

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



**CURSO DE ESPECIALIZACIÓN EN INGENIERÍA DE DATOS.**  
**ANÁLISIS DE INFORMACIÓN SOBRE VENTAS DE LA EMPRESA EL ÁNGEL**  
**GESTIONADA A TRAVÉS DEL SISTEMA MAGENTO COMMERCE**

**PRESENTADO POR:**

**ARÉVALO BELTRÁN, KEVIN ELISEO**

**DÍAZ SORTO, ELÍAS ARTURO**

**VIDES ROMERO, JESSICA ESMERALDA**

**PARA OPTAR AL TÍTULO DE:**

**INGENIERO DE SISTEMAS INFORMÁTICOS**

**CIUDAD UNIVERSITARIA, DICIEMBRE DE 2023**

**UNIVERSIDAD DE EL SALVADOR**

**RECTOR:**

**MSC. JUAN ROSA QUINTANILLA**

**SECRETARIO GENERAL:**

**LIC. PEDRO ROSALÍO ESCOBAR CASTANEDA**

**FACULTAD DE INGENIERIA Y ARQUITECTURA**

**DECANO:**

**ING. LUIS SALVADOR BARRERA MANCÍA**

**SECRETARIO:**

**ARQ. RAUL ALEXANDER FABIAN ORELLANA**

**ESCUELA DE INGENIERÍA DE SISTEMAS INFORMÁTICOS**

**DIRECTOR:**

**ING. CÉSAR AUGUSTO GONZÁLEZ**

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

Curso de Especialización previo a la opción al Grado de:  
**INGENIERO (A) DE SISTEMAS INFORMÁTICOS**

Título:

**ANÁLIS DE INFORMACIÓN SOBRE VENTAS DE LA  
EMPRESA EL ÁNGEL GESTIONADA A TRAVÉS DEL  
SISTEMA MAGENTO COMMERCE**

Presentado por:

**ARÉVALO BELTRÁN, KEVIN ELISEO**  
**DÍAZ SORTO, ELÍAS ARTURO**  
**VIDES ROMERO, JESSICA ESMERALDA**

Curso de Especialización Aprobado por:

Docente Asesor:

**ING. RENÉ FABRICIO QUINTANILLA GÓMEZ**

SAN SALVADOR, DICIEMBRE 2023

Curso de Especialización Aprobado por:

Docente Asesor:

**ING. RENÉ FABRICIO QUINTANILLA GÓMEZ**

## **AGRADECIMIENTOS**

Expreso mi profundo agradecimiento a todas las personas que contribuyeron significativamente en la culminación de mis estudios. En primer lugar, agradezco a mi amada madre que me apoya y constantemente me aconseja a no rendirme y siempre seguir adelante; a mi papá que Dios lo tenga en su santa gloria, por su eterno respaldo y confianza en mí; a mis hermanas, en especial a mi hermana Elisa cuya ayuda y apoyo fueron fundamentales en gran parte de la carrera.

Agradezco también a mi familia que siempre está al pendiente; a mis amigos que me han acompañado durante el recorrido de esta carrera. Finalmente, un profundo agradecimiento a mi equipo de trabajo, quiénes son un factor importante en la finalización de este importante logro personal. Gracias a todos.

**Elías Arturo Díaz Sorto**

Quiero expresar mis agradecimientos a las personas que me apoyaron a culminar esta etapa de mi vida, misma en la que estuvieron varias personas que directa e indirectamente estuvieron presentes en todo el tiempo que duro el proceso para cumplir la meta y por eso quiero agradecer a:

### **Mis padres**

“Agradecer a mis padres por siempre brindarme su apoyo para cumplir mi meta de terminar esta etapa académica, tengo que agradecerles el estar ahí para brindarme palabras de aliento, quiero agradecerles por motivarme a seguir, por haberme puesto en este camino y desde el primer día estar presentes para todos los retos que esto significo, realmente hay cosas que se quedan cortas explicarlas con palabras y esta es una de ellas, les agradezco de corazón el cariño y paciencia brindada y como ellos decían: si te vas a rendir que sea después de haberlo dado todo”.

### **Compañeros**

“Agradecer igualmente a todos mis compañeros que estuvieron presentes en todo el camino hacia esta meta, agradecer por las noches de desvelo que pasamos, por el apoyo y motivación brindada. Agradecer especialmente a mis compañeros de especialización con quienes inicie este último tramo del camino, agradecerles por ayudar a completar este proceso, por su paciencia, por siempre estar dispuestos a ayudarme y brindarme conocimiento. Hay un dicho muy popular el cual cobra sentido para este agradecimiento el cual dice: Reamente hay pocas cosas que puedas lograr tu solo, pero si tienes buenos compañeros apoyándote, puedes lograr cualquier cosa”.

### **Docentes**

“También cabe agradecer a los docentes que han sido parte de este camino, agradecer el conocimiento brindado para hoy estar acá finalizando esta meta”.

### **Dios**

“Durante toda mi vida mis padres me han inculcado el dar gracias a Dios por cada día y cada meta lograda, es por eso que quiero dedicar unas palabras para agradecer infinitamente a Dios por permitir empezar este proceso y por poder culminarlo”.

**Kevin Eliseo Arévalo Beltrán**

Agradecimientos a:

### **Dios todo poderoso**

Agradezco a Dios por concederme la capacidad de aprender y darme la perseverancia para poder superar los desafíos que surgieron a lo largo de la carrera.

### **Mis padres**

Juan y Flora que han formado una parte super importante en mi vida académica, mi mamá por siempre darme palabras de motivación cuando las situaciones se ponían difíciles y a enseñarme a no rendirme.

### **Hermanos**

Mis hermanos que han sido parte importante en el inicio de la carrera, mi hermana Rebeca que motivó a que iniciara la carrera universitaria y mi hermano Ismael que contribuyó al inicio de los estudios.

### **Amigos**

A todas las amistades que formé en este centro de estudios, en especial a Elsy y Arely que con su conocimiento y motivación superé diferentes circunstancias; a mi amigo Christian por su asesoramiento.

**Jessica Esmeralda Vides Romero**

## CONTENIDO

INTRODUCCIÓN .....	2
MARCO TEÓRICO.....	3
CAPÍTULO I: ESPECIFICACIÓN DEL PROYECTO .....	7
a. Situación actual.....	7
i. Antecedentes.....	7
ii. Descripción del problema .....	9
iii. Planteamiento del problema.....	10
b. Objetivos.....	11
c. Alcances.....	12
d. Justificación .....	13
e. Cronograma de actividades.....	14
f. Presupuesto .....	15
CAPÍTULO II: ANÁLISIS Y DISEÑO DE LA PROPUESTA DE SOLUCIÓN .....	17
a. Metodología de trabajo .....	17
b. Descripción de la propuesta de solución.....	26
c. Descripción de la tecnología a utilizar .....	29
d. Diagrama arquitectónico de la solución.....	33
e. Descripción de cada componente de la solución .....	34
CAPÍTULO III: ESTRATEGIA DE IMPLEMENTACIÓN DE PROPUESTA DE SOLUCIÓN .....	39
a. Estrategia de implementación .....	39
b. Presupuesto de implementación.....	57
c. Análisis de resultados.....	58
CONCLUSIONES Y RECOMENDACIONES.....	70
Conclusiones .....	70
Recomendaciones.....	70
BIBLIOGRAFÍA .....	72

## INTRODUCCIÓN

En el dinámico entorno empresarial actual, la capacidad de comprender y aprovechar la información generada por las actividades de ventas es fundamental para el éxito de cualquier organización. En este contexto, el presente documento tiene como fin realizar un estudio de análisis de ventas a través de la empresa “Importaciones el Ángel”, que se dedica a la importación y comercialización de producto decorativo para el hogar a través de tienda en línea y con presencia únicamente en El Salvador, aplicando la ingeniería de dato sobre las ventas de la empresa, cuyas operaciones son gestionadas a través del sistema Magento Commerce.

