los clientes de MySQL pueden conectarse a MySQL Server utilizando varios protocolos, incluyendo sockets TCP/IP en cualquier plataforma. MySQL también admite una serie de programas cliente y de utilidad, programas de línea de comandos y herramientas de administración como MySQL Workbench. En la *figura 13* podemos apreciar el logo de MySQL

Figura 13: Logo de MySQL

Magento

Magento es una plataforma de código abierto para comercio electrónico escrita en PHP. Fue desarrollada con apoyo de voluntarios por Varien Inc (ahora Magento Inc), una compañía privada con sede en Culver City, California.

Magento emplea el sistema de base de datos relacional MySQL/MariaDB, el lenguaje de programación PHP, y elementos de Zend Framework. Aplica las prácticas de la programación orientada a objetos y la arquitectura modelo–vista–controlador. También utiliza el modelo entidad–atributo–valor para almacenar los datos. En la *figura 14* podemos apreciar el logo de Magento

Figura 14: Logo de Magento

Power BI

Power BI es un conjunto de herramientas que pone el conocimiento al alcance de todos y nos brinda acceder a nuestros datos de forma segura y rápida, generando grandes beneficios para nosotros y para nuestra empresa. Es un sistema predictivo, inteligente y de gran apoyo, capaz de traducir los datos (simples o complejos) en gráficas, paneles o informes por sus cualidades como la capacidad gráfica de presentación de la información, o la integración de Power Query: el motor de extracción, transformación y carga (ETL) incluido en Excel. En la *figura 15* podemos apreciar el logo de Power BI.

Figura 15: Logo de Power BI

S3 Buckets AWS

Amazon Simple Storage Service (Amazon S3) es un servicio de almacenamiento de objetos que ofrece escalabilidad, disponibilidad de datos, seguridad y rendimiento líderes del sector. Los clientes de todos los tamaños y sectores pueden utilizar Amazon S3 para almacenar y proteger cualquier cantidad de datos para diversos casos de uso, tales como lagos de datos, sitios web, aplicaciones móviles, copia de seguridad y restauración, archivado, aplicaciones empresariales, dispositivos IoT y análisis de big data. Amazon S3 proporciona funciones de gestión para que pueda optimizar, organizar y configurar el acceso a sus datos para satisfacer sus requisitos empresariales, organizativos y de conformidad específicos. En la *figura 16* se puede apreciar la simbología de Bucket en AWS.

Figura 16: S3 Bucket en AWS

Redshift

Amazon Redshift es un producto de almacenamiento de datos que forma parte de la plataforma de computación en la nube más grande Amazon Web Services. Se basa en la tecnología de la empresa de almacenamiento de datos de procesamiento paralelo masivo (MPP) ParAccel (posteriormente adquirida por Actian), para manejar conjuntos de datos a gran escala y migraciones de bases de datos. Redshift se diferencia de otras bases de datos alojadas en Amazon, Amazon RDS, por su capacidad para manejar cargas de trabajo analíticas en grandes conjuntos de datos almacenados por un principio DBMS orientado a columnas. Redshift permite hasta 16 petabytes de datos en un clúster en comparación con el tamaño máximo de 128 tebibytes de Amazon RDS Aurora. En la *figura 17* se puede observar el logo de Amazon Redshift.

Figura 17: Logo Amazon Redshift

4. Diagrama arquitectónico de la solución

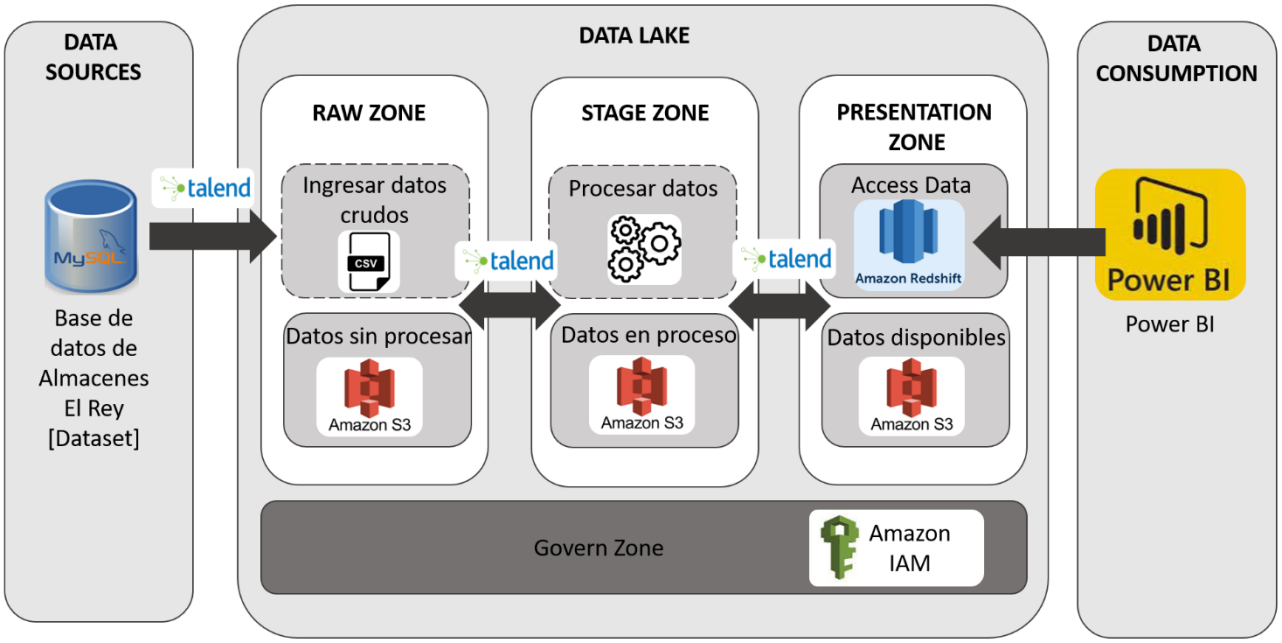


Figura 18: Diagrama arquitectónico de la solución – Arquitectura de Ralp Kimball

La figura 18 se muestra el Diagrama arquitectónico de la solución basado en la arquitectura de Ralp Kimball.

Data sources: Se refiere a las principales fuentes de datos de las cuales se alimenta al Data Lake. Estas fuentes de datos, por lo general, son múltiples, pero para el desarrollo del proyecto se utilizó una única fuente de datos el cual es un Dataset de la base de datos brindada por la tienda Almacenes El Rey, la cual esta alojada en el gestor de base de datos MySQL.

Procesos ETL (Extracción, Transformación y Carga): Se refiere a los procesos con los cuales se realiza el movimiento de datos desde las múltiples fuentes, así como también los procesos para reformatearlos, limpiarlos y cargarlos en otras bases de datos, para el proyecto se hace uso de la herramienta Talend Open Studio para extraer la información desde la fuente origen, así como hacer los procesos correspondientes entre las diferentes zonas que componen el Data Lake.

Data Lake: Es un repositorio centralizado diseñado para almacenar, procesar y proteger grandes cantidades de datos estructurados, semiestructurados o no estructurados. Para nuestro proyecto se utilizó Amazon S3 como repositorio para almacenar todos los archivos .CSV que vaya generando el proceso de Data Warehouse.

Dentro del Data Lake nos encontramos 4 zonas:

- **Raw Zone:** Área donde se almacenan los datos crudos extraídos de la base de datos origen. Estos datos son almacenados sin ningún tipo de transformación y conservando su formato original, tal cual están almacenados en las fuentes de datos. Para el proyecto se utilizó Amazon S3 dentro de la carpeta denominada RAW, donde se guardan los datos crudos en formato .CSV.
- **Stage Zone:** Área donde se almacenan los datos que están siendo procesados o han sufrido alguna transformación final (es decir, han sido limpiados, filtrados, se les ha convertido el tipo de dato o formato, etc, entre otro tipo de transformaciones), estos datos serán utilizados para la construcción de las respectivas tablas de hechos y dimensiones. Para el proyecto se utilizó Amazon S3 dentro de la carpeta denominada STAGE, donde se guardan los datos en proceso en formato .CSV.
- **Presentation Zone:** Área donde se almacenan los datos listos para su respectivo consumo por aplicaciones de BI. Estos datos ya están limpios, filtrados, transformados y organizados en las respectivas tablas de hechos y dimensiones. Para el proyecto se utilizó Amazon S3 dentro de la carpeta denominada PRESENTATION, donde se guardan los datos listos para ser consumidos en formato .CSV.

    Access Data: Dentro del área de Presentation Zone se hace uso de la herramienta Amazon Redshift la cual se encarga guardar la estructura del modelo dimensional y sus datos procesados y listos. A través de esta herramienta se da acceso a los datos para que diferentes aplicaciones BI puedan consumirlos.

- **Govern Zone:** Área donde se administran las reglas, políticas y usuarios que tendrán acceso a tanto a los datos que se utilizarán para la construcción del Data Warehouse, como a la administración de las diferentes herramientas utilizadas dentro del proceso. Es la capa de seguridad que se aplica en cada área que compone el Data Lake. Para el proyecto se hace uso de Amazon IAM.

Data consumption: Área donde se encuentran las aplicaciones BI las cuales son las que consumen los datos procesados por el Data Warehouse para dar respuesta a los requerimientos analíticos del negocio. Para el proyecto se hace uso de la herramienta Power BI en la cual se generan los respectivos Dashboards para cada proceso de negocio donde se analizan y diseñan los gráficos y reportes a presentar.

5. Descripción de cada componente de la solución.

En este apartado se describirán y demostrará algunos de los componentes descritos en la arquitectura anterior junto con un ejemplo de los procesos ETL que se llevaron a cabo para el desarrollo de este proyecto.

1. Data sources: Conexión a la base de datos origen.

Como se menciona a lo largo del documento, la fuente origen de este proyecto se encuentra alojada en el gestor de base de datos MySQL con el nombre de “almacén”.

Por lo cual para poder utilizar esos datos se establece una conexión con esa base de datos origen a través de Talend.

Esta herramienta tiene la opción para guardar diferentes tipos de componentes como: conexiones a bases de datos, esquemas de bases de datos, ficheros de todo tipo, servicios web, conexiones FTP, LDAP, etc., esa opción se llama “Metadatos” la cual se ocupará para guardar la conexión a la base de datos MySQL, la conexión a Redshift y esquemas de los archivos .CSV a utilizar a lo largo del proyecto. En la Figura 19 se puede ver la Conexión a Base de datos origen en Talend

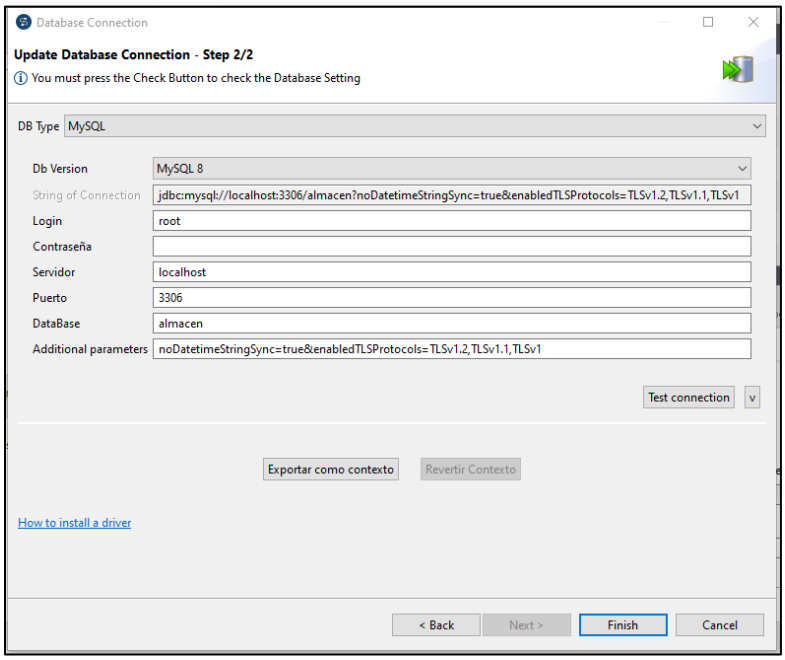


Figura 19: Talend – Conexión a Base de datos origen

2. ETL realizados en el proyecto.

Para la ejecución de los ETL descritos a continuación, la frecuencia deberá de ser cada martes dando inicio a las 00 horas, para que no interfiera con las compras de los clientes, debido a que se detectó que los martes por la madrugada hay menos afluencia de transacciones.

2.1 ETL de extracción de información.

Según la metodología de Ralph Kimball, lo primero a realizar es la extracción de información desde la base de datos origen, estos datos simplemente se extraerán sin ningún tipo de modificación o alteración.

El proceso de extracción de información que se muestra en la *Figura 20* es el mismo proceso que se realizó para cada una de las tablas que componen el Dataset.

Este proceso se realiza entre la zona de **Data Source** a **Raw Zone.**

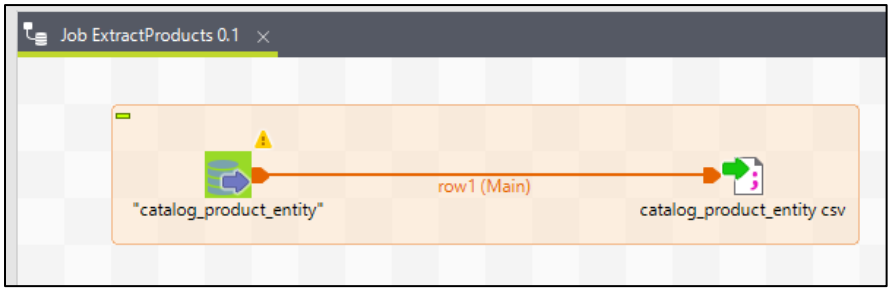


Figura 20: ETL de extracción de información. Catalog_product_entity

Este proceso está compuesto por los componentes:

- tDBInput el cual hace referencia a la conexión de base de datos antes realizada, pero especificando el nombre de la tabla se desea extraer, además de especificar los campos con la sentencia SQL y obtener el esquema con la opción "Guess Schema". En la *Figura 21* podemos observar el tDBinput-Extracción de datos

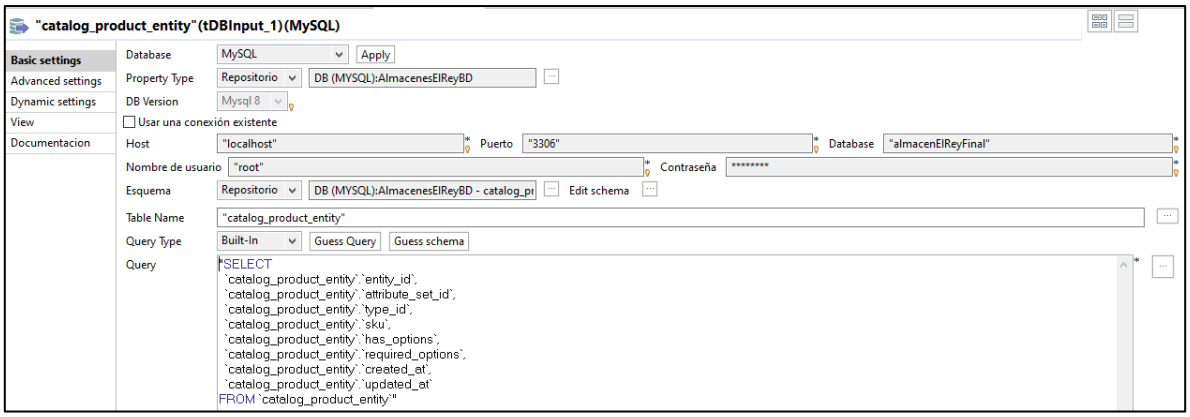


Figura 21: tDBinput-Extracción de datos

- tFileOutputDelimited, este componente genera un archivo .CSV al cual se le debe especificar una ruta donde va a almacenarse. En la *Figura 22* podemos observar tFileOutputDelimited -Extracción de datos

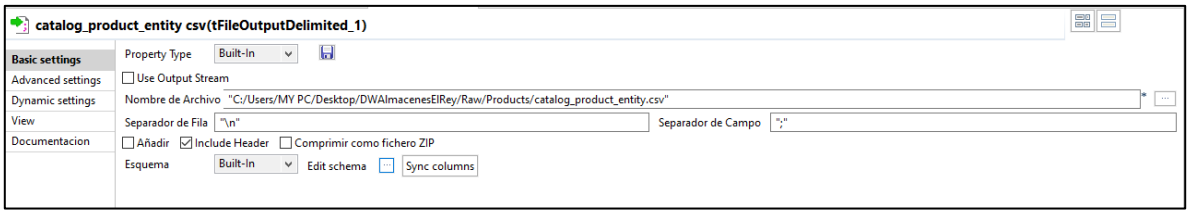


Figura 22: tFileOutputDelimited -Extracción de datos

2.2 ETL de limpieza y transformación de datos

Luego de la extracción de información se procede a la limpieza y transformación de los datos, en cuyo caso los procesos pueden ser diferentes y variantes dependiendo de la tabla que se esté analizando, por ejemplo, pueden existir tablas que requieran limpieza de nulos y transformación de sus tipos de dato de decimal a entero, en otros solo será necesario uno de estos procesos y en otros casos no se requiera realizar ningún proceso de limpieza y transformación debido a que los datos ya vienen limpios o se usarán como los proporciona el Dataset origen.

El proceso de limpieza y transformación que se muestra en la *Figura 23* es para el caso de la tabla salesOrderItem.

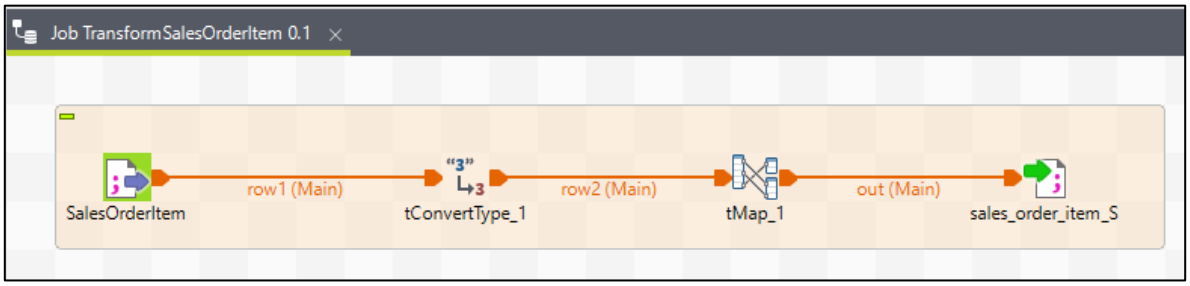


Figura 23: ETL limpieza y transformación. salesOrderItem

Este proceso está compuesto por:

- tFileInputDelimited llamado “SalesOrderItem” el cual es un archivo de entrada .CSV, este archivo contiene los datos obtenidos por el ETL: extracción de datos, es decir, sin ninguna modificación.
- tConvertType quien se encarga de definir los tipos de datos y convertir los espacios en blanco encontrados en el archivo “SalesOrderItem” en nulos para una mejor identificación y tratamiento en Talend. El cual se puede observar en la *figura 24*

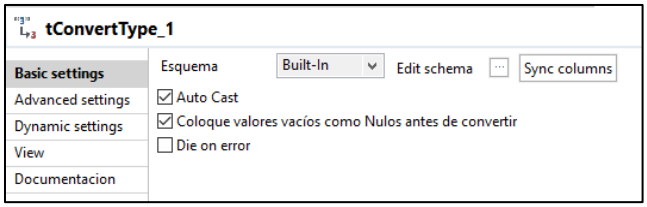


Figura 24: tConverType

- tMap componente donde se pueden realizar variedad de procesos entre los cuales están: el filtrado de campos y registros, el uso de variables y expresiones intermedias, transformación de tipos de datos o formatos de fecha, realizar joins entre tablas, entre otros. Se puede observar en la *figura 25*

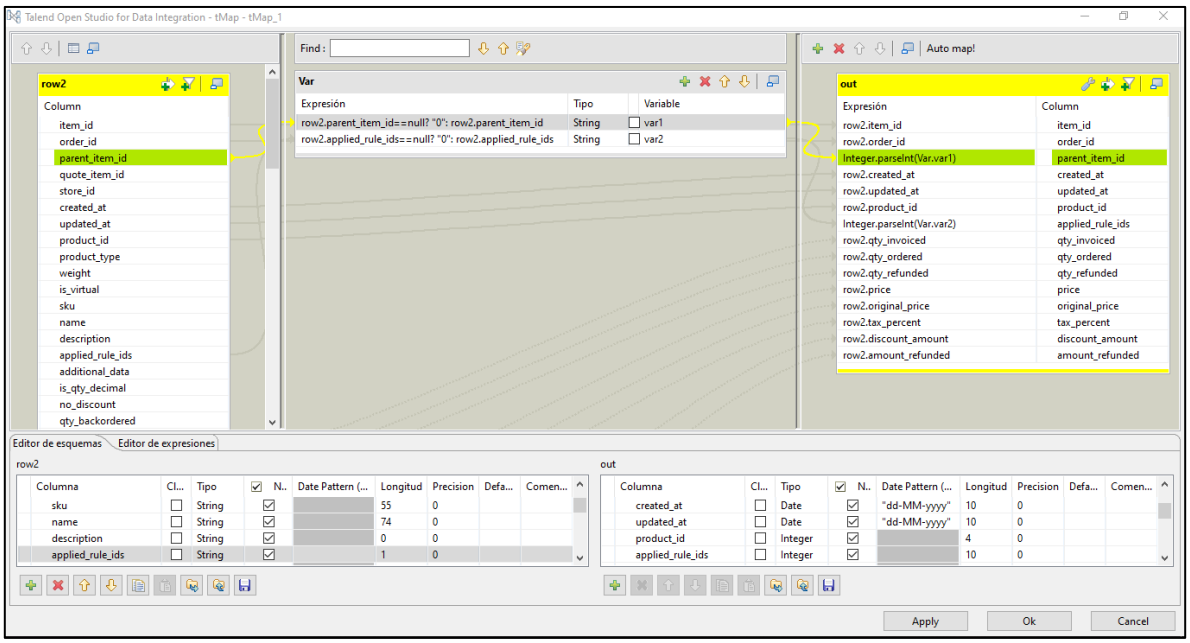


Figura 25: Tmap - salesOrderItem

Para este caso se utilizaron las expresiones y variables intermedias para realizar la limpieza de valores nulos y reemplazarlos por valores que sea válidos. Se pueden apreciar esas expresiones en la *Figura 26*

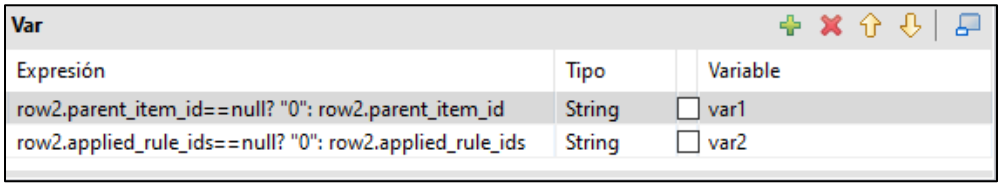


Figura 26: Tmap – expresiones y variables intermedias

Así como también, se realizaron conversiones de datos, en este caso se convirtió la variable string "parent_item_id" al tipo de dato entero. Se pueden observar algunas conversiones de datos en la *figura 27*

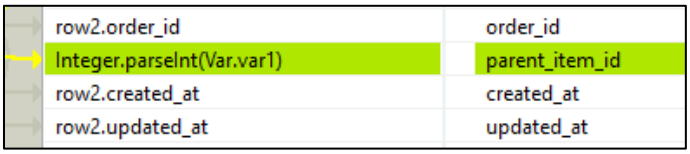


Figura 27: Tmap – conversión de datos

- tFileOutputDelimited, este componente genera un archivo .CSV con los datos limpios y transformados.

2.3 ETL cálculos y preparación de datos

Después de ejecutar los procesos de limpieza y transformación se realizan los procesos de cálculos y preparación de datos los cuales consisten en generar información a través de expresiones lógicas o matemáticas, volver a filtrar campos o registros, realizar joins y uniones entre tablas, ejecutar Jobs a detalle, entre otros procesos que se tuvieron que realizar a algunas tablas para obtener los datos listos para su consumo por aplicaciones de BI.

Los procesos ETL que conforman esta parte, son los procesos más largos y de los que se requirió mayor análisis, debido a que, por la recursividad de algunos datos y la normalización de la base de datos origen, se requirió realizar varios procesos para lograr llegar a la forma final de las tablas de hechos y dimensiones.

El proceso de cálculo y preparación de datos que se muestra en la *Figura 28* es para el caso de la tabla salesOrderItem.

Nota: este ejemplo es solo una muestra del proceso completo que se llevó a cabo para obtener la estructura final de la tabla DimCategoria.

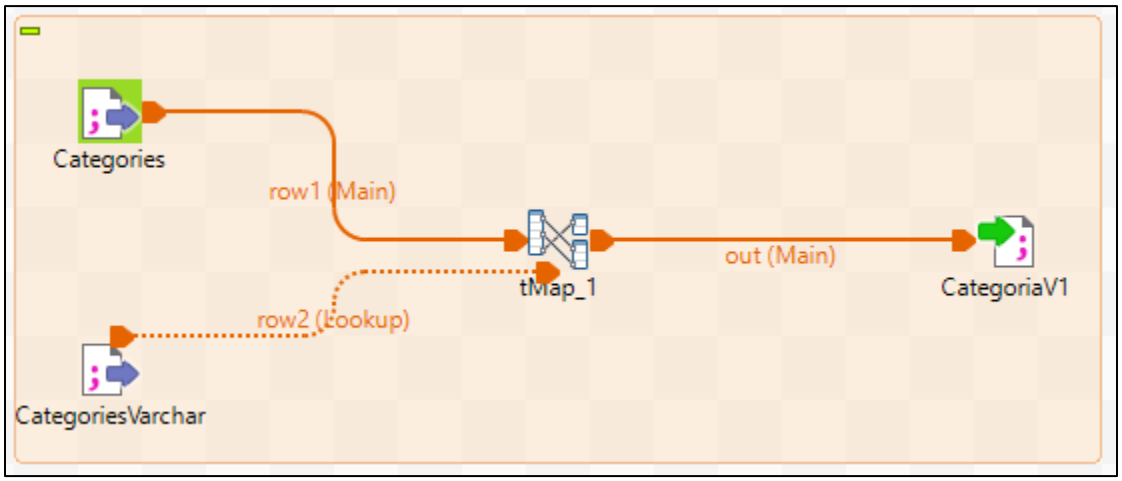


Figura 28: ETL calculo y preparación de DimCategoria

Este proceso está formado por:

- Dos tFileInputDelimited los cuales son dos archivos .CSV limpios y transformados.
- tMap con el cual se realizan, para este caso:

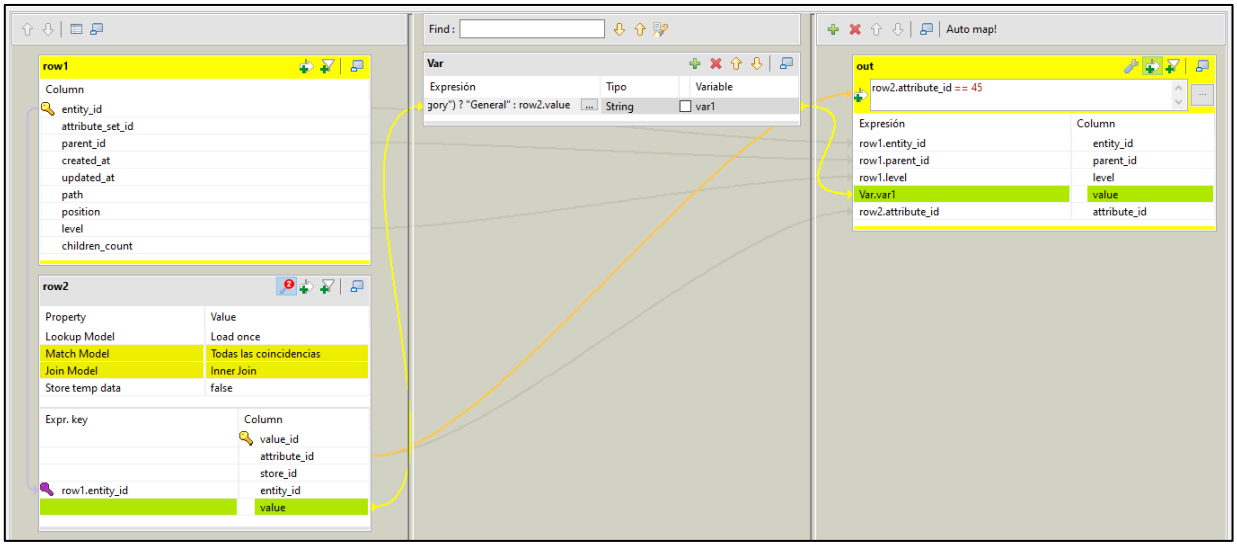


Figura 29: Tmap- DimCategoria

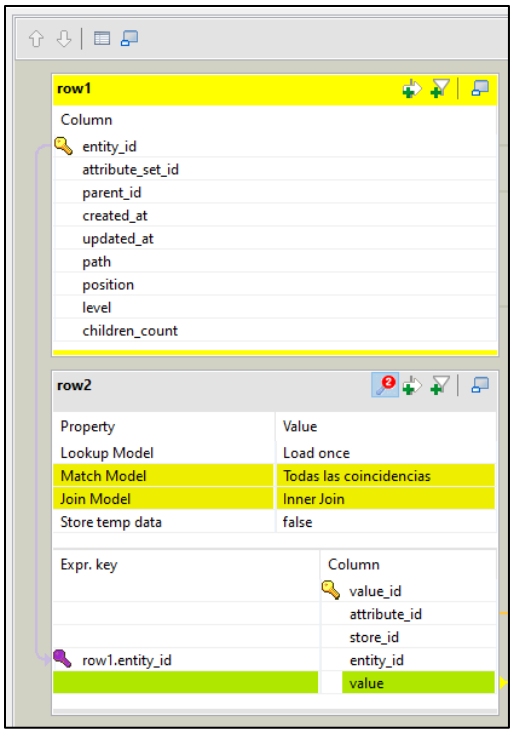


Figura 31: Tmap-Ejem. Inner Join

Procesos de Inner Join (como se muestra en la *Figura 31*) o Left Outer Join para unir archivos con información requerida.

Así como también, formulación de expresiones y variables intermedias que apoyan para esclarecer información, como se muestra en la *figura 30*.

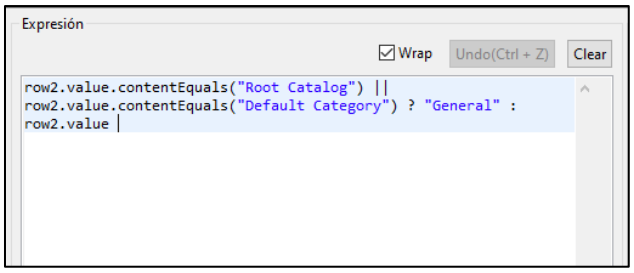
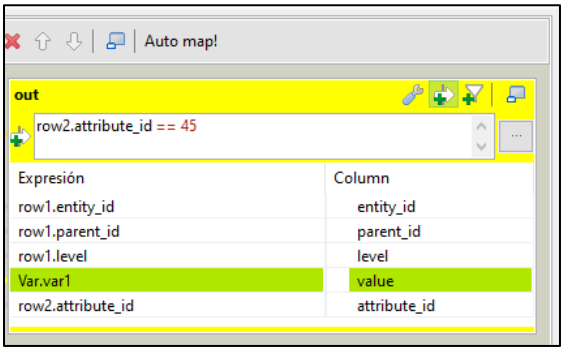


Figura 30: Tmap-Ejem. Expresión



Además de aplicar filtros a los registros, ya que no todos los datos dentro de las tablas son requeridos, como se muestra en la *figura 32*.

Figura 32: Tmap-Ejem. Filtros a registros

- tFileOutputDelimited, este componente genera un archivo .CSV con los datos, para este caso, medianamente preparados para su consumo.

Todo este proceso se realizó en algunas tablas para formar conjuntos de datos que servirán para hacer otros conjuntos de datos y realizar otros análisis, o para llegar a la presentación final de las tablas de hechos y dimensiones correspondientes.

2.4 ETL de carga y descarga de datos

- **ETL: descarga de archivos de S3**

El proceso de descarga de datos se realiza para las zonas de Stage y Presentation, debido a que estas zonas requieren la información de la zona anterior para poder desarrollar sus procesos ETL, es decir, la zona de Stage necesita los archivos generados por la zona Raw y la Zona Presentation necesita los archivos generados por la zona Stage.