La ingeniería de datos se encarga de diseñar, desarrollar y mantener los sistemas y procesos que permiten la adquisición, almacenamiento, procesamiento y análisis de datos de manera eficiente y efectiva. Esta juega un papel fundamental en la gestión de grandes volúmenes de información, asegurando que los datos estén disponibles, sean confiables y estén listos para ser utilizados en diversas aplicaciones y análisis.

Analizar la información de las ventas es una tarea clave para evaluar el desempeño comercial, identificar patrones y tendencias, y tomar decisiones estratégicas fundamentadas. En este sentido, se busca examinar los datos de ventas, para obtener una visión general de su rendimiento y entender los factores que influyen en su éxito de mercado.

Pueden existir ciertos factores que impidan el correcto análisis de ventas de una empresa, entre los cuales se pueden encontrar la coherencia y precisión en los datos de las ventas, esto puede afectar en la toma de decisiones poniendo en riesgo la salud financiera y el crecimiento sostenible del negocio, otros factores que pueden obstaculizar el correcto análisis de ventas incluyen falta de integración de sistemas, errores humanos en la entrada de datos, inconsistencias en los registros y falta de análisis predictivo.

Para superar estos desafíos, la empresa puede impulsar la implementación de prácticas sólidas de gestión de datos, uso de herramientas que ayuden a la extracción, transformación, carga y visualización de los datos de entrada y salida, implementar integración de sistemas internos, capacitar al personal sobre el uso de las herramientas y la interpretación de los datos para garantizar una comprensión eficiente de la información y mantenerse actualizado sobre las tendencias de mercado como lo son las operaciones en línea. Un análisis de ventas sólido y confiable es esencial para tomar decisiones informadas que impulsen éxito a largo plazo.

## MARCO TEÓRICO

### **Ingeniería de datos**

La ingeniería de datos es una disciplina dentro del campo de la ciencia de datos y la ingeniería de software que se centra en el diseño, desarrollo y gestión de sistemas y arquitecturas para la recopilación, almacenamiento, procesamiento y análisis de datos. El objetivo principal es garantizar que los datos estén disponibles, accesibles, confiables y listos para ser utilizados por los profesionales de datos, científicos de datos y otros usuarios finales.

Las actividades específicas en la ingeniería de datos pueden incluir la extracción de datos de diversas fuentes, la transformación de datos para asegurar su calidad y coherencia, la carga de datos en almacenes de datos o bases de datos, la implementación de pipelines de datos para el flujo continuo de información, y la creación de modelos de datos eficientes y escalables. Además, la ingeniería de datos a menudo implica el uso de tecnologías y herramientas especializadas, como bases de datos distribuidas, sistemas de procesamiento de datos en tiempo real, frameworks de Big Data y herramientas de orquestación de flujos de trabajo.

La ingeniería de datos puede ser utilizada en un escenario de la vida real, para realizar análisis de información sobre ventas, es fundamental para comprender el rendimiento de un negocio, identificar tendencias, tomar decisiones operativas y optimizar estrategias de comercio. Para lograr esto se debe seguir una base teórica que proporciona los puntos claves y consideraciones que se deben de tomar en cuenta para un correcto análisis del modelo de estudio.

En la aplicación de la ingeniería de datos se ven involucrados diferentes conceptos, algunos de los más fundamentales se describen a continuación.

### **DataWarehouse**

Un almacén de datos, o DataWarehouse en inglés, es un sistema de almacenamiento de datos diseñado para consolidar y almacenar grandes cantidades de información de diversas fuentes en un solo lugar, con el propósito de facilitar el análisis y la toma de decisiones. Es una parte clave de la gestión de la información empresarial y se utiliza comúnmente en entornos empresariales y organizativos.

### **Base de datos transaccional**

Son bases de datos que tiene como fin el envío y recepción de datos a gran velocidad. Están destinadas generalmente al entorno de análisis de calidad, datos de producción e industrial, y su objetivo principal es asegurar las transacciones dentro de una base de datos relacional o, en caso de que no se puedan asegurar, revertirlas, de manera que evitan que las transacciones queden incompletas, es decir, o se realiza la transacción o no pasa nada.

## **Almacén de datos y modelo dimensional**

El método de Kimball aboga por el uso de modelos dimensionales para el diseño del almacén de datos. En este enfoque, se utiliza una estructura de modelo de estrella, donde una tabla central de hechos está conectada a varias tablas de dimensiones. Este diseño simplifica el análisis y la consulta de datos.

### **Modelo dimensional**

Es un enfoque de diseño de bases de datos utilizado principalmente en almacenes de datos (DataWarehouse) para facilitar el análisis y la generación de informes. Se centra en la presentación y la comprensión de los datos de manera intuitiva, lo que lo hace especialmente útil para el análisis de negocios.

Las características clave del modelo dimensional incluyen:

- **Hechos:** representan los datos numéricos que se quieren analizar, como las ventas, el ingreso, la cantidad de productos vendidos, etc.
- **Dimensiones:** son las categorías descriptivas que proporcionan contexto a los hechos. Por ejemplo, una dimensión podría ser el tiempo, el lugar, el producto, el cliente, etc.
- **Esquema en estrella:** en el modelo dimensional, los hechos están en el centro de un esquema en estrella, rodeados por las dimensiones. Esto facilita las consultas y el análisis porque los datos relevantes para un análisis específico están organizados de manera clara.
- **Esquema de copo de nieve:** una variación del esquema en estrella es el esquema de copo de nieve, que normaliza aún más las dimensiones dividiéndolas en niveles más detallados.

La principal ventaja del modelo dimensional es su capacidad para proporcionar una visión clara y sencilla de los datos, lo que facilita el análisis y la generación de informes. Esto lo hace especialmente útil para las necesidades de negocio, donde la comprensión rápida de los datos es esencial. El modelo dimensional se utiliza comúnmente en combinación con herramientas de inteligencia empresarial (BI) para permitir a los usuarios explorar y analizar datos de manera efectiva.

### **Perfilado de datos**

Mejor conocido como Data Profiling, es un proceso que implica analizar y descubrir los datos en un conjunto específico. Este análisis proporciona una comprensión detallada de la calidad de los datos que servirán como insumo para las transformaciones en ETL.

A continuación, se muestran una serie de pasos para realizar el perfilado de datos según Kimball:

- Realizar una evaluación de las fuentes de datos para asegurarse de que son viables.
- Identificación de problemas de calidad en los datos.

- Evaluación de los problemas de calidad encontrados y definición de las posibles soluciones a aplicar en los ETL.
- Identificación de reglas de negocio que sean implícitas.

### **Big Data**

El termino Big Data, en español “gran volumen de datos”, comenzó a aparecer cuando los métodos tradicionales de almacenamiento comenzaron a no ser tan eficientes en este nuevo entorno que requería mucho más de lo que una herramienta de almacenamiento podía manejar. En general, Big Data se puede definir con tres V principales:

- Volumen.
- Variedad.
- Velocidad.

### **Herramientas ETL (Extracción, Carga y Transformación).**

Para la implementación del proceso ETL, se hace uso de herramientas como Talend Open Studio. Estas herramientas facilitan la extracción de datos desde diversas fuentes, la transformación según las necesidades del negocio y la carga eficiente en el almacén de datos.

Además de los conceptos anteriormente descritos, es importante conocer cuáles son algunas tecnologías utilizadas para el apoyo de la aplicación de la ingeniería de datos.

### **Amazon S3**

Amazon Simple Storage Service (S3) es un servicio de almacenamiento en la nube de Amazon que permite almacenar y recuperar cualquier cantidad de datos en cualquier momento. Se utiliza comúnmente para almacenar datos no estructurados, como archivos y documentos, y es una opción popular para almacenar datos antes de cargarlos en un almacén de datos.

### **Amazon Redshift**

Amazon Redshift es un servicio de almacenamiento de datos en la nube basado en arquitectura de almacén de datos columnares. Está diseñado para ejecutar consultas analíticas complejas sobre grandes conjuntos de datos con un rendimiento rápido. Redshift se integra fácilmente con otras herramientas de AWS y se utiliza comúnmente como el almacén de datos principal en arquitecturas de análisis en la nube.

### **Integración de Power BI para visualización y análisis**

Power BI se integrará como la herramienta principal de visualización y análisis de datos. Permite la creación de paneles interactivos y reportes basados en datos del almacén, proporcionando a los usuarios finales la capacidad de tomar decisiones informadas.

# CAPÍTULO I: ESPECIFICACIÓN DEL PROYECTO

## a. Situación actual

### i. Antecedentes

Importaciones El Ángel es una empresa salvadoreña que apoya el diseño y la decoración del hogar. La misión de la empresa, desde su inicio, ha sido proporcionar a los clientes en El Salvador acceso a una amplia gama de productos decorativos para el hogar de alta calidad y estilo, importados de diversas partes del mundo.

La empresa comenzó como una pequeña tienda física en San Salvador, especializándose en productos decorativos exclusivos y de lujo. Con el tiempo, la entidad ha evolucionado y ha diversificado su negocio para incluir una tienda en línea, capitalizando la creciente tendencia de compras en línea en El Salvador.

A partir de esto las operaciones de Importaciones El Ángel se realizan en su mayoría de forma digital, haciendo uso del sistema de comercio en línea Magento Commerce. Esto le ha permitido persistir una gran cantidad de datos sobre sus clientes y las compras que estos realizan a través de la plataforma electrónica mencionada.

Los usuarios pueden ingresar a la plataforma como invitados y visualizar el catálogo de productos. Sin embargo, para realizar compras es necesario que los usuarios realicen un registro de usuario dentro de la plataforma, este registro se realiza ingresando datos para la identificación univoca de él mismo, así como datos de dirección de envío. Para realizar la compra los usuarios registrados pueden añadir los productos deseados al carrito de compras, modificando las cantidades o eliminar productos del carrito en caso de ser necesario. Además, la cuenta de usuario permite el uso de aplicación de cupones sobre los productos puestos en el carrito, siempre y cuando las reglas de los cupones así lo permitan.

Una vez finalizada la adición de productos al carrito, se procede a la selección del método de pago y la posterior especificación de los datos necesarios para el método de pago seleccionado. Al verificar el método de pago se procede a la confirmación del pedido.

El usuario puede realizar el seguimiento de sus pedidos consultando los estados de los mismos.

Para la gestión de las diferentes tiendas en línea se utiliza el usuario administrador el cual se encarga de la gestión de productos que consiste en agregar productos, editar información existente, precios, entre otros.

Adicional a esto el usuario administrador realiza la gestión de pedidos como revisar los pedidos diferenciados por sus estados que son pendientes, en proceso, completados y cancelados. También es capaz de actualizar el estado de los pedidos según sea necesario

Este usuario es el encargado de la configuración de las reglas de precios y promociones sobre los productos, ya sea de forma directa o mediante la configuración de cupones.

## ii. Descripción del problema

Importaciones El Ángel requiere de un medio para realizar consultas sobre los datos históricos relacionados con las ventas realizadas de manera electrónica a través de su plataforma de comercio en línea Magento Commerce.

Se ha planteado una lista de necesidades de información que puedan ser mostrados en un formato de fácil lectura y acceso para los usuarios gerenciales. Es importante tomar en cuenta que los pedidos a proveedores para la reposición de inventario se realizan cada tres meses. Por esto es importante conocer el comportamiento de las ventas en este período de tiempo.

Los requerimientos establecidos por Importaciones El Ángel son:

- Volumen de ventas totales cada trimestre por tienda.
- Producto con mayores ventas durante los últimos 3 meses por tienda.
- Producto con mayores ventas durante los últimos 3 meses. (Acumulado de todas las tiendas).
- Servicio de pago más usado en los últimos 3 meses.
- Meses con mayores ventas al año por tienda.
- Porcentaje de clientes a los que se les completa una venta mayor a \$600.
- Monto total de ventas mensuales por cliente.
- Ventas totales por cliente de un mismo departamento.
- Total, de descuentos aplicados en los últimos meses por tienda.

Se requiere que la visualización de la información anteriormente planteada, tenga un aspecto amigable, entendible, de fácil manipulación y de carácter dinámico. Se desea que la presentación de la información se muestre a través del uso de los gráficos que mejor se adecuen y representen las necesidades de información. El componente dinámico, se debe establecer a través de diferentes filtros de datos que contribuyan de pasar de una visualización general a información con un nivel de detalle más específico.

### iii. Planteamiento del problema

En un acercamiento inicial a la situación actual de Importaciones El Ángel, se han detectado los siguientes planteamientos que deben ser solucionados mediante el desarrollo del proyecto:

Se debe mostrar la información requerida por Importaciones El Ángel, en un formato amigable y de fácil visualización para los usuarios gerenciales.

Los datos necesarios para responder a los requerimientos planteados por el cliente, se encuentran actualmente dispersos en las tablas del modelo relacional dentro de la base de datos transaccional. Estos datos deben relacionarse a través de consultas SQL que den como resultado tablas consolidadas que respondan a las necesidades establecidas.

La información debe visualizarse en un formato amigable a través de tableros con gráficos, que tengan una complejidad de uso baja y que posea un comportamiento dinámico en base a filtros de datos.

La consulta de datos a la base transaccional debe tener un impacto mínimo sobre el rendimiento del servidor de bases de datos, de modo que las peticiones realizadas por los clientes no se vean afectadas causando que la página web tenga un rendimiento lento.

El proceso de transformación de los datos desde el sistema transaccional, debe ser automatizado en la medida de lo posible; debe ser diseñado para su reutilización posterior; así mismo, el diseño de las tareas de automatización debe tomar en cuenta los posibles cambios que pueda sufrir los datos, evitando conflictos y duplicidad de información.

La solución diseñada debe poder ejecutarse bajo el sistema operativo Windows 7 o superior.

Se debe garantizar un amplio grado de disponibilidad sobre los tableros de datos, así como de los datos utilizados para generarlos; esto a través del uso de servicios de almacenamiento en la nube.

## **b. Objetivos**

### **Objetivo General.**

Diseñar el modelo dimensional para un DataWarehouse, así como los respectivos tableros de visualización de métricas, mediante el análisis y comprensión del esquema de base de datos de las ventas realizadas a través del sistema transaccional Magento Commerce, con el fin de satisfacer las necesidades analíticas de “Importadora El Ángel” otorgando una herramienta para la toma de decisiones estratégicas del negocio

### **Objetivos específicos.**

- Identificar las necesidades analíticas del cliente comprendiendo sus objetivos comerciales, preguntas clave que desean responder y los desafíos que buscan resolver mediante el análisis de datos.
- Analizar la estructura de la base de datos transaccional de Magento Commerce, identificando las tablas y relaciones más relevantes involucradas en los procesos de venta, teniendo en cuenta las necesidades analíticas del cliente.
- Determinar las dimensiones necesarias para el diseño del modelo dimensional, identificando los atributos clave en el modelo transaccional agrupándolos de manera lógica bajo estructura semántica.
- Diseñar el esquema dimensional acoplando las dimensiones en torno a la tabla de hechos.
- Construir rutinas automatizadas para la extracción y transformación de los datos, así como para la posterior carga dentro del esquema dimensional que corresponde al DataWarehouse.
- Integrar de manera efectiva la herramienta Power BI con los datos procesados a través del proceso ETL, con el objetivo de crear elementos visuales y dashboards interactivos. Estos recursos visuales proporcionarán a la gerencia una representación gráfica y comprensible de los datos empresariales clave.