El proceso de descarga de archivos de S3 que se muestra en la *Figura 33* es para la zona de stage la cual requiere de los datos almacenados en S3, dentro de la carpeta raw/.

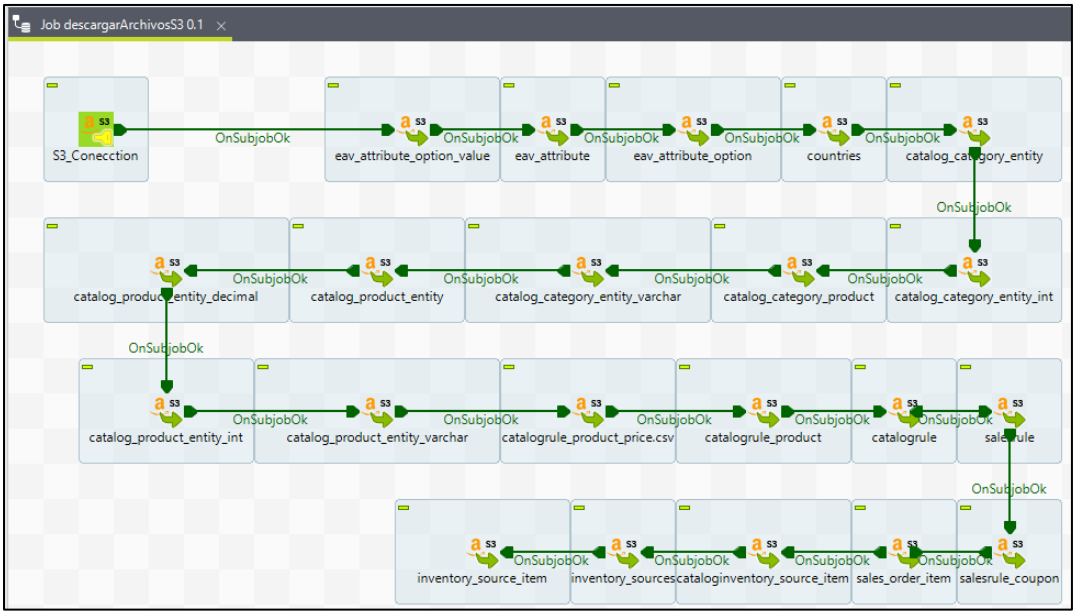


Figura 33: ETL-Descarga de archivos de S3

El proceso este compuesto por los siguientes componentes:

- tS3Connection, este componente sirve para configurar la conexión con Amazon S3 colocando la llave de acceso, la llave secreta y la región del repositorio.
- tS3Get, para este componente se debe especificar desde que bucket o repositorio se ira a descargar la información, el nombre del archivo .CSV que se obtendrá y en que carpeta del sistema operativo se almacenará. Para el caso de descarga, hace referencia a la carpeta raw, para obtener los datos, se puede observar en la *figura 34*.

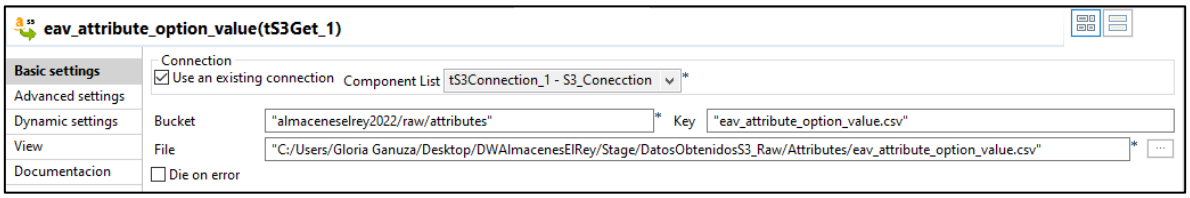


Figura 34: tS3Get

Cada uno de los componentes tS3Get cuenta con la misma lógica.

- **<u>ETL: carga de archivos a S3</u>**

Para cada una de las zonas que componen el Data Lake (raw, stage y presentation) se realizo este proceso el cual consiste en cargar los datos, de la zona que se este analizando en ese momento, y almacenarlos al repositorio de datos en Amazon S3. Este proceso se realiza luego de:

- Finalizar los procesos de ETL: extracción de datos (procesos entre BD origen - raw)
- Finalizar los procesos de ETL: limpieza y transformación (procesos entre raw - stage)

- Finalizar los procesos de ETL: cálculos y presentación de datos (stage - presentation)

El proceso de carga de archivos a S3 que se muestra en la *Figura 35* es luego de haber finalizado los procesos de la zona de stage. Este proceso es similar en el caso de la zona raw, pero para la zona de presentación es diferente, debido a que solo se suben los archivos .CSV.

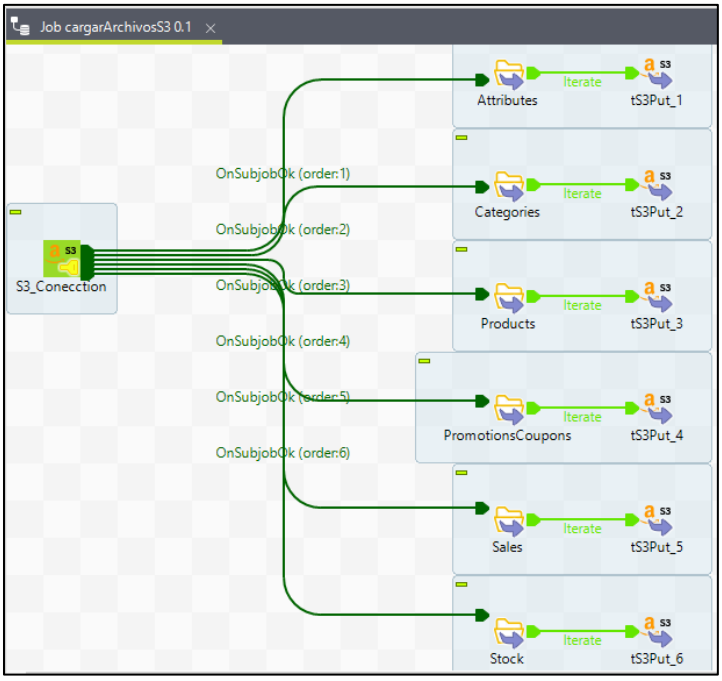


Figura 35: ETL-Carga de archivos a S3

Este proceso está compuesto por los siguientes componentes:

- tS3Connection, este componente sirve para configurar la conexión con Amazon S3 colocando la llave de acceso, la llave secreta y la región del repositorio, se puede observar el la *figura 36*

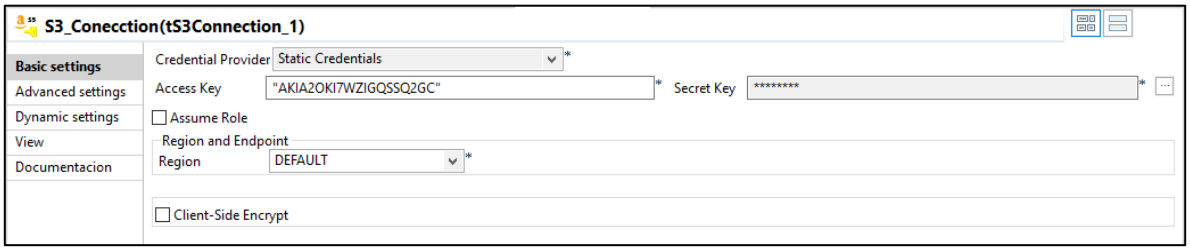


Figura 36: tS3Connection

- tFileList, el cual es un componente que hace referencia a la carpeta dentro del sistema operativo donde se han almacenado los archivos .CSV que se generan en los procesos ETL anteriores, se puede observar en la *figura 37*.

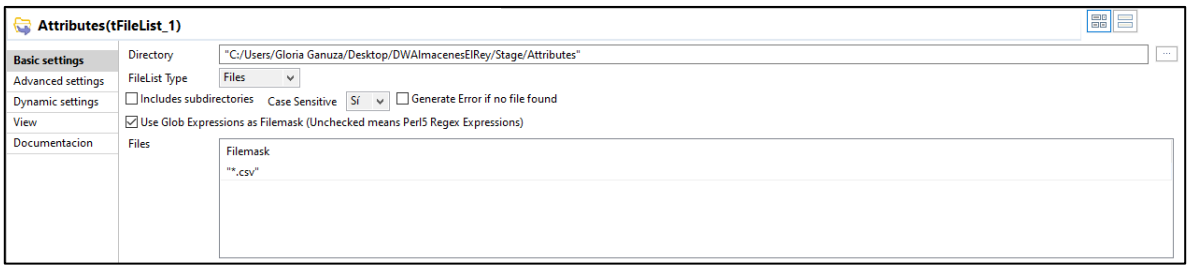


Figura 37: tFileList – Tabla: Attributes

- tS3Put, componente donde se configura en que bucket se ira a almacenar los datos. Para ello se especifica el nombre del bucket o repositorio, en este caso se incluye el nombre de la carpeta y se coloca el nombre que tendrá el archivo en S3, el cual debe ser único, además, se debe colocar desde que carpeta dentro del sistema operativo se va a subir. Para el caso de carga, hace referencia a la carpeta stage, donde se almacenan los datos generados en esta zona, se puede observar en la *figura 38*.
  En este caso, al utilizar el componente tFileList se puede hacer uso de las variables que este genera, para este caso se utiliza "Current File Name" el cual captura el nombre de cada uno de los archivos que se encuentran en la carpeta configurada en tFileList, lo cual, de forma automática, cargará todos los archivos de esa capeta a S3.

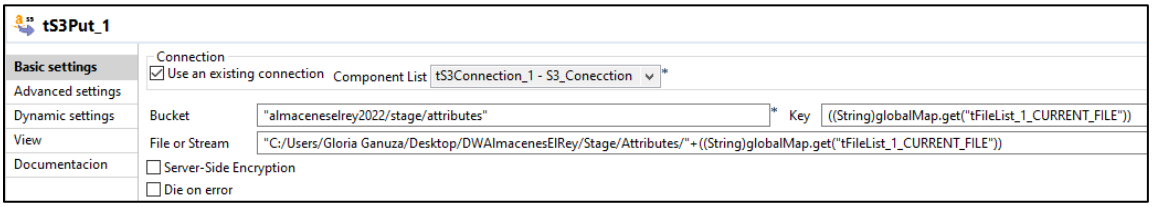


Figura 38: tS3Put

Cada uno de los otros tFileList y tS3Put tienen la misma lógica.

- **<u>ETL: carga de archivos a Redshift</u>**

Este proceso consiste en cargar la información que finalmente está limpia, transformada y en su formato de tablas de hechos y dimensiones a la estructura de la modelo dimensional alojada en Redshift.

Para ello se hace uso del comando COPY el cual se ejecuta en Redshift y está compuesta por:

COPY [*nombre de la tabla*] **FROM** ['*URI de objeto en S3*']
access_key_id ['*llave de acceso del usuario*']
secret_access_key ['*llave secreta del usuario*']
DELIMITER '*;*'
IGNOREHEADER *1*;

COPY: comando para cargar la información de los archivos .CSV alojados en S3 (en presentation zone) a Redshift
[*nombre de la tabla*]: nombre de la tabla en Redshift, para esto es necesario haber ejecutado el script ubicado en anexos (Anexo a.)
[‘*URI de objeto en S3*’]: identificador del recurso que se quiere cargar en Redshift.

access_key_id y secret_access_key: llave de acceso y llave secreta de acceso del usuario asignado.

DELIMITER: delimitador la información del recurso que se quiere cargar. Se coloca el carácter que hace de delimitador.

IGNOREHEADER: para ignorar los encabezados.

El comando de la *Figura 39* corresponde a la tabla dimCategoria. Este mismo proceso se utilizó para llenar cada una de las demás tablas de hechos y dimensiones.

```
COPY dimCategoria FROM 's3://almaceneselrey2022/presentation/DimCategoria.csv'
access_key_id 'AKIA2OKI7WZIDMN7EK7K'
secret_access_key '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'
DELIMITER ';'
IGNOREHEADER 1;
```

Figura 39: comando COPY

3. Data Lake: Repositorio en AWS

El repositorio de datos está alojado en Amazon S3 y tiene de nombre: “almaceneselrey2022”. Se creó una estructura de carpeta para una mejor organización de la información (Ver *Figura 40* y *Figura 41*):

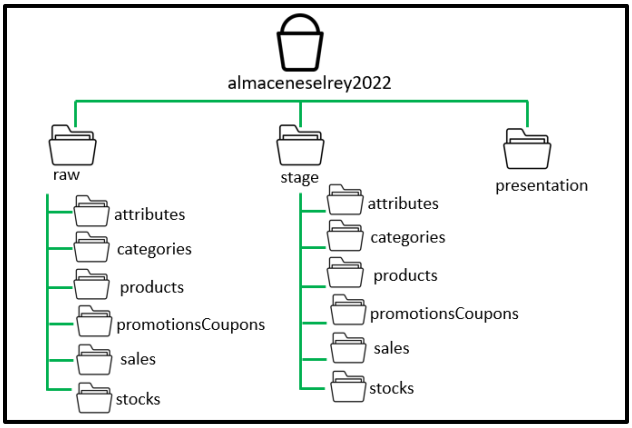


Figura 40: Estructura de carpetas dentro de Amazon S3

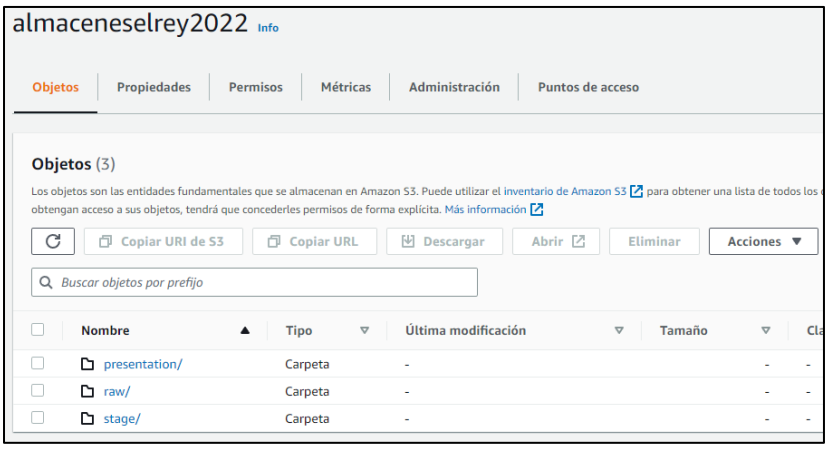


Figura 41: Repositorio en S3

Raw Zone: En la zona de raw o data cruda se almacena la información sin modificar, ni alterar, por lo cual, luego del proceso de ETL: extracción de datos, se realiza el ETL: carga de archivos a S3, los cuales se almacenan de acuerdo a la información a la que están relacionadas y siguiendo el diseño establecido en la *Figura 42*, dentro de la carpeta raw/ se encuentra la siguiente estructura de carpetas:

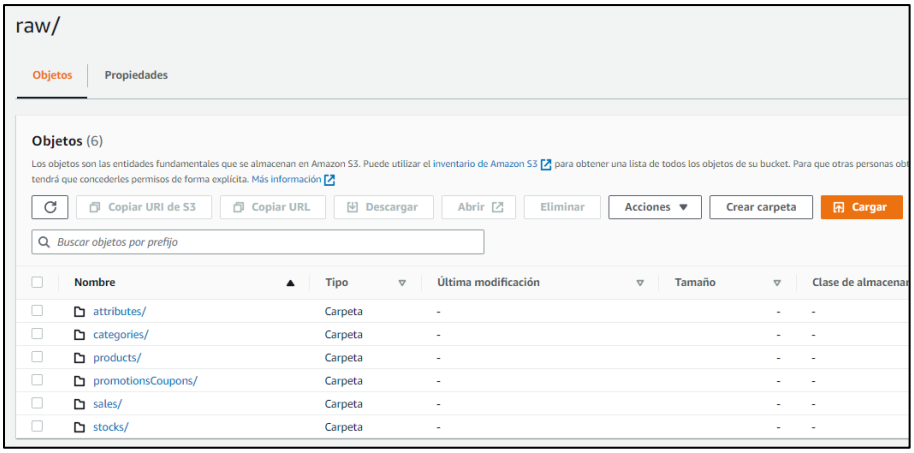


Figura 42: Repositorio en S3-Carpeta raw/

Stage Zone: En la zona de stage se almacena la información que está en proceso ya sea limpieza, filtrado, transformación, entre otros, para ello, primero es necesario ejecutar los procesos de ETL: descarga de archivo de S3 el cual, para este caso, extrae información desde la carpeta raw/ alojada en S3, luego realizar los procesos de ETL: limpieza y transformación de datos y al finalizar se realiza el ETL: carga de archivos a S3, los cuales se almacenan de acuerdo a la información a la que están relacionadas y siguiendo con el diseño establecido en la *Figura 43*, dentro de la carpeta stage/ se encuentra la siguiente estructura de carpetas:

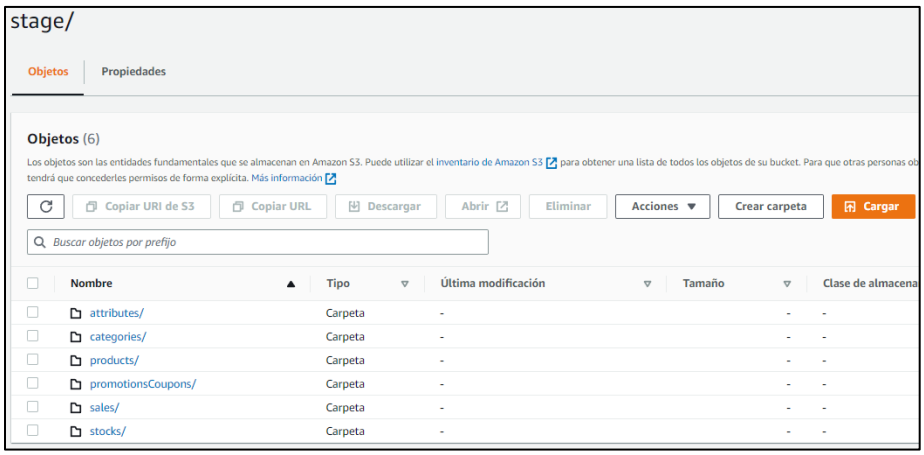


Figura 43: Repositorio en S3-Carpeta raw/

Presentation Zone: En la zona de presentación se almacena la información lista para ser consultada por las aplicaciones de BI, de igual forma que en Stage Zone, primero deben ejecutarse los ETL: descarga de archivos de S3, el cual para este caso, extrae información desde la carpeta stage/ alojada en S3, luego realiza los respectivos procesos de ETL: cálculos y preparación de datos, para finalizar ejecutando los procesos de ELT: carga de archivos a S3, almacenados en sus respectivos formatos de tablas de hechos y tablas de dimensión, siguiendo con el diseño establecido en la *Figura 43*, dentro de la carpeta presentation/ solo se guarda los archivos .CSV y se puede confirmar en la *figura 44*:

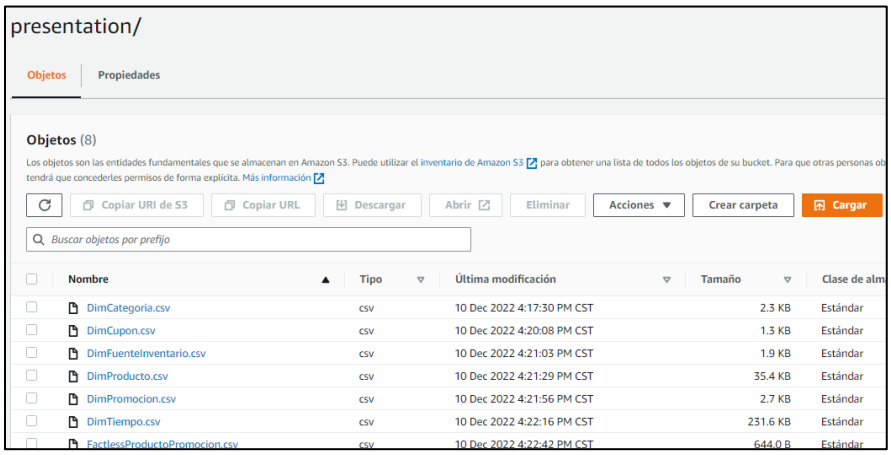


Figura 44: Repositorio en S3-Carpeta presentation/

- **Access Data: Redshift**

Amazon Redshift es la herramienta que se utilizó para guardar la estructura del modelo dimensional y dar acceso a la aplicación de BI, para ello se ejecutó el script que se encuentra en los anexos (Anexo a.) y seguido a ello se ejecutó los procesos ETL: carga de archivos a

Redshift para llenar su estructura. En la *figura 45* se pude observar el modelo dimensional en Redshift

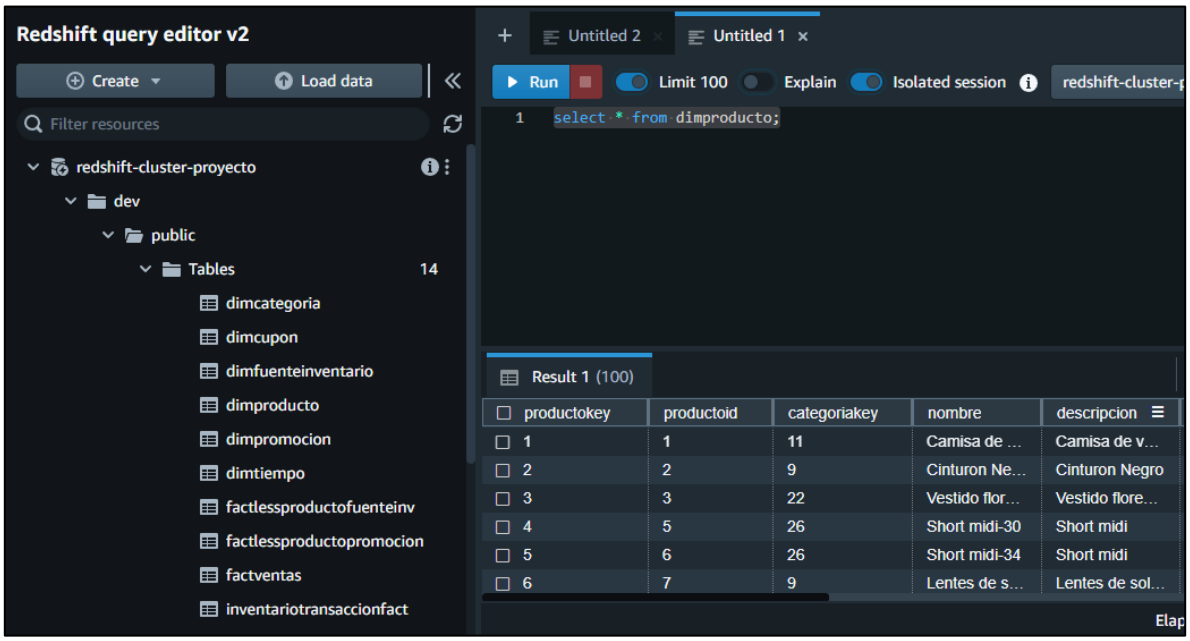


Figura 45: Redshift – Modelo Dimensional

4. Govern Zone: Configuración en Amazon IAM

Para llevar un mejor control del acceso a la información almacenada en el repositorio de datos S3 se crearon 3 usuarios a los cuales se les asignaron diferentes políticas. En la *figura 46* se pueden observar los usuarios IAM creados.

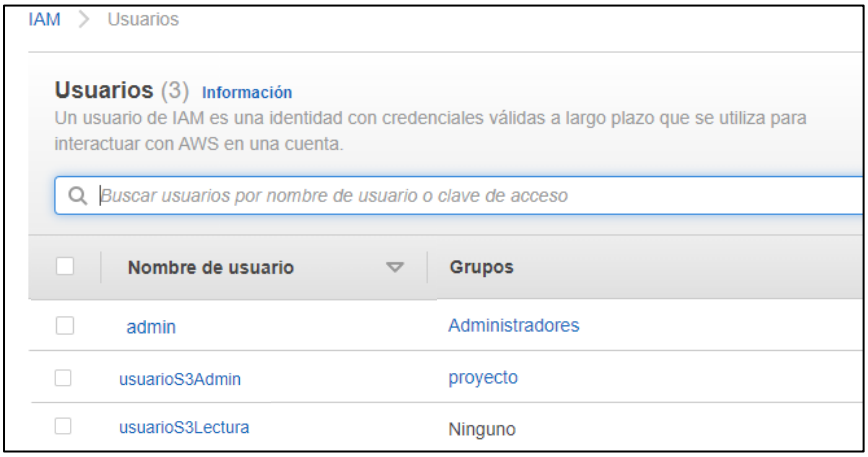


Figura 46: IAM – Usuarios creados

- **Nombre de usuario:** admin
  Descripción: Administrador general de AWS

Política: AdministratorAccess
Descripción de la política: Proporciona acceso completo a los servicios y recursos de AWS.

- **Nombre de usuario:** usuarioS3Admin
  Descripción: usuario administrador de Amazon S3
  Política: AmazonS3FullAccess
  Descripción de la política: Proporciona acceso completo a todos los repositorios a través de la Consola de administración de AWS.

- **Nombre de usuario:** usuarioS3Lectura
  Descripción: usuario solo de lectura para Amazon S3
  Política: AmazonS3ReadOnlyAccess
  Descripción de la política: Proporciona acceso de solo lectura a todos los repositorios a través de la Consola de administración de AWS.

5. Data consumption: Conexión a Power BI

Para finalizar, se desarrolla el consumo de datos listos y limpios en aplicaciones de BI, en este caso se realiza con la herramienta Power Bi Desktop, para lo cual es necesario realizar la conexión desde Power BI hasta Amazon Redshift.

La conexión se realiza ingresando a Power BI, en el ícono 

y buscar la opción de Amazon Redshift para poder ingresar las configuraciones correspondientes y lograr conectarse. En las *figura 47 y 48* puede observar las configuraciones en Power Bi para establecer conexión con el servidor Redshift

Figura 47: Power BI - Servidor Redshift

Figura 48: Power BI – Credenciales Redshift

6. Mapping por tablas

Nombre de la tabla: DimProducto

Tipo de tabla: Tabla de Dimensión

Descripción: Dimensión de producto, que incluye todos los productos de la base transaccional

Llave primaria: productoKey

Llave foránea: categoriaKey, hace referencia a DimCategoria

Política de nulidad: Todos los campos son requeridos.

Política de SCD:

- **SCD tipo 0:** productoID
- **SCD tipo 1:** categoriaKey, nombre, descripción, talla, color, material, clima, sku, precio, estado
- **SCD tipo 2:** costo, fechaColumnaEfectiva, fechaColumnaExpiracion, estadoColumna

La tabla 41 se puede observar el mapping de DimProducto

Nombre de la columna	Descripción	Grupo de atributos	Tipo de dato	Sistema Origen	Tabla Origen	Columna Origen
productoKey	Llave primaria subrogada	Identificador	int(10)	Llave Subrogada (autoincremental)	-	-
productoID	Llave primaria del sistema fuente	Identificador	int(10)	AlmacenesElReyDB	catalog_product_entity	entity_id
categoriaKey	Llave foránea de la categoría al que pertenece el producto	Identificador	int(10)	DWAlmacen	DimCategoria	categoriaKey

nombre	Nombre del producto	Detalles del producto	varchar(255)	AlmacenesElReyDB	catalog_product_entity_varchar	value
descripción	Descripción del producto	Detalles del producto	varchar(255)	AlmacenesElReyDB	catalog_product_entity_varchar	value
talla	Talla o tamaño del producto	Detalles del producto	varchar(255)	AlmacenesElReyDB	eav_attribute_option_value	value
color	Color del producto	Detalles del producto	varchar(255)	AlmacenesElReyDB	eav_attribute_option_value	value
material	Material de producto	Detalles del producto	varchar(255)	AlmacenesElReyDB	eav_attribute_option_value	value
clima	Clima para el cual el producto es óptimo	Detalles del producto	varchar(255)	AlmacenesElReyDB	eav_attribute_option_value	value
sku	Código del producto	Detalles del producto	varchar(64)	AlmacenesElReyDB	catalog_product_entity	sku
precio	Precio del producto	Detalles del producto	decimal(20,6)	AlmacenesElReyDB	catalog_product_entity_decimal	value
costo	Costo del producto	Detalles del producto	decimal(20,6)	AlmacenesElReyDB	catalog_product_entity_decimal	value
estado	Estado disponible/no disponible del producto	Detalles del producto	varchar(50)	AlmacenesElReyDB	catalog_product_entity_int	value
fechaColumnaEfectiva	Fecha vigente para el registro histórico del producto	Estado histórico producto	datetime	Derivado	Calculado ETL	-
fechaColumnaExpiracion	Fecha de expiración para el registro histórico del producto	Estado histórico producto	datetime	Derivado	Calculado ETL	-
estadoColumna	Estado del registro	Estado histórico producto	varchar(50)	Derivado	Calculado ETL	-

	histórico del producto					

Tabla 41: Mapping de la tabla DimProducto

Nombre de la tabla: DimCategoria

Tipo de tabla: Tabla de Dimensión

Descripción: Dimensión de categoría, que incluye todas las categorías de productos de la base transaccional

Llave primaria: categoriaKey

Llave foránea: no hay

Política de nulidad: Todos los campos son requeridos.

Política de SCD:

- **SCD tipo 0:** categoriaID
- **SCD tipo 1:** nombre, tipoProducto, departamento, estado

La tabla 42 se puede observar el mapping de DimCategoria

Nombre de la columna	Descripción	Grupo de atributos	Tipo de dato	Sistema Origen	Tabla Origen	Columna Origen
categoriaKey	Llave primaria subrogada	Identificador	int(10)	Llave Subrogada (autoincremental)	-	-
categoriaID	Llave primaria del sistema fuente	Identificador	int(10)	AlmacenesElReyDB	catalog_category_entity	entity_id
nombre	Nombre de la categoría	Detalles de la categoría	varchar(255)	AlmacenesElReyDB	catalog_category_entity_varchar	value
tipoProducto	Tipo de producto para la categoría	Detalles de la categoría	varchar(255)	AlmacenesElReyDB	catalog_category_entity_varchar	value
departamento	Departamento al que pertenece la categoría	Detalles de la categoría	varchar(255)	AlmacenesElReyDB	catalog_category_entity_varchar	value
estado	Estado de la categoría	Detalles de la categoría	varchar(50)	AlmacenesElReyDB	catalog_category_entity_int	value

Tabla 42: Mapping de la tabla DimCategoria

Nombre de la tabla: DimPromocion

Tipo de tabla: Tabla de Dimensión

Descripción: Dimensión de promoción, que incluye todas las promociones de la base transaccional

Llave primaria: promocionID

Llave foránea: no hay

Política de nulidad: Todos los campos son requeridos.

Política de SCD:

- **SCD tipo 0:** promocionID
- **SCD tipo 1:** nombre, descripción, fechaInicio, fechaFin, estaActiva

La tabla 43 se puede observar el mapping de DimPromocion

Nombre de la columna	Descripción	Grupo de atributos	Tipo de dato	Sistema Origen	Tabla Origen	Columna Origen
promocionKey	Llave primaria subrogada	Identificador	int(10)	Llave Subrogada (autoincremental)	-	-
promocionID	Llave primaria del sistema fuente	Identificador	int(10)	AlmacenesElReyDB	Catalogrule	rule_id
nombre	Nombre de la promoción	Detalles de la promoción	varchar(255)	AlmacenesElReyDB	Catalogrule	name
descripcion	Descripción de la promoción	Detalles de la promoción	Text	AlmacenesElReyDB	Catalogrule	description
fechaInicio	Fecha de inicio de la promoción	detalles de la promoción	datetime	AlmacenesElReyDB	Catalogrule	from_date
fechaFin	Fecha finalización de la promoción	Detalles de la promoción	datetime	AlmacenesElReyDB	Catalogrule	to_date
estaActiva	Estado de la promoción	Detalles de la promoción	varchar(50)	AlmacenesElReyDB	Catalogrule	is_active

Tabla 43: Mapping de la tabla DimPromocion

Nombre de la tabla: DimCupon

Tipo de tabla: Tabla de Dimensión

Descripción: Dimensión de cupón, que incluye todos los cupones de la base transaccional.

Llave primaria: cuponKey

Llave foránea: no hay

Política de nulidad: Todos los campos son requeridos.