### c. Alcances

Con la realización del presente proyecto de desarrollo, se pretende realizar lo siguiente:

- Diseño de modelo dimensional para un DataWarehouse de análisis de ventas de la empresa Importaciones El Ángel, utilizando un esquema arquitectónico de estrella.
- El origen de datos para el proceso de transformación será la base de datos transaccional de Magento Commerce a través de la cual se registran las transacciones de ventas en las tiendas de Importaciones El Ángel; Las credenciales de acceso proporcionadas serán de solo lectura; y el acceso a las tablas del origen de datos es únicamente para la lectura de las tablas: catalog\_product\_entity, catalog\_product\_entity\_text, catalog\_product\_entity\_varchar, customer\_address\_entity, customer\_entity, sales\_order, sales\_order\_address, sales\_order\_item, sales\_order\_status, sales\_order\_payment y store.
- Utilizando Talend Open Studio se realizarán tareas automatizadas para la extracción de datos del sistema transaccional, procesamiento de datos que permitan transformar la porción del modelo transaccional extraído, en la estructura del modelo dimensional de estrella que representa el DataWarehouse diseñado.
- Utilizando Talend Open Studio se realizarán tareas automatizadas para la transformación y carga del modelo dimensional para ser resguardado en las bases de datos destino, tanto local como en la nube.
- El esquema de estrella resultante del proceso de transformación de datos, se almacenará en dos lugares; en primer lugar, dentro de un servidor local de bases de datos bajo MySQL del cual se proporcionará acceso de lectura y escritura. Se proporcionarán los scripts SQL necesarios para la creación del esquema, así como el orden de ejecución de los mismos. En segundo lugar, se realizará un resguardo de los datos resultantes de la transformación dentro de la plataforma de Amazon S3, en el que se cargarán los archivos .CSV correspondientes a cada dimensión y tabla de hechos. Así mismo, se creará el DataWarehouse utilizando el servicio de Amazon Redshift.
- Utilizando PowerBI, se presentará un tablero de visualización en el cual se podrá realizar lectura y análisis de las métricas establecidas por el cliente.

#### **d. Justificación**

El desarrollo del presente proyecto surge a raíz de la detección de una oportunidad de mejora para suplir las necesidades de información de Importaciones El Ángel, quienes han expresado su interés en adquirir una solución de software que le permita visualizar comportamientos de las ventas registradas por su sistema informático de comercio en línea Magento Commerce.

A través de los tableros de análisis de datos solicitados, se pretende:

Optimizar las estrategias de mercadeo, evaluando el impacto de las estrategias aplicadas en los meses anteriores vs el volumen de ventas obtenido para los productos sobre los que se ha aplicado la estrategia.

Ofrecer un mejor servicio a los clientes poniendo a disponibilidad los productos más populares. Realizando un análisis de las tendencias de ventas, es posible predecir la popularidad de los productos, abasteciendo el stock para mantener una alta disponibilidad de estos.

Importaciones El Ángel pretende potenciar el impacto de las decisiones gerenciales al plantearlas bajo las sólidas bases estadísticas que proporcionan los datos históricos recolectados y la presentación correcta de los mismos a través de tableros de datos.

### e. Cronograma de actividades.

Cronograma de actividades  
Read-only view, generated on 03 Dec 2023

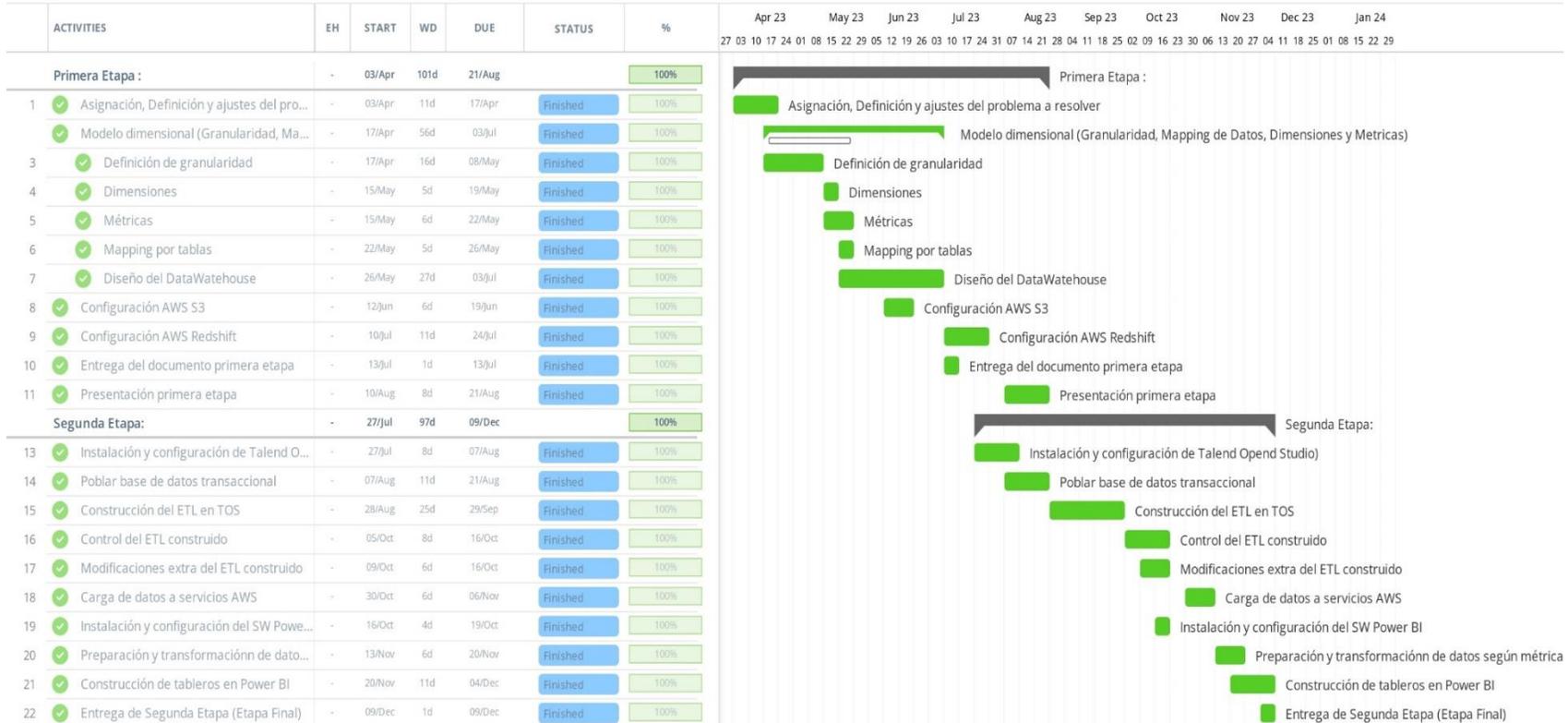


Ilustración 1. Cronograma de actividades

## f. Presupuesto

El presupuesto utilizado se ha dividido en tres rubros diferentes, los cuales se detallan a continuación:

- **Recurso humano**

Recurso humano				
Cargo	Salario por mes (USD)	Meses	Cantidad de recurso	Costo sub total (USD)
Informático	\$800	8	3	\$19,200.00
<b>Costo total</b>				<b>\$19,200.00</b>

Tabla 1. Costo de recurso humano

El monto asignado al costo del recurso humano está basado en las ofertas de empleo de TI del mercado laboral salvadoreño, esto se encuentra indicado en la bibliografía.

**Cálculo:**

$$\text{Costo Total}_{RRHH_T} = (\text{Sueldo al mes en USD}) * (\text{Meses}) * (\text{Personas})$$

$$\text{Costo Total}_{RRHH_T} = 800 * 8 * 3$$

$$\therefore \text{Costo Total}_{RRHH_T} = 19,200 \text{ USD}$$

- **Hardware y software**

Hardware y software				
Elemento	Descripción	Cantidad	Costo individual (USD)	Costo sub total (USD)
RAM	Capacidad: 8GB	3	\$36.00	\$108.00
Procesador	Especificación: Intel Core i511th Gen	3	\$174.00	\$522.00
Disco duro	Especificación: SSD 500 GB	3	\$40.00	\$120.00
<b>Costo total</b>				<b>\$750.00</b>

Tabla 2. Costo de hardware y software

- **Costos indirectos asociados**

Costos indirectos asociados			
Servicio	Tarifa por mes	Meses	Costo sub total (USD)
Energía eléctrica	\$36.41	24	\$873.84
Internet residencial	\$30.00	24	\$720.00
<b>Costo total</b>			<b>\$1,593.84</b>

Tabla 3. Costos indirectos asociados

**Cálculo:**

$$\text{Consumo por mes}_{Energía} = (\text{Consumo en horas al mes}) * (Kvh) * (\text{Días del mes})$$

$$\text{Consumo por mes}_{Energía} = 8 * 0.1517 * 30$$

$$\therefore \text{Consumo por mes}_{Energía} = 36.41 \text{ USD}$$

Considerando que el consumo de energía se realizará durante 8 meses por cada uno de los miembros del equipo, se realiza la sumatoria, llegando al resultado de 24 meses en total durante el proyecto.

- **Servicios de terceros**

<b>Servicios de terceros</b>			
Servicio	Tarifa por mes	Meses	Costo sub total (USD)
Amazon Web Services	\$9.60	8	\$76.80
S3 Bucket	\$0.15	8	\$1.20
<b>Costo total</b>			<b>\$78.00</b>

*Tabla 4. Costos de servicios de terceros*

Para el cálculo de los servicios a terceros se hizo uso de la calculadora que provee AWS indicada en la bibliografía.

Con los cálculos anteriores se considera una estimación de costos totales de:

<b>Estimación total</b>	
<b>Rubro</b>	<b>Costo sub total (USD)</b>
Recurso humano	\$19,200.00
Hardware y software	\$750.00
Costos indirectos asociados	\$1,593.84
Servicios de terceros	\$78.00
<b>Total</b>	<b>\$21,621.84</b>

*Tabla 5. Estimación de costos totales*

## CAPÍTULO II: ANÁLISIS Y DISEÑO DE LA PROPUESTA DE SOLUCIÓN

### a. Metodología de trabajo

Este proyecto busca realizar el análisis de la información de las ventas de la empresa “Importaciones El Ángel” que utiliza la tienda en línea administrada a través de Magento Commerce, mediante la cual se recolectan los datos de las ventas, siendo estos almacenados en la base de datos transaccional.

Como primer paso es necesario establecer las necesidades de información que se alineen con los objetivos que “Importaciones El Ángel” desea cumplir y sobre las cuales se fundamentará el análisis y diseño tanto del DataWarehouse como de los tableros de datos.

Las necesidades analíticas encontradas se plantean a continuación:

- Volumen de ventas totales cada trimestre por tienda.
- Producto con mayores ventas durante los últimos 3 meses por tienda.
- Producto con mayores ventas durante los últimos 3 meses. (Acumulado de todas las tiendas).
- Servicio de pago más usado en los últimos 3 meses.
- Meses con mayores ventas al año por tienda.
- Porcentaje de clientes a los que se les completa una venta mayor a \$600.
- Monto total de ventas mensuales por cliente.
- Ventas totales por cliente de un mismo departamento.
- Total, de descuentos aplicados en los últimos meses por tienda.

Para el desarrollo de la solución se opta por utilizar el modelo de Ralph Kimball, siguiendo el ciclo de vida de un proyecto para DataWarehouse como se presenta en la siguiente imagen.

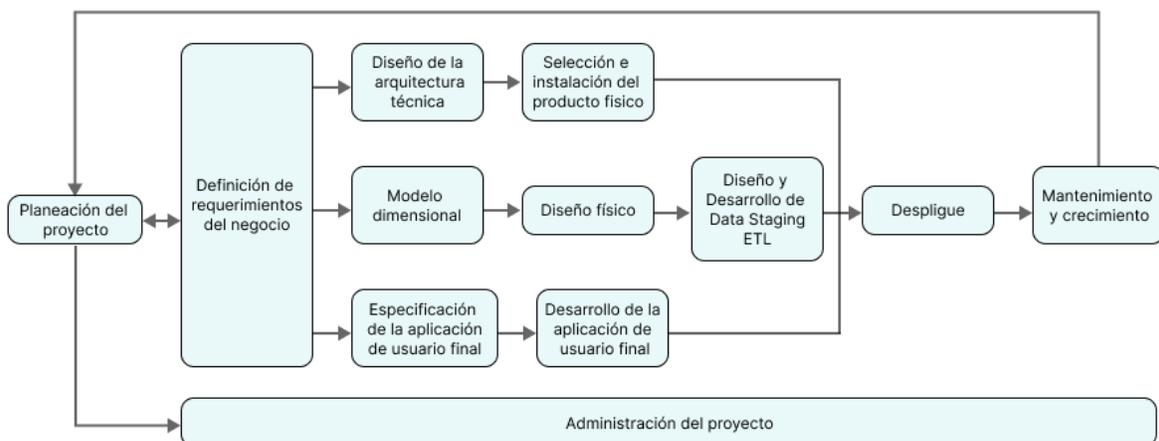


Ilustración 2. Modelo de ciclo de vida de un proyecto de DW

### **Planeación del Proyecto.**

Se establece la definición y el alcance de lo que se pretende alcanzar con el desarrollo del proyecto, las justificaciones y evaluaciones de factibilidad. Se centra en la gestión de recursos, establecimiento de perfiles, tareas, duración y secuencia de ejecución. Esta etapa identifica el escenario del proyecto para saber por qué surge la necesidad del DataWarehouse.

A lo largo del ciclo de vida, la planificación y la dirección de las tareas del proyecto mantiene las actividades en marcha. Entre las tareas que abarca esta etapa están:

- Identificación de los usuarios.
- Motivaciones del negocio
- Análisis de factibilidad (es el análisis financiero, económico y social de una inversión).

### **Definición de los requerimientos del negocio.**

La técnica utilizada para nivelar los requerimientos de los analistas de negocio difiere de los enfoques tradicionalistas guiados por los datos. Los diseñadores de los DataWarehouse deben entender los factores claves que guían al negocio para determinar efectivamente los requerimientos y traducirlos en consideraciones de diseño apropiadas, pues son la base para las tres etapas paralelas subsiguientes focalizadas en la tecnología, los datos y las aplicaciones, por lo cual es altamente crítica y el diseño es el centro de atención del BDL (Business Dimensional Lifecycle).

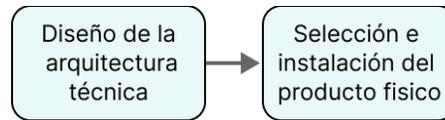
Así pues, al momento de empezar con el análisis del diseño, es necesario entender las necesidades del negocio, así como la realidad de las fuentes de datos existentes. Existen cuatro decisiones clave de diseño que deberán ser tomadas en conjunto con los representantes del negocio, estas serán:

- Selección del proceso de negocio.
- La granularidad
- Identificación de dimensiones
- Identificaciones de métricas.

Una vez establecido el proceso de negocio, granularidad, dimensiones y métricas que se necesitaran, el equipo de desarrollo puede empezar con el diseño del modelo dimensional y su respectiva implementación.

### **Eje tecnológico del modelo**

Las iniciativas analíticas del DataWarehouse como el de Power BI, suelen enfrentarse a múltiples problemas, entre ellos: integración de datos y compatibilidad de tecnologías. Es por eso que antes de adquirir algún producto, la recomendación es diseñar una arquitectura adecuada para la compañía.



*Ilustración 3. Eje tecnológico*

### **Diseño de la Arquitectura Técnica**

Los ambientes de Data Warehousing requieren la integración de numerosas tecnologías. Se debe tener en cuenta tres factores: los requerimientos del negocio, los actuales ambientes técnicos y las directrices técnicas estratégicas planificadas para de esta forma poder establecer el diseño de la arquitectura técnica del ambiente de DataWarehousing.