Política de SCD:

- **SCD tipo 0:** cuponID
- **SCD tipo 1:** nombre, descripción, código, cantidadCupones, cantidadUsados, fechaInicio, fechaFin, estado

La tabla 44 se puede observar el mapping de DimCupon

Nombre de la columna	Descripción	Grupo de atributos	Tipo de dato	Sistema Origen	Tabla Origen	Columna Origen
cuponKey	Llave primaria subrogada	Identificador	int(10)	Llave Subrogada (autoincremental)	-	-
cuponID	Llave primaria del sistema fuente	Identificador	int(10)	AlmacenesElReyDB	salesrule_coupon	coupon_id
nombre	Nombre del cupón	Detalles del cupón	varchar(255)	AlmacenesElReyDB	salesrule	name
descripcion	Descripción del cupón	Detalles del cupón	text	AlmacenesElReyDB	salesrule	description
codigo	Código identificador del cupón	Detalles del cupón	varchar(255)	AlmacenesElReyDB	salesrule_coupon	code
cantidadCupones	Cantidad de cupones disponibles	Detalles del cupón	int(10)	AlmacenesElReyDB	salesrule_coupon	usage_limit
cantidadUsados	Cantidad de veces que se ha utilizado el cupón	Detalles del cupón	int(10)	AlmacenesElReyDB	salesrule_coupon	times_used
fechaInicio	Fecha inicial que el cupón es válido y podrá ser utilizado	Detalles del cupón	date	AlmacenesElReyDB	salesrule	from_date

fechaFin	Fecha final que el cupón es válido y podrá ser utilizado	Detalles del cupón	date	AlmacenesElReyDB	salesrule_coupon	expiration_date
estado	Estado activo/inactivo del cupón	Detalles del cupón	varchar(10)	AlmacenesElReyDB	salesrule	is_active

Tabla 44: Mapping de la tabla DimCupon

Nombre de la tabla: DimTiempo

Tipo de tabla: Tabla de Dimensión

Descripción: Dimensión de tiempo, que incluye la fecha y su detalle diario, para un período de tiempo

Llave primaria: fechaKey

Llave foránea: no hay

Política de nulidad: Todos los campos son requeridos.

Política de SCD:

- **SCD 0:** Todas las columnas de esta dimensión

La tabla 45 se puede observar el mapping de DimTiempo

Nombre de la columna	Descripción	Grupo de atributos	Tipo de dato	Sistema Origen	Tabla Origen	Columna Origen
fechaKey	Llave primaria (Formato ISO de la fecha)	Identificador	int(10)	Script de fechas	-	-
fecha	Fecha	Detalles de la fecha	smalldatetime	Script de fechas	-	-
fechaCompleta	Fecha en formato timestamp	Detalles de la fecha	timestamp	Script de fechas	-	-
diaDeSemana	Día de la semana correspondiente a la fecha	Detalles de la fecha	int(10)	Script de fechas	-	-
numeroDiaDelMes	Número del día del mes correspondiente a la fecha	Detalles de la fecha	int(10)	Script de fechas	-	-
numeroDiaDelAnio	Número del día del año	Detalles de la fecha	int(10)	Script de fechas	-	-

	correspondiente a la fecha					
nombreDia	Nombre del día del de la semana correspondiente a la fecha	Detalles de la fecha	varchar(255)	Script de fechas	-	-
diaLaboralNoLaboral	Si es día laboral o no laboral de la semana	Detalles de la fecha	varchar	Script de fechas	-	-
numeroSemanaAlAnio	Número de semana en el año correspondiente a la fecha	Detalles de la fecha	int(10)	Script de fechas	-	-
numeroDeSemana	Número de semana desde el inicio general correspondiente a la fecha	Detalles de la fecha	int(10)	Script de fechas	-	-
fechaInicioDeLaSemanaKey	Identificador del inicio de semana de la fecha correspondiente (Formato ISO)	Detalles de la fecha	int	Script de fechas	-	-
fechaInicioDeLaSemana	Fecha de inicio de la semana correspondiente a la fecha	Detalles de la fecha	timestamp	Script de fechas	-	-
mes	Número del mes en el año correspondiente a la fecha	Detalles de la fecha	int(10)	Script de fechas	-	-
numeroDelMes	Número del mes desde el inicio general correspondiente a la fecha	Detalles de la fecha	int(10)	Script de fechas	-	-
nombreMes	Nombre del mes correspondiente a la fecha	Detalles de la fecha	varchar(255)	Script de fechas	-	-
anio	Año correspondiente a la fecha	Detalles de la fecha	int(10)	Script de fechas	-	-

trimestre	Número trimestre correspondiente a la fecha	Detalles de la fecha	int(10)	Script de fechas	-	-
numeroTrimestre	Número del trimestre desde el inicio general correspondiente a la fecha	Detalles de la fecha	int(10)	Script de fechas	-	-
semestre	Número semestre correspondiente a la fecha	Detalles de la fecha	int(10)	Script de fechas	-	-
numeroSemestre	Número del semestre desde el inicio general correspondiente a la fecha	Detalles de la fecha	int(10)	Script de fechas	-	-

Tabla 45: Mapping de la tabla DimTiempo

Nombre de la tabla: DimFuenteInventario

Tipo de tabla: Tabla de Dimensión

Descripción: Dimensión de Fuente de inventario, que incluye todas las fuentes de inventario o proveedores de la base transaccional

Llave primaria: fuenteInventarioKey

Llave foránea: no hay

Política de nulidad: Todos los campos son requeridos.

Política de SCD:

- **SCD tipo 0:** fuenteInventarioID
- **SCD tipo 1:** nombre, descripción, país, ciudad

La tabla 46 se puede observar el mapping de DimFuenteInventario

Nombre de la columna	Descripción	Grupo de atributos	Tipo de dato	Sistema Origen	Tabla Origen	Columna Origen
fuenteInventarioKey	Llave primaria subrogada	Identificador	int	Llave Subrogada (autoincremental)	-	-

fuenteInventarioID	Llave primaria del sistema fuente	Identificador	varchar(255)	AlmacensElReyDB	inventory_source	source_code
nombre	Nombre de la fuente de inventario	Detalles de la fuente de inventario	varchar(255)	AlmacensElReyDB	inventory_source	name
descripcion	Descripción de la fuente de inventario	Detalles de la fuente de inventario	varchar(255)	AlmacensElReyDB	inventory_source	description
pais	País de la fuente de inventario	Detalles de la fuente de inventario	varchar(255)	AlmacensElReyDB	msp_tfa_country_codes	name
ciudad	Ciudad de la fuente de inventario	Detalles de la fuente de inventario	varchar(255)	AlmacensElReyDB	inventory_source	city

Tabla 46: Mapping de la tabla DimFuenteInventario

Nombre de la tabla: FactlessProductoFuenteInv

Tipo de tabla: Tabla de Dimensión

Descripción: Dimensión tipo Factless que incluye la relación de productos con fuentes de inventario/proveedores

Llave primaria: Llave compuesta por: fuenteInventarioKey, productoKey

Llave foránea:

- fuenteInventarioKey hace referencia a DimFuenteInventario
- productoKey hace referencia a DimProducto

Política de nulidad: Todos los campos son requeridos.

Política de SCD:

- **SCD tipo 1:** cantidadPorFuente, estado

La tabla 47 se puede observar el mapping de FactlessProductoFuenteInv

Nombre de la columna	Descripción	Grupo de atributos	Tipo de dato	Sistema Origen	Tabla Origen	Columna Origen
fuenteInventarioKey	Llave primaria subrogada	Identificador	int(10)	DWAlmacen	DimFuenteInventario	fuenteInventarioKey
productoKey	Llave primaria subrogada	Identificador	int(10)	DWAlmacen	DimProducto	productoKey

cantidadPorFuente	Cantidad en existencias del producto para una fuente de inventario	Detalles de producto fuente inventario	int(10)	AlmacenesElReyDB	inventory_source_item	quantity
estado	Estado del producto por sku	Detalles de producto fuente inventario	varchar(255)	AlmacenesElReyDB	inventory_source_item	status

Tabla 47: Mapping de la tabla FactlessProductoFuenteInv

Nombre de la tabla: FactVentas

Tipo de tabla: Tabla de Hechos

Descripción: Contiene los hechos para el proceso de ventas

Llave primaria: Llave compuesta por: productoKey, cuponKey, promocionKey, fechaKey

Llave foránea:

- productoKey, hace referencia a la DimProducto
- cuponKey, hace referencia a la DimCupon
- promocionKey, hace referencia a la DimPromocion
- fechaKey, hace referencia a la DimFecha

Política de nulidad: Todos los campos son requeridos.

La tabla 48 se puede observar el mapping de FactVentas

Nombre de la columna	Descripción	Tipo de dato	Sistema Origen	Tabla Origen	Columna Origen	Formula ETL
productoKey	Llave primaria a la dimensión Producto	int(10)	DWAlmacen	DimProducto	productoKey	-
cuponKey	Llave primaria a la dimensión Cupón	int(10)	DWAlmacen	DimCupon	cuponKey	-
promocionKey	Llave primaria a la dimensión Promoción	int(10)	DWAlmacen	DimPromocion	promocionKey	-
fechaKey	Llave primaria a la dimensión Fecha	int(10)	DWAlmacen	DimFecha	fechaKey	-

numeroOrden	Numero de orden	int(10)	AlmacenesElReyDB	sales_order_item	order_id	-
cantidadVendida	Cantidad de productos vendidos	int(10)	AlmacenesElReyDB	sales_order_item	qty_ordered	-
precioOriginal	Precio original del producto individual	decimal(12,4)	AlmacenesElReyDB	sales_order_item	original_price	-
precioVendido	Precio al que se vendió un producto individual	decimal(12,4)	AlmacenesElReyDB	sales_order_item	price	-
subtotal	precioVendido*cantidadVendida	decimal(12,4)	Derivado	-	-	precioVendido*cantidadVendida
totalImpuestos	Monto total de los impuestos en la venta	decimal(20,5)	AlmacenesElReyDB	sales_order_item	tax_amount	-
cantidadDescuentoxCupon	Cantidad a descontar por el uso de un cupón	decimal(20,5)	AlmacenesElReyDB	sales_order_item	discount_amount	-
totalVendidoAntesImpuesto	Monto total de la venta sin incluir impuestos	decimal(20,5)	Derivado	-	-	subtotal-cantidadDescuentoxCupon
totalVendidoConImpuesto	Monto total de la venta con el impuesto	decimal(20,5)	Derivado	-	-	(subtotal-cantidadDescuentoxCupon)+totalImpuesto
totalReembolso	Monto total de reembolsos en la venta	decimal(20,5)	AlmacenesElReyDB	sales_order_item	amount_refunded	-
cantidadReembolso	Cantidad de productos con reembolsos o devueltos en la venta	int(10)	AlmacenesElReyDB	sales_order_item	qty_refunded	-
montoFinal	Monto total final de la venta incluyendo impuestos,	decimal(20,5)	Derivado	-	-	totalVendidoConImpuesto-totalReembolso

	descuentos y reembolsos					

Tabla 48: Mapping de la tabla FactVentas

Nombre de la tabla: FactlessProductoPromocion

Tipo de tabla: Tabla de Hechos

Descripción: Contiene todos los productos que están en promoción en una fecha determinada

Llave primaria: Llave compuesta por: productoKey, promocionKey, fechaKey

Llave foránea:

- productoKey, hace referencia a la DimProducto
- promocionKey, hace referencia a la DimPromocion
- fechaKey, hace referencia a la DimFecha

Política de nulidad: Todos los campos son requeridos.

La tabla 49 se puede observar el mapping de FactlessProductoPromocion

Nombre de la columna	Descripción	Tipo de dato	Sistema Origen	Tabla Origen	Columna Origen	Formula ETL
productoKey	Llave primaria a la dimensión Producto	int(10)	DWAlmacen	DimProducto	productoKey	-
promocionKey	Llave primaria a la dimensión Promoción	int(10)	DWAlmacen	DimPromocion	promocionKey	-
fechaKey	Llave primaria a la dimensión Fecha	int(10)	DWAlmacen	DimFecha	fechaKey	-
promocionContador	Contador de relación producto-promoción	int(10)	Valor por defecto	-	-	Valor int: 1
precioPromocion	Precio del producto en promoción	decimal(20,6)	AlmacenesElReyDB	catalogrule_product_price	rule_price	-

Tabla 49: Mapping de la tabla FactlessProductoPromocion

Nombre de la tabla: InventarioTransaccionFact

Tipo de tabla: Tabla de Hechos

Descripción: Contiene los hechos para el proceso de inventario

Llave primaria: Llave compuesta por: productoKey, fechaKey, fuenteInventarioKey

Llave foránea:

- productoKey, hace referencia a la DimProducto
- fechaKey, hace referencia a la DimFecha
- fuenteInventarioKey hace referencia a DimFuenteInventario

Política de nulidad: Todos los campos son requeridos.

La tabla 50 se puede observar el mapping de InventarioTransaccionFact

Nombre de la columna	Descripción	Tipo de dato	Sistema Origen	Tabla Origen	Columna Origen	Formula ETL
productoKey	Llave primaria a la dimensión Producto	int(10)	DWAlmacen	DimProducto	productoKey	-
fechaKey	Llave primaria a la dimensión Fecha	int(10)	DWAlmacen	DimFecha	fechaKey	-
fuenteInventarioKey	Llave primaria a la dimensión FuenteInventario	int(10)	DWAlmacen	DimFuenteInventario	fuenteInventarioKey	-
tipoTransaccion	Tipo de la transacción a registrar en Inventario	varchar(255)				Se manejan los tipos de transacción: -Compra -Venta -Devolución
cantidad	Cantidad total de producto en stock	int(10)	Derivado	FactlessProductoFuenteInventario	cantidadPorFuente	-

umbralFueraStock	Cantidad con la cual el producto es considerado en escasez	int(10)	AlmacenesElReyDB	cataloginventory_stock_item	min_qty	-
pedidosPendientes	Si existen pedidos pendientes del producto	int(10)	AlmacenesElReyDB	cataloginventory_stock_item	backorders	-
cantidadMinimaVenta	Cantidad mínima que se debe vender del producto	int(10)	AlmacenesElReyDB	cataloginventory_stock_item	min_sale_qty	-
cantidadMaximaVenta	Cantidad máxima que se debe vender del producto	int(10)	AlmacenesElReyDB	cataloginventory_stock_item	max_sale_qty	-
cantidadVendida	Cantidad de unidades del producto que se han vendido a la fecha	int(10)	AlmacenesElReyDB	sales_order_item	qty_invoiced	-
cantidadComprada	Cantidad de unidades del producto que se han comprado en una fecha	int(10)	AlmacenesElReyDB	inventory_source_item	quantity	-
cantidadDevuelta	Cantidad de unidades del producto que se devolvieron	int(10)	AlmacenesElReyDB	sales_order_item	qty_refunded	-
costoPromedio	Costo promedio del producto a la fecha	decimal(20,5)	Derivado	DimProducto	costo	-
estaEnStock	Estado de disponibilidad del producto	varchar(255)	AlmacenesElReyDB	cataloginventory_stock_item	is_in_stock	-

Tabla 50: Mapping de la tabla InventarioTransaccionFact

c. Capitulo III: Estrategia de implementación de propuesta de solución

1. Estrategia de implementación

La solución que se ha diseñado ha sido pensando en empresas como Almacenes El Rey, empresas de pequeñas a mediado tamaño, cuyo modelo de negocio es B2C. Este modelo de negocio en el que las empresas venden a particulares, ya sea de manera física o digital.

Tomando en cuenta que la cantidad de datos que estas empresas maneja es cambiante dependiendo la temporada de venta, hacer uso de herramientas como AWS permite a las empresas adecuar su presupuesto de acuerdo al uso mensual que se le dan a los servicios.

En la solución planteada la empresa no debe hacer un esfuerzo extra para brindarnos su información de transacciones, haciendo uso de las herramientas descritas a continuación, accedemos directamente a la base de datos para llevar a cabo todo el análisis de datos y poder entregarle al usuario de negocio una visualización de los datos que le permita realizar una toma de decisión de manera oportuna.

En el caso del proceso ETL, si bien existe una amplia variedad de herramientas, Talend brinda una curva de aprendizaje menor, lo cual nos permite enfocarnos en el análisis y la solución a ejecutar.

Estrategias para implementar el proyecto.