La arquitectura de un DataWarehouse consiste en cuatro servidores: Un servidor de ETL, un servidor de bases de datos, un servidor OLAP y un servidor de reportes.

### **Selección e instalación del producto físico**

Utilizando el diseño de arquitectura técnica como marco, es necesario evaluar y seleccionar componentes específicos de la arquitectura, como la plataforma de hardware, el motor de base de datos, la herramienta de ETL o el desarrollo pertinente, herramientas de acceso, etc. Una vez evaluados y seleccionados los componentes determinados, se procede con la instalación y prueba de estos en un ambiente integrado de Data Warehousing.

### **Eje de datos del modelo**

Una vez definido el eje tecnológico del modelo seleccionado e instalado las herramientas/servicios que contendrán los datos, es momento de definir esos datos y cómo se modelan.



*Ilustración 4. Eje de datos del modelo*

## **Modelado dimensional.**

Se inicia con una matriz en la que se determina la dimensionalidad de cada métrica y luego se especifican los diferentes grados de detalle (atributos), dentro de cada concepto del negocio (dimensión), así como la granularidad de cada métrica y las jerarquías que dan forma al modelo dimensional del negocio.

El modelo dimensional puede ser físico o lógico. Un modelo dimensional físico generalmente se representa en forma de esquema de estrella o de copo de nieve, en el que los objetos que contiene son tablas de base de datos. El esquema dimensional puede incluso adoptar la forma de una sola tabla o vista, en la que todos los hechos y dimensiones están en columnas distintas de dicha tabla o vista.

En un esquema dimensional lógico, los hechos, las medidas y las dimensiones se representan como entidades y atributos independientes a un proveedor de base de datos y, por lo tanto, se pueden transformar en un esquema dimensional físico para cualquier proveedor de base de datos.

Durante la fase de diseño se debe identificar los procesos de negocio que desea modelar, después identificar la granularidad y las dimensiones y medidas del proceso de negocio.

En la definición de requerimientos se deben definir las siguientes cuatro decisiones:

- **Identificar los procesos de negocio:** Seleccionar el proceso de negocio para el que se asignará el modelo dimensional. Según la selección, se reúnen los requisitos del proceso de negocio. En ocasiones un proceso de negocio requiere más de un modelo dimensional.
- **Identificar la granularidad:** Identificar la granularidad de cada tabla de hechos y proceso de negocio. Durante este proceso se deben identificar los tipos de tablas de hechos y los candidatos preliminares para las dimensiones y medidas.
- **Identificar las dimensiones:** Una vez se haya determinado la granularidad del modelo, identificar las dimensiones verdaderas para ese nivel de granularidad. Para el esquema de copo de nieve se deben crear columnas, jerarquías y casos.
- **Identificar las medidas:** Durante este paso del ciclo de diseño de modelo dimensional, se identifican las medidas y el tipo de medidas incluidas en el modelo dimensional.

A continuación, se detallan los conceptos claves utilizados en modelamiento dimensional.

- **Tablas y entidades de hechos (fact table):** Es una tabla o entidad de un esquema de estrella o copo de nieve que almacena medidas para medir el negocio, como las ventas, el coste de las mercancías o las ganancias.

- **Tablas y entidades de dimensiones(dimensión):** Es una tabla o entidad de un esquema de estrella, copo de nieve o constelación que almacena detalles acerca de hechos. Por ejemplo, una tabla de dimensión de hora almacena los distintos aspectos del tiempo, como el año, trimestre, mes y día.
- **Jerarquías:** Es una relación de muchos a uno entre los miembros de una tabla o entre tablas. Una jerarquía consta básicamente de distintos niveles, y cada uno corresponde a un atributo de dimensión.
- **Medidas.** Definen un atributo de medida y se utilizan en las tablas de hechos. Puede calcular medidas correlacionándolas directamente con un valor numérico en una columna o atributo. Una función de agregación resume el valor de las medidas para el análisis dimensional.
- **Modelo de estrella.** Está conformado por varias dimensiones relacionadas con una fact table, por lo cual un modelo estrella representa un proceso de negocio. Un modelo estrella tiene las siguientes características:
  - Una fact table que contiene métricas del proceso de negocio rodeado de dimensiones que proveen del contexto de cuando sucedió el evento.
  - Fácil de comprender por los usuarios del negocio
  - Simplicidad y simetría
  - Incrementa el rendimiento
  - Altamente prolongable con nuevas dimensiones y métricas.

### **Diseño físico.**

Este se focaliza en la selección de estructuras necesarias para soportar el diseño lógico. Los elementos principales de este proceso son la definición de convenciones estándares de nombres y formas de acceder específicas del ambiente de la base de datos. La indexación y las estrategias de particionamiento son también determinadas etapas.

Los procesos de ETL consideran la extracción de datos de diversos sistemas, aplican criterios de calidad y consistencia de datos. Luego los unifican para que puedan ser utilizados en conjunto y finalmente entregan los datos, en un formato dimensional, para que los desarrolladores/analistas BI puedan crear aplicaciones y los usuarios finales puedan tomar decisiones.

### **Diseño y desarrollo de Data Staging ETL.**

Las principales subetapas de esta zona del ciclo de vida son: la extracción, la transformación y la carga de datos (proceso ETL). Se definen como procesos de extracción a aquellos requeridos para obtener los datos que permitirán efectuar la carga del modelo físico acordado. Los procesos de

transformación sirven para convertir o codificar los datos fuente para cargar el modelo físico. Los procesos de carga de datos sirven para poblar el DataWarehouse.

Para realizar el proceso de ETL “en el eje técnico” se debe definir la herramienta que permita realizar el proceso de extracción, transformación y carga, apoyándose de tres criterios para su elección definitiva.

- Inicialmente conocer las posibilidades técnicas de cada herramienta ya que permite de manera directa entender si la herramienta será lo suficientemente potente y capaz para implementar las cosas que deseamos construir.
- La consideración de los conocimientos del equipo de trabajo, es decir tomar en cuenta las opiniones en cuanto a la experiencia con la que cuentan los miembros del equipo en el uso de las herramientas.
- La estimación de los precios existentes, así como también el respectivo licenciamiento que cada una de las herramientas posee.

Algunas herramientas para la implementación de ETL son:

**Talend Open Studio.** Es un popular software de integración de datos de código abierto que cuenta con una interfaz gráfica de usuario fácil de usar. Los usuarios pueden arrastrar y soltar componentes, configurarlos y conectarlos para crear canalizaciones de datos. Entre bastidores, Open Studio convierte la representación gráfica en código Java y Perl.

**PowerCenter.** Es una de las mejores herramientas ETL del mercado. Dispone de una amplia gama de conectores para almacenes y lagos de datos en la nube, como AWS, Azure, Google Cloud y Salesforce. Sus herramientas de bajo y ningún código están diseñadas para ahorrar tiempo y simplificar los flujos de trabajo.

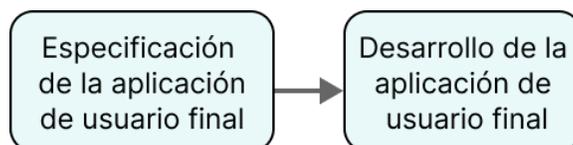
**Apache Airflow.** Es una plataforma de código abierto para crear, programar y supervisar flujos de trabajo mediante programación. La plataforma cuenta con una interfaz de usuario basada en web y una interfaz de línea de comandos para gestionar y activar flujos de trabajo.

**Oracle Data Integrator .** Es una herramienta ETL que ayuda a los usuarios a construir, desplegar y gestionar almacenes de datos complejos. Viene con conectores listos para usar para muchas bases de datos, como Hadoop, ERPs, CRMs, XML, JSON, LDAP, JDBC y ODBC.

**Hadoop.** Es un marco de código abierto para procesar y almacenar BigData en clústeres de servidores informáticos. Se considera la base del BigData y permite almacenar y procesar grandes cantidades de datos.