Para que la solución pueda implementarse en un ambiente de producción es necesario instalar las herramientas requeridas para su correcto funcionamiento, así como también realizar las configuraciones necesarias para evitar errores.

- **Instalar y configurar Talend Open Studio**

Talend es una herramienta esencial para este proyecto debido a que en ella se desarrollan la mayor cantidad de procesos ETL necesarios para generar un modelo dimensional adecuado.

Instalar Java

Antes de instalar Talend Open Studio se debe verificar que se cuenta con el JDK de Java, el cual es un software que provee herramientas de desarrollo para la creación de programas en Java y como en Talend, se compilaran archivos fuentes basados en ese lenguaje, es necesario contar con él.

En la documentación oficial de Talend establece como requisito que el JDK[9] que se instale sea: el JDK 8 o el JDK 11, este último siento el más recomendado para ejecutar Talend Open Studio 8.0.

Si no se cuenta con el JDK, realizar los siguientes pasos:

- Descargar el instalador del JDK 11 en Zulu, distribución recomendada por la documentación de Talend: https://www.azul.com/downloads/?package=jdk
- Se procede a instalarlo en el sistema operativo.
- Luego se configuran las variables de entorno del sistema operativo para asegurarse que exista la ruta de instalación donde se encuentran los archivos de java. En la *figura 49* se puede observar la configuración de la variable de entorno Java Home

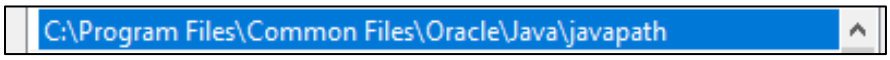


Figura 49: Variable de entorno Java Home

Instalación y configuración de Talend

Una vez instalado el JDK 11, la instalación de Talend no es complicada:

- Se debe ir a la página oficial de Talend en: https://www.talend.com/products/talend-open-studio/
- Para poder descargar el software de forma gratuita, Talend solicita a los usuarios registrarse colocando el correo electrónico entre otra información necesaria, luego de ello se procede a recibir un correo donde se proporcionan los diferentes links de descargas para los diferentes sistemas operativos.
- Se procede a descargar e instalar.
- Cuando se ingrese a Talend por primera vez solicitará la descarga de ciertas librerías requeridas, al aceptar, el software las instalará automáticamente. En la *figura 50* se puede observar la configuración de Talent

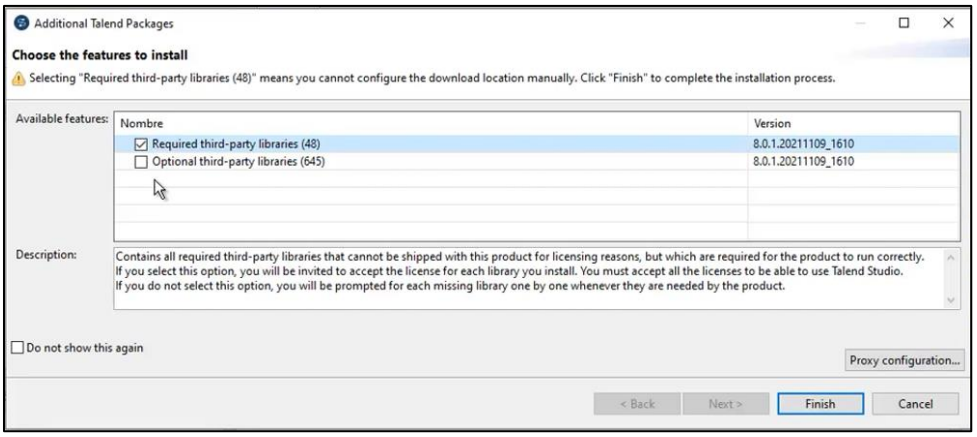


Figura 50: Configuración en Talend

[9] https://help.talend.com/r/en-US/8.0/installation-guide-linux/compatible-java-environments

- Talend también requerirá de los .jar para utilizar los componentes relacionado con MySQL y para tratar los componentes de Amazon (en este caso S3 y Redshift), de igual forma, Talend lo solicita y al aceptar los descargará e instalará automáticamente.

- **Instalar y configurar Power BI Desktop**

Power BI es la herramienta de visualización de datos que se ocupa en este proyecto, con el cual se crean los dashboards y reportes a presentar.

- Para descargar Power BI Desktop de forma gratuita solo es necesario dirigirse a la página oficial que se encuentra en: https://powerbi.microsoft.com/es-es/desktop/
- Luego proceder a instalarlo

Se recomienda que si se cuenta con algún correo electrónico con el cual se tenga acceso a Power BI Web se utilice para iniciar sesión en Power BI Desktop y así tener una mejor colaboración, organización y seguridad en los dashboards que se generen, además de que es más fácil compartirlos con los usuarios de negocio al momento de presentar los resultados obtenidos. En la *figura 51* pude ver la configuración inicial en Power BI.

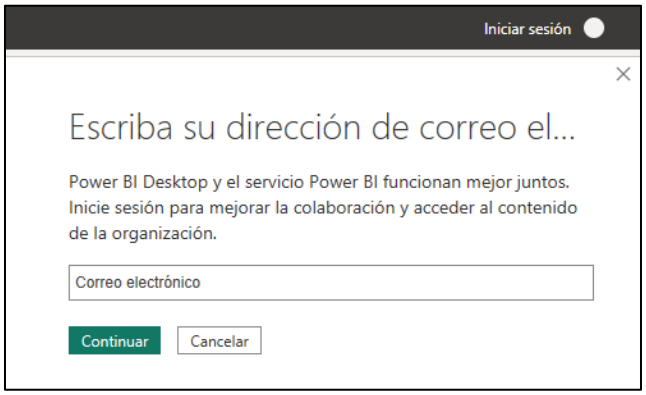


Figura 51: Configuración en Power BI

- **Configurar componentes en Amazon Web Services**

Para poder utilizar los componentes de Amazon Web Services es necesario crear una cuenta con esta plataforma.

  - **Amazon S3**

    Es el repositorio utilizado para almacenar los archivos en estado crudo, los archivos en procesos y los archivos finalizados para su consumo. Las configuraciones que se realizaron fueron las siguientes:

- Primero inicia sesión en la consola de AWS y se busca el componente de Amazon S3, en la *figura 52,* puede observar el símbolo del componente Amazon S3.

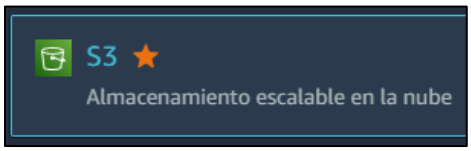


Figura 52: Seleccionar Amazon S3

- Al ingresar se crea un bucket o repositorio con el botón 

- Se le asigna un nombre de acuerdo a las reglas de nomenclatura[10] proporcionadas por AWS y se seleccionó una región, en cuyo caso fue: EE. UU. Este (Norte de Virginia) us-east-1
- Se desactivan las propiedades ACL para controlar el acceso del bucket y sus objetos mediante políticas.
- Se bloquea todo el acceso público para el bucket creado. Así como tampoco se lleva un control de versiones por lo cual esta opción esta desactivada.
- Para el cifrado de los objetos se selecciona Claves administras por Amazon S3 (SSE-S3) y se habilitado la creación de Clave de Buckets
- Luego de creado el bucket, se ingresa en el mismo y se selecciona el botón 

donde solo se ingresa el nombre de las carpetas y el tipo de clave de cifrado, en este caso: Claves administradas por Amazon S3 (SSE-S3)
- Se crean las carpetas hasta llegar a la estructura de carpetas que se muestra en la *Figura 40.*

- **Amazon Redshift**

Amazon Redshift a pesar de ser un producto de almacenamiento de información para este proyecto se utilizará como herramienta para manejar el acceso a los datos, además de guardar la estructura del modelo dimensional y su información lista para consumo. Las configuraciones que se realizaron fueron las siguientes:

- Se inicia sesión en la consola de AWS y se busca el componente Amazon Redshift, en la *figura 53,* puede observar el símbolo del componente Amazon Redshift.

Figura 53: Seleccionar Amazon Redshift

[10]Reglas de nomenclatura de buckets: https://docs.aws.amazon.com/es_es/AmazonS3/latest/userguide/bucketnamingrules.html

- Al ingresar se crea un clúster en 

- Se le asigna un nombre al clúster y se selecciona para que se quiere utilizar y de acuerdo a ello se elige la configuración que tendrá
- Se selecciona el tipo y la cantidad de nodos que tendrá el clúster. Esto dependerá de la capacidad que necesite la empresa.
- Luego se asigna el nombre del usuario de administrador y su contraseña.
- Después de crear el clúster se debe configurar la accesibilidad pública, para ello se debe ingresar al clúster y en la desglosar el botón "Acciones" y seleccionar la opción correspondiente. Esta acción se puede observar en la *figura 54.*

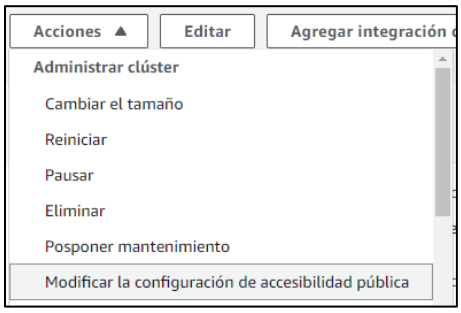


Figura 54: Redshift - Modificar la configuración de acceso público

- Marcar con un chequecito que se permita la accesibilidad pública. Esta acción se puede observar en la *figura 55.*

Figura 55: Redshift - Habilitar acceso público

- **Usuarios IAM**

Amazon IAM (Identity and Access Management) permite administrar usuarios y permisos de usuario en AWS. El usuario de IAM se crea para alguien que necesite acceder a la consola de AWS, o cuando se tenga una nueva aplicación que necesite hacer llamadas API a AWS. Esto es para agregar una capa adicional de seguridad a la cuenta de AWS. Presupuesto de implementación

- Iniciar sesión en la consola de AWS y seleccionar IAM en la lista de servicios, en la *figura 56,* puede observar el símbolo del componente Amazon Usuario IAM.

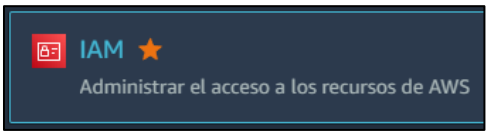


Figura 56: Seleccionar Amazon IAM

- Seleccione política y clic en el btn Crear política
- En servicios buscamos S3 y seleccionamos el nivel de acceso que queremos asignarle
- En la sección de recursos en el apartado de Bucket agregamos el ARN del bucket al que le queremos dar acceso
- Clic en revisar política y le asignamos un nombre a la política recién creada.
- Seleccione usuarios y clic en el btn añadir usuario
- Configurar el detalle del nuevo usuario, el nombre y la contraseña
- Agregar los permisos al nuevo usuario que estamos por crear, seleccionamos la política anteriormente creada y continuamos con el siguiente paso
- Luego dar clic en "Crear usuario"
- En este punto AWS genera una URL la cual se debe guardar En este punto AWS genera una URL la cual será usada por el usuario IAM recién creado para acceder a la consola de AWS

2. Presupuesto de implementación

Teniendo en cuenta los elementos esenciales que intervienen en la elaboración de la solución planteada, y pretendiendo que la misma se implemente y se lleve a un ambiente de producción, se presenta el a nivel general el siguiente presupuesto:

Recursos humanos

En la tabla 51 se presenta el presupuesto de recursos humanos.

Cargo	Funciones	Salario mensual	Cantidad	Costo
Ingeniero de datos	• Creación del modelo dimensional. • Diseño e implementación de las ETL en Talend. • Manejo de los servicios de AWS.	$5000.00	1	$5,000.00
Analista de datos	• Creación de reportes en Power BI. • Análisis de resultados.	$2500.00	1	$2,500.00
			Total mensual	**$7,500.00**

Tabla 51: Presupuesto Implementación – Recursos humanos

Hardware

En la tabla 52 se presenta el presupuesto de hadware.

Equipo	Cantidad	Costo	Subtotal
PC procesador I7 11th, 16 GB RAM, 2TB SSD	2	$2,100.00	$4,200.00
		Total	**$4,200.00**

Tabla 52: Presupuesto Implementación – Hardware

Las características del equipo son las mínimas, para tener un óptimo rendimiento en las operaciones realizadas.

Licencias

En la tabla 53 se presenta el presupuesto de licencias.

Licencia	Cantidad	Costo	Subtotal
Licencia premium para Power BI	20	$18.02	$360.04
		Total mensual	**$360.04**

Tabla 53: Presupuesto Implementación – Licencias

Servicios AWS

Según calculadora de servicios AWS

En la tabla 54 se presenta el presupuesto del Servicio AWS.

Servicio	Descripción	Costo	Subtotal
S3 Standard	• Almacenamiento y datos devueltos/escaneados para 10 TB. • 20000 solicitudes PUT, COPY, POST y LIST a S3 y Solicitudes GET, SELECT desde S3.	$360.04	$360.04
Instancia de Redshift	• 96 GiB de almacenamiento, disponible el 100% del tiempo.	$2,379.80	$2,379.80
		Total mensual	**$2,739.84**

Tabla 54: Presupuesto Implementación – Servicio AWS

Los servicios de AWS están sujetos a cambios en necesidad de mayor almacenamiento o accesibilidad de datos. Con los elementos anteriormente mencionados, se estima un costo inicial el cual se puede observar en la tabla 55:

Elemento	Costo
Recursos humanos	$7,500.00
Hardware	$4,200.00
Licencias	$360.04
Servicios AWS	$2,739.84
Total	**$14,799.88**

Tabla 55: Presupuesto Implementación – Total

3. Análisis de resultados

Reportes de resultados finales.

Para cada uno de las métricas requeridas por el negocio, se genera en Power BI un reporte, el cuál responde a lo solicitado. Cabe mencionar que este puede contener más elementos y configuraciones extras para enriquecer la visualización y por ende el análisis de los resultados finales. Los reportes se pueden filtrar por fechas.

1. Título: Informe del total de ventas por producto en una fecha determinada.

Descripción: En el informe se listan los productos con el acumulado de sus ventas por cada uno y se presenta la sumatoria de todas las ventas realizadas, ambos resultados están filtrados por las fechas que decida el usuario. Al lado derecho se muestra una gráfica que representa a los 10 productos con mayor monto de venta para las fechas indicadas. Este informe puede apreciarse en la *figura 57*.

Figura 57: Informe del total de ventas por producto en una fecha determinada.

2. Título: Informe de la cantidad de ventas por producto en una fecha determinada.

Descripción: En el informe se presenta en una tabla con cada uno de los productos y la cantidad de unidades vendidas para cierta fecha. Se visualiza también la totalidad de unidades vendidas para dicho lapso, además una gráfica que representa a los 10 productos con mayor cantidad de unidades vendidas en las fechas indicadas. Este informe puede apreciarse en la *figura 58*.

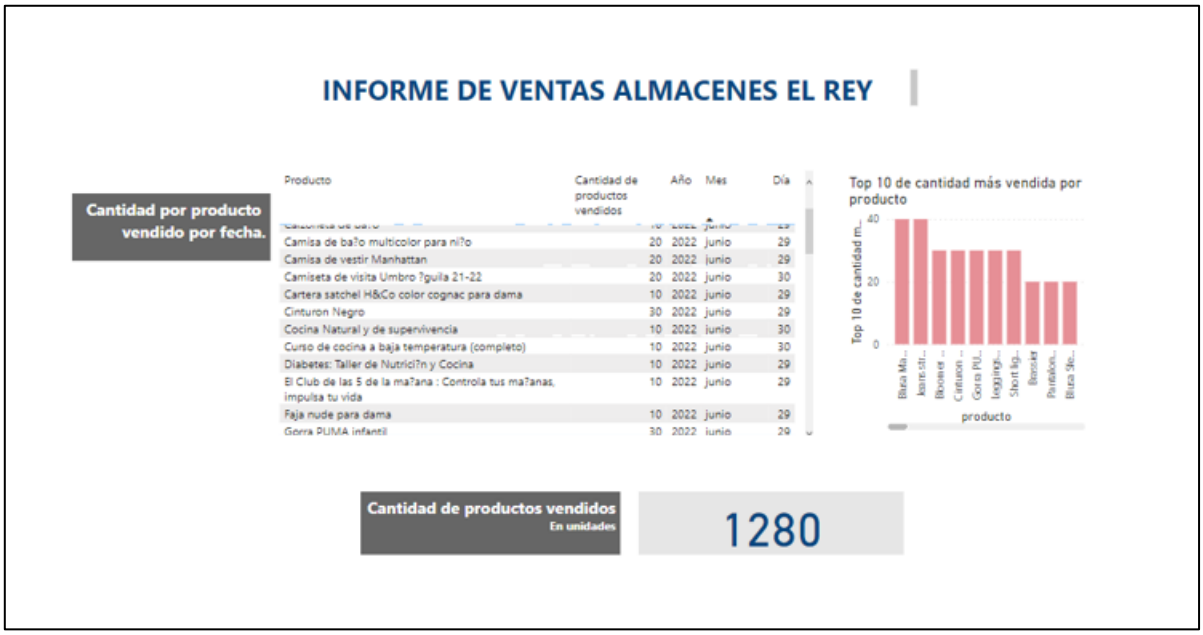


Figura 58: Informe de la cantidad de ventas por producto en una fecha determinada

3. Título: Informe del monto de descuentos por producto en una fecha determinada.

Descripción: En el informe se listan en una tabla los productos y el monto total de descuentos que se han aplicado en una fecha específica. De forma sumarizada se presenta también el monto total de descuentos realizados y el monto total vendido con descuentos incluídos para dicho período de tiempo, adicional una gráfica que contiene a las categorías de productos con mayor monto de descuentos aplicados. Este informe puede apreciarse en la *figura 59*.

Figura 59: Informe del monto de descuentos por producto en una fecha determinada.

4. Título: Informe del monto de impuestos cobrados por producto en una fecha determinada.

Descripción: En el informe se listan en una tabla los productos y el monto total de impuestos para cada producto en una fecha específica. En resumen se presenta también el monto total de impuestos cobrados y el monto total vendido con impuestos incluídos para dicho período de tiempo, y una gráfica que contiene a las categorías de productos con mayor monto de impuestos cobrados. Este informe puede apreciarse en la *figura 60*.

Figura 60: Informe del monto de impuestos cobrados por producto en una fecha determinada.

5. Título: Informe del monto de total ganado por producto en una fecha determinada.

Descripción: En el informe se listan en una tabla los productos y el monto total ganado para cada uno, esto se obtiene de las ventas, incluyendo impuestos, descuentos, devoluciones y costo promedio para una fecha específica. En resumen se presenta también el monto total de ganancias generales para dicho período de tiempo, y una gráfica que contiene a las categorías de productos con mayores ganancias obtenidas. Este informe puede apreciarse en la *figura 61*.

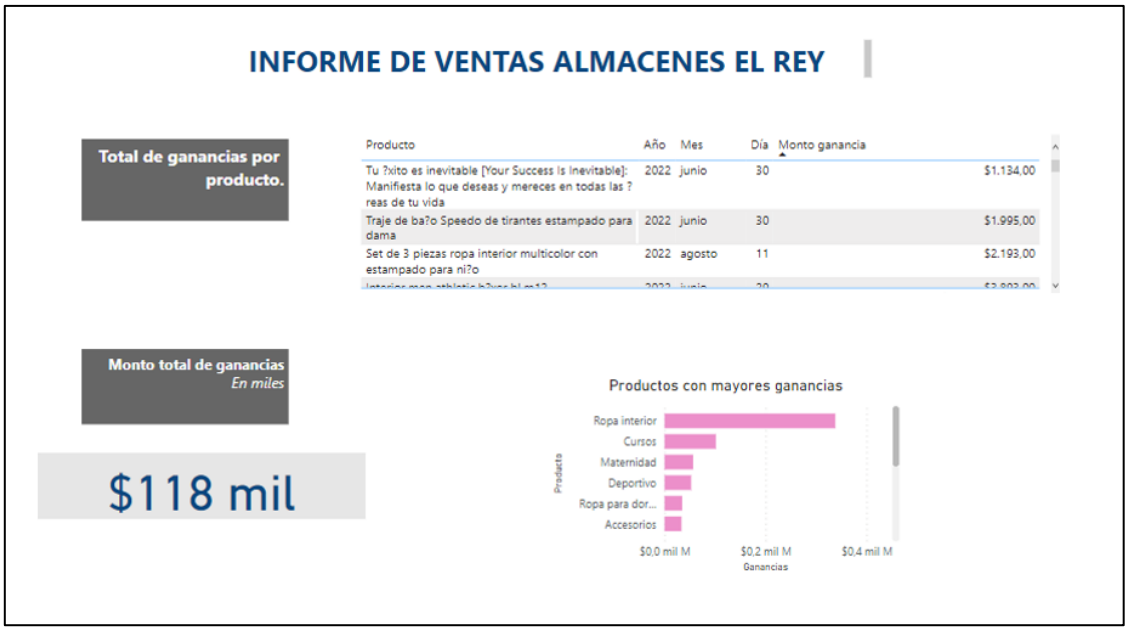


Figura 61: Informe del monto de total ganado por producto en una fecha determinada.

6. Título: Informe de existencias en inventario por producto en una fecha determinada.

Descripción: En el informe se listan en una tabla los productos y la cantidad de unidades disponibles en inventario, para una fecha determinada. En gráficas se presentan los productos con mayores y menores existencias en inventarios para la misma fecha especificada. Este informe puede apreciarse en la *figura 62*.

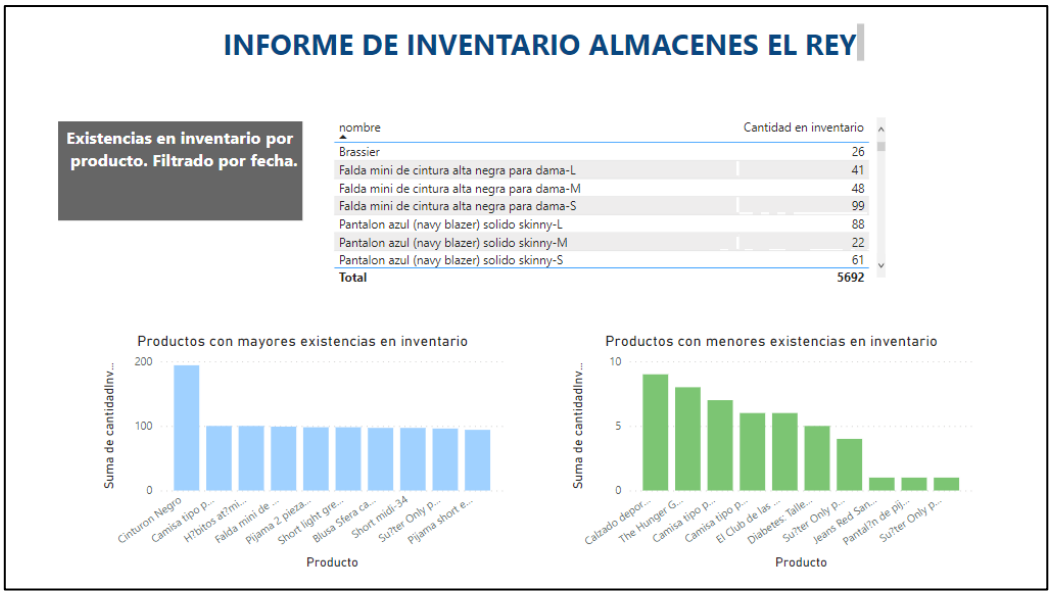


Figura 62: Informe de existencias en inventario por producto en una fecha determinada.

7. Título: Informe del costo promedio por producto, en una fecha determinada.

Descripción: En el informe se listan en una tabla los productos y el costo promedio para el tiempo indicado y dos gráficas que contienen los productos con mayor y menor costo promedio con lo que cuenta el almancen en inventario. Este informe puede apreciarse en la *figura 63*.

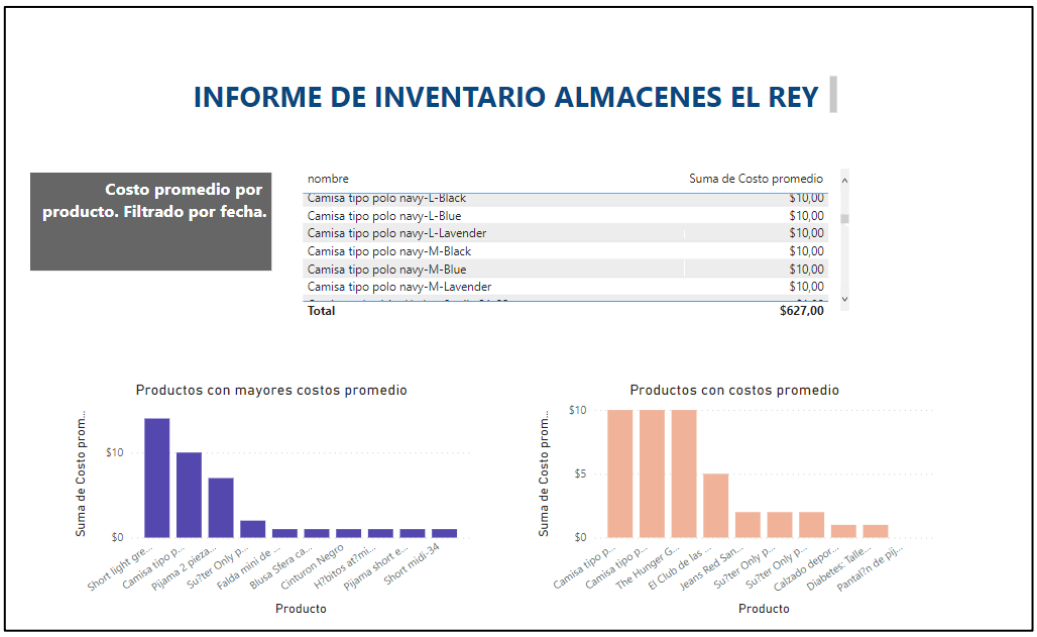


Figura 63: Informe del costo promedio por producto, en una fecha determinada.

5. Conclusiones y Recomendaciones

a. Conclusiones

Gracias al correcto estudio de los elementos brindados por la base de datos transaccional, la comprensión del aplicativo y los procesos de negocio venta e inventario se pudo identificar los componentes participantes que fueron insumos para el modelo dimensional creado, así como también, se logró extraer satisfactoriamente el conjunto de tablas con información relevante para la actividad, creando así el dataset y, posteriormente, elaborar un diccionario de datos y perfilado del mismo.

Luego de analizar los elementos transaccionales, lo cual permitió tener un amplio panorama de las reglas del negocio, se logró diseñar un modelo dimensional correspondiente a cada proceso de negocio requerido (ventas e inventario) que incluyera los elementos necesarios para dar respuesta a los requerimientos dados por el usuario.

La herramienta Talend Open Studio fue de gran utilidad para el desarrollo de este proyecto ya que nos permitió realizar los procesos de extracción, limpieza y carga de datos (ETL), logrando así, desarrollar con éxito la transformación de información desde su fuente de origen hasta su forma final.
Se motiva a seguir explorando la herramienta para encontrar formas más optimas y sofisticadas de implementar la solución, así como también para aprendizaje y conocimiento personal.

Los procesos ETL implementados en este proyecto generaban datos pertinentes, diversos y relevantes, por lo cual fue necesario su correcto almacenamiento y respaldo. Gracias a los servicios de S3 de Amazon Web Services se logró almacenar, organizar y estructurar exitosamente esta información en tres diferentes zonas: Raw, Stage y Presentation para mayor claridad y mejor organización.

Amazon Redshift fue otra herramienta de Amazon Web Services muy útil y necesaria para el desarrollo de esta solución, gracias a ella se logró implementar exitosamente el modelo dimensional diseñado, logrando crear las tablas de hechos y dimensiones correspondientes al modelo, estas tablas se cargaron con la información almacenada en Amazon S3, realizando esta acción dentro de Redshift de forma satisfactoria.

Redshift también fue útil para realizar exitosamente la comunicación entre Amazon S3 y las aplicaciones de Business Intelligence.

Finalmente, la creación de reportes, gráficos y tablas fueron exitosas y de utilidad para el usuario, gracias a la herramienta de Power BI, utilizada para diseñar dichos reportes. Como grupo se recomienda mucho el uso de Power BI para desarrollar y presentar análisis de datos debido a que es muy completa, es fácil de usar y cuenta con accesibilidad a los datos.

b. Recomendaciones

- El diseño del modelo dimensional y el nivel de detalle que se pretenda manejar debe ser determinado después de analizar la base de datos transaccional y la lógica del negocio, pues se debe gestionar la información al nivel de granularidad que los datos lo permitan.

- Antes de la adquisición de herramientas o licencias de las mismas, se debe estimar la cantidad de datos que se van a manejar, así se cuenta con la capacidad adecuada según sea el caso y se evitan inconvenientes como el bajo rendimiento en las operaciones, información inaccesible, o por otro lado, herramientas sobrecalificadas para el ejercicio, que llevan a gastos innecesarios.

- El diseño de la solución debe centrarse en los requerimientos que se quieren satisfacer y en la obtención de resultados o métricas que sean de valor para el usuario, aunque el diseño puede brindar datos suficientes para obtener otras métricas en un futuro, deben cumplirse primordialmente las que se han establecido y que son el motivo de la implementación del modelo dimensional.

- Los datos o información que maneja una entidad son de gran valor y por lo tanto deben protegerse y resguardarse. Se recomienda gestionar usuarios, roles y permisos para el debido acceso a la información, y un mejor control de la misma.

6. Bibliografía

- AWS Pricing Calculator. (s. f.). Calculadora de precios de AWS. https://calculator.aws
- Introducción a Amazon Redshift. Documentación oficial Amazon Redshift. https://docs.aws.amazon.com/es_es/redshift/latest/gsg/getting-started.html
- Kimball, R., & Ross, M. (2013). The data warehouse toolkit (3rd ed.). Wiley.
- Paula Nicole Roldán, 31 de julio, 2017. Negocio. Economipedia.com
- Peñafiel, G. E. S., Yánez, V. M. Z., Guamán, K. P. M., & Padilla, L. M. T. (2019). Análisis de metodologías para desarrollar Data Warehouse aplicado a la toma de decisiones. Ciencia digital, 3(3.4.), 397-418.
- René Cuchillas, Oscar Hernández, Yuri Mejía, Héctor Silva (2010). Tesis DESARROLLO DE UN "DATA WAREHOUSE" PARA EL PROCESO DE DENUNCIAS DE LA DEFENSORÍA DEL CONSUMIDOR
- Rivadera, G. R. (2010). La metodología de Kimball para el diseño de almacenes de datos (Data warehouses).
- Rosado Gómez, A. A., & Rico Bautista, D. W. (2010). Inteligencia de negocios: Estado del arte. *Scientia Et Technica*, *1*(44), 321–326. https://doi.org/10.22517/23447214.1803
- ¿Qué es Amazon S3?. Documentación oficial de Amazon S3. https://docs.aws.amazon.com/es_es/AmazonS3/latest/userguide/Welcome.html
- ¿Qué es IAM?. Documentación oficial de Amazon IAM. https://docs.aws.amazon.com/es_es/IAM/latest/UserGuide/introduction.html
- ¿Qué es Power BI?. Documentación oficial de Power BI. https://learn.microsoft.com/es-es/power-bi/fundamentals/power-bi-overview
- What is Talend Studio?. Documentación official de Talend Open Studio. https://help.talend.com/r/en-US/8.0/studio-user-guide/what-is-talend-studio

7. Glosario

- **Data warehouse:** En el contexto de la informática, un almacén de datos o repositorio de datos es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza. Se usa para realizar informes (reports) y análisis de datos y se considera un componente fundamental de la inteligencia empresarial. Se trata, sobre todo, de un expediente completo de una organización, más allá de la información transaccional y operacional, almacenado en una base de datos diseñada para favorecer el análisis y la divulgación eficiente de datos (especialmente OLAP, *procesamiento analítico en línea*). El almacenamiento de los datos no debe usarse con datos de uso actual. Los almacenes de datos contienen a menudo grandes cantidades de información que se subdividen a veces en unidades lógicas más pequeñas dependiendo del subsistema de la entidad del que procedan o para el que sea necesario.

- **Datos:** Los datos requieren una interpretación para convertirse en información. Para traducir datos a información, debe haberse considerado varios factores conocidos. Los factores involucrados están determinados por el creador de los datos y la información deseada. El término metadatos se usa para hacer referencia a los datos sobre los datos. Los metadatos pueden estar implícitos, especificados o dados. Los datos relacionados con eventos o procesos físicos también tendrán un componente temporal.