### **Eje de Business Intelligence del modelo**

A pesar de que se ha descrito el modelo en 3 ejes de forma secuencial, es importante recalcar que los 3 ejes se pueden trabajar en paralelo. Es importante aclarar esto, pues cuando se ejecuta el ciclo de vida del modelo, deben trabajarse los diferentes ejes en paralelo, optimizando tiempos y no dejando que cada componente sea aislado uno del otro.



*Ilustración 5. Eje de Business Intelligence del modelo*

### **Especificación de Aplicaciones para Usuarios Finales**

Los diferentes roles o perfiles de usuarios determinan la interfaz o ventana al DataWarehouse. Herramientas de diseño de reportes y consultas avanzadas para analistas, tableros de control para gerentes, acceso mediante internet para usuarios internos/externos remotos, envío de información por dispositivos no estándares para usuarios internos/externos, etc.

Se clasifican los usuarios según su perfil de consulta, desde usuarios con un perfil más estratégico y menos predecibles, hasta usuarios netamente operacionales que consumen una serie de reportes estándares pasando por los usuarios gerenciales con uso de interfaces push-button.

### **Desarrollo de Aplicaciones para Usuarios Finales**

Siguiendo a la especificación de las aplicaciones para usuarios finales, el desarrollo de las aplicaciones de los usuarios finales involucra configuraciones del metadata y construcción de reportes específicos.

Algunas herramientas para la visualización y reportes son:

**Microsoft Power BI.** Es un servicio de análisis de datos de Microsoft orientado a proporcionar visualizaciones interactivas y capacidades de inteligencia empresarial con una interfaz lo suficientemente simple como para que los usuarios finales puedan crear por sí mismos sus propios informes y paneles.

**Google Charts** .Es una herramienta basada en la web que puede crear visualizaciones sencillas a partir de conjuntos de datos pequeños y grandes.

**Tableau** .Es una potente y conocida herramienta de visualización de datos que permite analizar datos de múltiples fuentes a la vez.

**Google Analytics** . Es una herramienta de BI de probada eficacia, perfecta para pequeñas, medianas y grandes empresas que desean analizar la actividad de su sitio web. Google Analytics puede hacer un seguimiento de cifras críticas como la tasa de rebote, la duración media de la sesión y las páginas por sesión.

### **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. Hay varios factores extras que aseguran el correcto funcionamiento de todas estas piezas, entre ellos se encuentran la capacitación, el soporte técnico, la comunicación, las estrategias de feedback. Todas estas tareas deben tenerse en cuenta antes de que cualquier usuario pueda tener acceso al DataWarehouse.

Para que el despliegue sea exitoso, cada etapa anterior debe ser analizada y probada para detectar errores y sean resueltos en cada uno de los ejes mencionados y el problema no pase a la parte de visualización.

### **Mantenimiento y crecimiento**

DataWarehousing es un proceso de etapas bien definidas, con comienzo y fin, pero de naturaleza espiral pues acompaña a la evolución de la organización durante toda su historia. Se necesita continuar con los relevamientos de forma constante para poder seguir la evolución de las metas por conseguir. Según afirma Kimball, “si se ha utilizado el proceso anterior de forma completa, el DataWarehouse está preparado para evolucionar y crecer”.

Al contrario de los sistemas tradicionales, los cambios en el desarrollo deben ser vistos como signos de éxito y no de falla. Es importante establecer las prioridades para poder manejar los nuevos requerimientos de los usuarios y de esa forma poder evolucionar y crecer.

En la actualidad la mayoría de las industrias se enfocan en el uso de metodologías ágiles. Siendo el ciclo de vida de Kimball una buena solución, ya que permite que el modelo siga creciendo al compartir ciertas características con otras metodologías ágiles como enfocarse en aportar valor al negocio, colaboración directa con el mismo y un desarrollo incremental.

## Administración del Proyecto

La administración del proyecto asegura que las actividades del ciclo de vida del modelo dimensional se ejecuten en forma sincronizada. Como lo indica el diagrama, la administración acompaña todo el ciclo de vida. Entre sus actividades principales se encuentra el monitoreo del estado del proyecto y la comunicación entre los requerimientos del negocio y las restricciones de información para poder manejar correctamente las expectativas en ambos sentidos.

Con una buena administración se pueden detectar puntos problemáticos de un DataWarehouse. Los principales puntos de atención que pueden llegar a complicar un proyecto de Data Warehousing se clasifican en las siguientes tres áreas:

- **Rutinas de Carga:** Incluye programas de extracción y limpieza de datos. Surgen problemas en este punto dada la falta de integración y estructura consistente alineada entre los sistemas fuentes.
- **Mantenimiento:** Dados los diferentes períodos de almacenamiento para OLTP y el OLAP y el hecho de que los DW son sistemas secundarios de información, otro problema surge para sincronizar los datos entre los sistemas operacionales fuentes y los DataWarehouse.
- **Afinación:** En los patrones de uso y los métodos de acceso típicos de los sistemas OLAP, diseñadores y administradores deben realizar cambios significativos a los implementados en la afinación de sistemas OLTP.

## **b. Descripción de la propuesta de solución**

Para este punto se han considerado los siguientes elementos:

a) Definición de métricas.

Las métricas de interés establecidas por Importaciones El Ángel son:

- Volumen de ventas totales cada trimestre por tienda.
- Producto con mayores ventas durante los últimos 3 meses por tienda.
- Producto con mayores ventas durante los últimos 3 meses. (Acumulado de todas las tiendas).
- Servicio de pago más usado en los últimos 3 meses.
- Meses con mayores ventas al año por tienda.
- Porcentaje de clientes a los que se les completa una venta mayor a \$600.
- Monto total de ventas mensuales por cliente.
- Ventas totales por cliente de un mismo departamento.
- Total, de descuentos aplicados en los últimos meses por tienda.

b) Análisis del modelo transaccional.

A continuación, se presenta la porción del modelo de la base de datos transaccional extraída desde Magento Commerce. Esta contiene las tablas relevantes que contribuyen al modelado dimensional para las métricas establecidas.