- **Análisis de datos:** El análisis de datos es un proceso que consiste en inspeccionar, limpiar y transformar datos con el objetivo de resaltar información útil, para sugerir conclusiones y apoyo en la toma de decisiones. El análisis de datos tiene múltiples facetas y enfoques, que abarca diversas técnicas en una variedad de nombres, en diferentes negocios, la ciencia, y los dominios de las ciencias sociales. Los datos se coleccionan y analizan para indagar en cuestiones, probar conjeturas o refutar teorías

- **Sistema transaccional:** Se entiende por sistema de información transaccional aquel diseñado para recolectar, modificar, almacenar y recuperar información generada por las transacciones en una organización

- **Enfoque de sistemas:** Su propósito es estudiar los principios aplicables a los sistemas en cualquier nivel en todos los campos de la investigación.1 Un sistema se define como una entidad con límites y con partes interrelacionadas e interdependientes cuya suma es mayor a la suma de sus partes. El cambio de una parte del sistema afecta a las demás y, con esto, al sistema completo, generando patrones predecibles

de comportamiento. El crecimiento positivo y la adaptación de un sistema dependen de cómo se ajuste este a su entorno. Además, a menudo los sistemas existen para cumplir un propósito común (una función) que también contribuye al mantenimiento del sistema y a evitar sus fallos.

El objetivo de la teoría de sistemas es el descubrimiento sistemático de las dinámicas, restricciones y condiciones de un sistema, así como de principios (propósitos, medidas, métodos, herramientas, etc.) que puedan ser discernidos y aplicados a los sistemas en cualquier nivel de anidación y en cualquier campo, con el objetivo de lograr una equifinalidad optimizada

- **Base de datos:** es una herramienta para recopilar y organizar información. Las bases de datos pueden almacenar información sobre personas, productos, pedidos u otras cosas. Muchas bases de datos comienzan como una lista en una hoja de cálculo o en un programa de procesamiento de texto.

- **Base de datos relacional:** «Una base de datos relacional es un tipo de base de datos que cumple con el modelo relacional». Así, según esta definición de base de datos relacional, se trata de una base de datos que almacena y da acceso a puntos de datos relacionados entre sí. El modelo relacional es una forma intuitiva y directa de representar datos sin necesidad de jerarquizarlos.

8. Anexos

a. Cronograma de actividades

Para el proyecto desarrollado en la especialización puede observar las actividades realizadas en la *figura 65* y en la tabla 56:

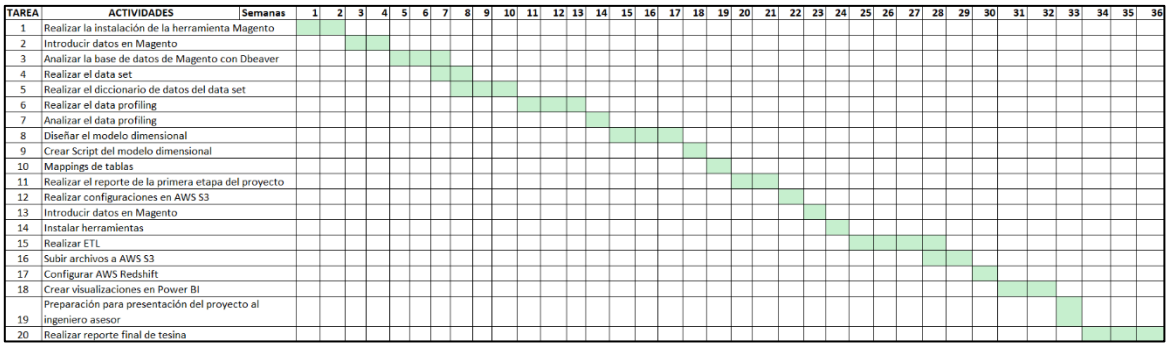


Figura 64: Cronograma de actividades en semanas

Tarea	Descripción	Duración
1	Realizar la instalación de la herramienta Magento	2 semanas
2	Introducir datos en Magento	2 semanas
3	Analizar la base de datos de Magento con DBeaver	3 semanas
4	Realizar el data set	2 semanas
5	Realizar el diccionario de datos del data set	3 semanas
6	Realizar el data profiling	3 semanas
7	Analizar el data profiling	1 semana
8	Diseñar el modelo dimensional	3 semanas
9	Crear Script del modelo dimensional	1 semana
10	Mappings de tablas	1 semana
11	Realizar el reporte de la primera etapa del proyecto	2 semanas
12	Realizar configuraciones en AWS S3	1 semana
13	Introducir datos en Magento	1 semana
14	Instalar herramientas	1 semana
15	Realizar ETL	4 semanas
16	Subir archivos a AWS S3	2 semanas
17	Configurar AWS Redshift	1 semana
18	Crear visualizaciones en Power BI	2 semanas
19	Preparación para presentación del proyecto al ingeniero asesor	1 semana
20	Realizar reporte final de tesina	2 semanas

Tabla 56: Cronograma de actividades en semanas

b. Presupuesto

El costo de la realización de la solución planteada se divide en los siguientes elementos:

El salario propuesto a continuación es para un analista de datos junior, como sería nuestro caso, además de que se buscaron publicaciones para obtener un rango salarial (Anexo d).

Recursos humanos

En la tabla 57 se presenta el presupuesto de recursos humanos.

Cargo	Funciones	Salario mensual	Cantidad	Costo
Ingeniero/Analista de datos	• Creación del modelo dimensional. • Diseño e implementación de las ETL en Talend. • Manejo de los servicios de AWS. • Creación de reportes en Power BI. • Análisis de resultados.	$800.00	3	$2,400.00
			Total mensual	**$2,400.00**

Tabla 57: Presupuesto desarrollo – Recursos humanos

Hardware

En la tabla 58 se presenta el presupuesto de hadware.

Equipo	Cantidad	Costo	Subtotal
PC procesador I5 10th, 16 GB RAM, 500 GB SSD	3	$1,000.00	$3,000.00
		Total	**$3,000.00**

Tabla 58: Presupuesto desarrollo – Hardware

Licencias

Se utilizaron licencias gratuitas para las herramientas utilizadas (Talend Open Studio y Microsoft Power BI).

Servicios AWS

Se utilizó la capa gratuita que ofrece AWS, pero cabe mencionar que se presentaron múltiples restricciones, por lo que se recomienda que para un mejor uso de dichos servicios se emplee la siguiente configuración (ver tabla 59):

Servicio	Descripción	Costo	Subtotal
S3 Standard	• Almacenamiento y datos devueltos/escaneados para 500 GB. • 10000 solicitudes PUT, COPY, POST y LIST a S3 y Solicitudes GET, SELECT desde S3.	$12.90	$12.90
Instancia de Redshift	• 32 GiB de almacenamiento, disponible el 100% del tiempo.	$792.78	$792.78
		Total mensual	**$805.68**

Tabla 59: Presupuesto desarrollo – Servicio AWS

Con los elementos anteriormente mencionados, se estima un costo inicial (primer mes) el cual se puede observar en la tabla 60:

Elemento	Costo
Recursos humanos	$2,400.00
Hardware	$3,000.00
Servicios AWS	$805.68
Total	**$6,205.68**

Tabla 60: Presupuesto desarrollo – Total

c. Script del modelo dimensional para Redshift

```
create table DimCategoria(
categoriaKey INTEGER NOT NULL PRIMARY KEY DISTKEY,
categoriaId  INTEGER NOT NULL,
nombre       VARCHAR(255) NOT NULL,
tipoProducto VARCHAR(255) NOT NULL,
departamento VARCHAR(255) NOT NULL,
estado       VARCHAR(50) NOT NULL
);
```

```
create table DimProducto(
productoKey INTEGER NOT NULL PRIMARY KEY DISTKEY,
productoId INTEGER NOT NULL,
categoriaKey INTEGER NOT NULL,
nombre VARCHAR(MAX) NOT NULL,
descripcion VARCHAR(MAX) NOT NULL,
talla VARCHAR(255) NOT NULL,
color VARCHAR(255) NOT NULL,
material VARCHAR(255) NOT NULL,
clima VARCHAR(255) NOT NULL,
sku VARCHAR(64) NOT NULL,
precio DECIMAL(20,6) NOT NULL,
costo DECIMAL(20,6) NOT NULL,
estado VARCHAR(50) NOT NULL,
scd_start VARCHAR(10) NOT NULL,
scd_end VARCHAR(10),
scd_active VARCHAR(7) NOT NULL

FOREIGN KEY (categoriaKey) REFERENCES DimCategoria(categoriaKey)
);


create table DimPromocion(
promocionKey INTEGER NOT NULL PRIMARY KEY DISTKEY,
promocionID INTEGER NOT NULL,
nombre VARCHAR(255) NOT NULL,
descripcion TEXT NOT NULL,
fechaInicio VARCHAR(10) NOT NULL,
fechaFin VARCHAR(10) NOT NULL,
cantidadDescuento DECIMAL(10,2) NOT NULL,
estaActiva VARCHAR(50) NOT NULL
);


create table DimCupon(
cuponKey INTEGER NOT NULL PRIMARY KEY DISTKEY,
cuponID INTEGER NOT NULL,
nombre VARCHAR(255) NOT NULL,
descripcion TEXT NOT NULL,
codigo VARCHAR(255) NOT NULL,
cantidadCupones INTEGER NOT NULL,
cantidadUsados INTEGER NOT NULL,
```

```
fechaInicio VARCHAR(10) NOT NULL,
fechaFin VARCHAR(10) NOT NULL,
estado VARCHAR(50) NOT NULL
);


create table dimTiempo (
fechaKey INTEGER NOT NULL PRIMARY KEY DISTKEY,
fechaCompleta VARCHAR (25) NOT NULL,
diaDeSemana INTEGER NOT NULL,
numeroDiaDelMes INTEGER NOT NULL,
numeroDiaDelAnio INTEGER NOT NULL,
nombreDia VARCHAR (25) NOT NULL,
diaLaboralNoLaboral VARCHAR (25) NOT NULL,
numeroSemanaAlAnio INTEGER NOT NULL,
numeroDeSemana INTEGER NOT NULL,
fechaInicioDeLaSemana VARCHAR (25) NOT NULL,
fechaInicioDeLaSemanaKey INTEGER NOT NULL,
mes INTEGER NOT NULL,
numeroMes INTEGER NOT NULL,
nombreMes VARCHAR (25) NOT NULL,
anio INTEGER NOT NULL,
trimestre VARCHAR (25) NOT NULL,
numeroTrimestre INTEGER NOT NULL,
semestre VARCHAR (25) NOT NULL,
numeroSemestre INTEGER NOT NULL)

sortkey (fechaKey, mes);


create table factVentas(
productoKey INTEGER NOT NULL PRIMARY KEY,
cuponKey INTEGER NOT NULL PRIMARY KEY,
promocionKey INTEGER NOT NULL PRIMARY KEY,
fechaKey INTEGER NOT NULL PRIMARY KEY,
numeroOrden INTEGER NOT NULL DISTKEY,
cantidadVendida INTEGER NOT NULL,
precioOriginal DECIMAL(5,2) NOT NULL,
precioVendido DECIMAL(5,2) NOT NULL,
porcentajeImpuesto DECIMAL(5,2) NOT NULL,
cantidadDescuentoxCupon DECIMAL(5,2) NOT NULL,
totalReembolso DECIMAL(5,2) NOT NULL,
cantidadReembolso INTEGER NOT NULL
```

```
FOREIGN KEY (productoKey) REFERENCES DimProducto(productoKey),
FOREIGN KEY (cuponKey) REFERENCES DimCupon(cuponKey),
FOREIGN KEY (promocionKey) REFERENCES DimPromocion(promocionKey),
FOREIGN KEY (fechaKey) REFERENCES DimFecha(fechaKey))

sortkey (productoKey,cuponKey,promocionKey,fechaKey);


create table factlessProductoPromocion(
productoKey INTEGER NOT NULL PRIMARY KEY,
promocionKey INTEGER NOT NULL PRIMARY KEY,
fechaKey INTEGER NOT NULL PRIMARY KEY,
promocionContador INTEGER NOT NULL,
precioPromocion DECIMAL(5,2) NOT NULL

FOREIGN KEY (productoKey) REFERENCES DimProducto(productoKey),
FOREIGN KEY (promocionKey) REFERENCES DimPromocion(promocionKey),
FOREIGN KEY (fechaKey) REFERENCES DimFecha(fechaKey))

sortkey (productoKey, promocionKey,fechaKey);


create table DimFuenteInventario(
fuenteInventarioKey INTEGER NOT NULL PRIMARY KEY,
fuenteInventarioID VARCHAR(255) NOT NULL,
nombre VARCHAR(255) NOT NULL,
descripcion VARCHAR(255) NOT NULL,
pais VARCHAR(255) NOT NULL,
ciudad VARCHAR(255) NOT NULL
);


create table FactlessProductoFuenteInv (
fuenteInventarioKey INTEGER NOT NULL PRIMARY KEY,
productoKey INTEGER NOT NULL PRIMARY KEY,
cantidadPorFuente INTEGER NOT NULL,
estado VARCHAR(255) NOT NULL

FOREIGN KEY (productoKey) REFERENCES DimProducto(productoKey),
FOREIGN KEY (fuenteInventarioKey) REFERENCES DimFuenteInventario (fuenteInventarioKey),

sortkey (fuenteInventarioKey, productoKey);
```

```
create table InventarioTransaccionFact(
productoKey INTEGER NOT NULL PRIMARY KEY,
fechaKey INTEGER NOT NULL PRIMARY KEY,
fuenteInventarioKey INTEGER NOT NULL PRIMARY KEY,
tipoTransaccion VARCHAR(255) NOT NULL,
cantidad INTEGER NOT NULL,
umbralFueraStock INTEGER NOT NULL,
pedidosPendientes INTEGER NOT NULL,
cantidadMinimaVenta INTEGER NOT NULL,
cantidadMaximaVenta INTEGER NOT NULL,
cantidadVendida INTEGER NOT NULL,
cantidadComprada INTEGER NOT NULL,
cantidadDevuelta INTEGER NOT NULL,
costoPromedio DECIMAL(5,2) NOT NULL,
estaEnStock VARCHAR(255) NOT NULL

FOREIGN KEY (productoKey) REFERENCES DimProducto(productoKey),
FOREIGN KEY (fuenteInventarioKey) REFERENCES DimFuenteInventario (fuenteInventarioKey),
FOREIGN KEY (fechaKey) REFERENCES DimFecha(fechaKey))

sortkey (productoKey, fuenteInventarioKey, fechaKey);
```

d. Publicación de oferta analista de datos

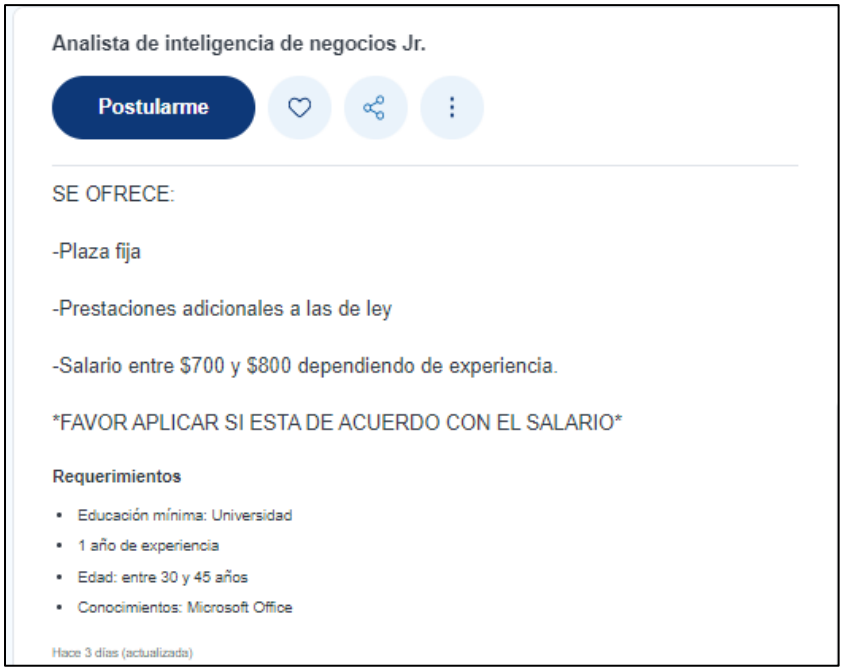


Figura 65: Oferta laboral