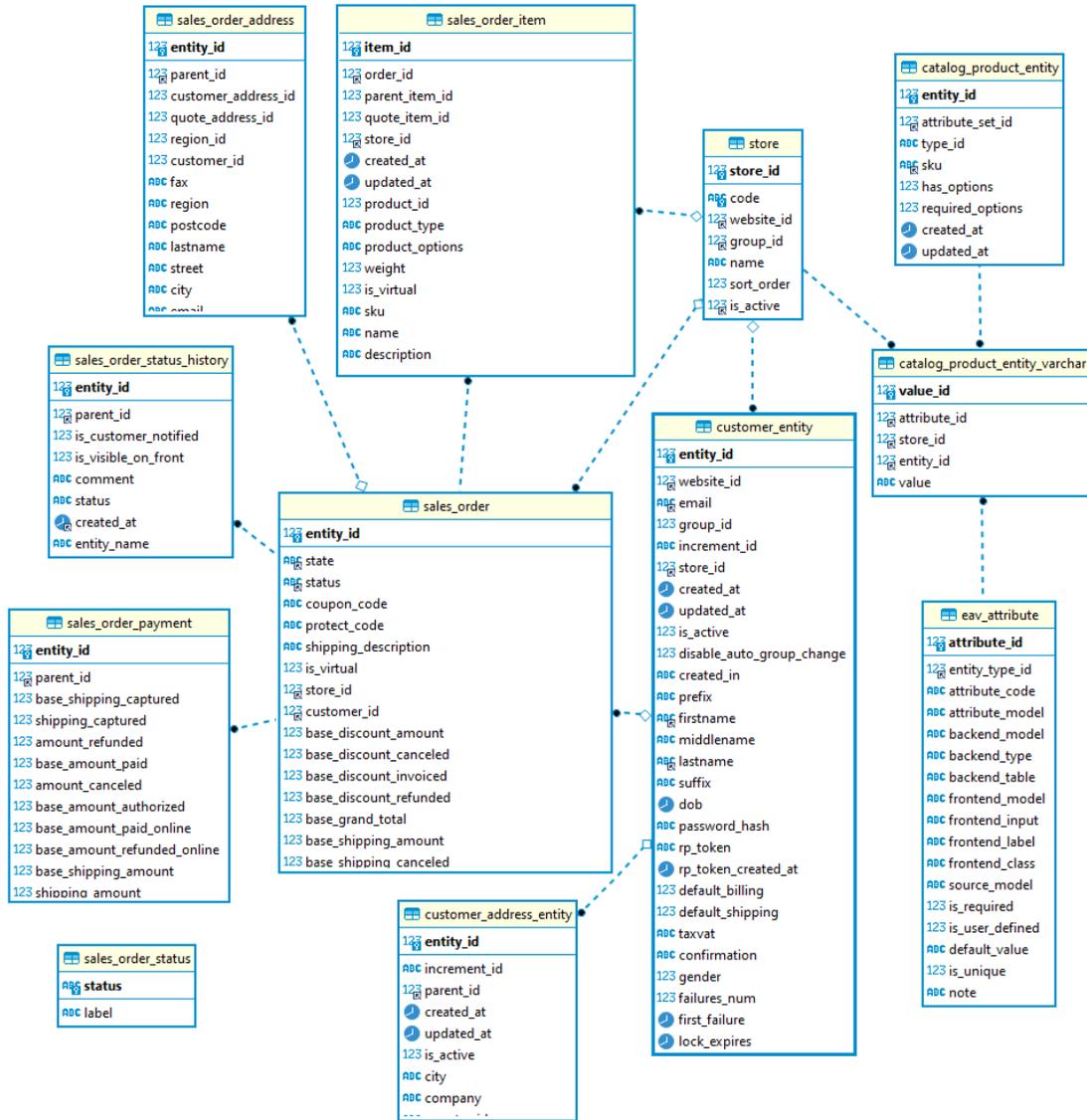


Ilustración 6. Modelo de la base de datos transaccional

c) Granularidad.

Posterior al análisis del modelo transaccional en conjunto con las métricas planteadas se ha definido el nivel de granularidad donde una fila representa una línea de venta de un producto por cliente, tienda, fecha y método de pago.

d) Diseño del modelo dimensional.

El modelo dimensional a partir del análisis de las métricas y la base de datos transaccional se muestra a continuación.

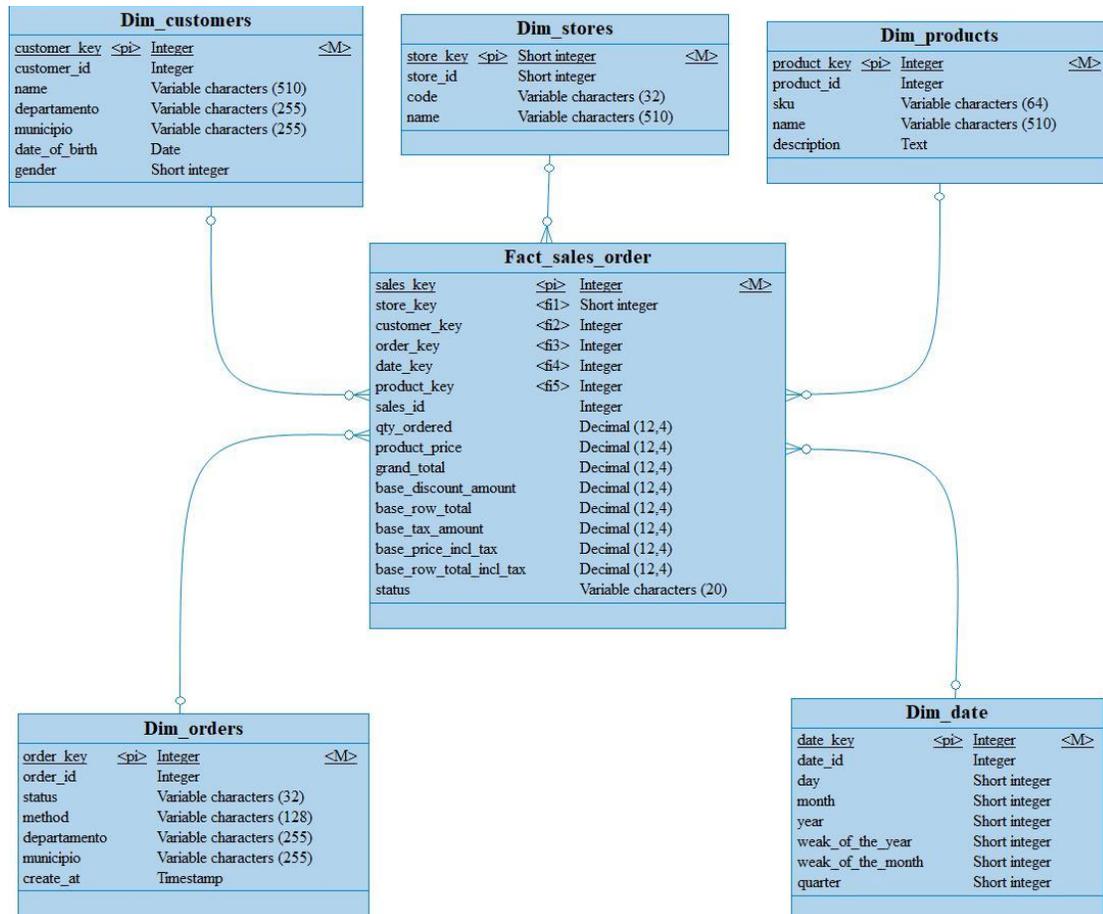


Ilustración 7. Modelo dimensional de estrella

e) Diseño de ETL.

Para el diseño de este se hace de la herramienta Talend Open Studio la que facilita el desarrollo y la automatización de las tareas de Extracción, transformación y carga. Se han establecido dos grupos de Jobs.

- Jobs de extracción y transformación.
- Jobs de carga.

f) **Presentación.**

Se utiliza la herramienta PowerBI para el diseño y la creación de los tableros cuyas visualizaciones serán alimentadas con los datos del DataWarehouse que han sido cargados a la plataforma de Amazon Redshift con la cual se establecerá la conexión remota.

**c. Descripción de la tecnología a utilizar**

A continuación, se detallan las herramientas de las que se hizo uso para el desarrollo del proyecto, y su respectiva descripción.

**Excel**

Microsoft comercializó originalmente un programa para las hojas de cálculo llamado Multiplan en 1982, que fue muy popular en los sistemas CP/M (programas de control para microcomputadoras), pero en los sistemas MS-DOS perdió popularidad frente al Lotus 1-2-3. Microsoft publicó la primera versión de Excel para Mac en 1985, y la primera versión de Windows (numeradas 2-05 en línea con el Mac y con un paquete de tiempo de ejecución de entorno de Windows) en noviembre de 1987.

Y se ha decidido trabajar con esta herramienta para el manejo de archivos en forma de tablas. Además de poder trabajar de forma colaborativa en la nube sin costo, su facilidad de uso así mismo todo el equipo tiene un previo conocimiento de la misma.



**Magento Commerce**

Es una plataforma de código abierto para comercio electrónico escrita en PHP. Fue desarrollada con apoyo de voluntarios por “Varien Inc”, una compañía privada con sede en Culver City, California. “Varien” publicó la primera versión del software el 31 de marzo de 2008.

Magento es la herramienta tecnológica utilizada por Importaciones El Ángel, por tanto, será la que proporcione la base de datos transaccional.



## MySQL

Es un sistema de gestión de bases de datos relacionales (RDBMS, por sus siglas en inglés) de código abierto. Fue desarrollado por Oracle Corporation y está distribuido bajo la Licencia Pública General de MySQL, que es una licencia de software libre y de código abierto. Este sistema de gestión de bases de datos utiliza el lenguaje de consulta estructurado (SQL) para administrar y manipular datos. MySQL es ampliamente utilizado en aplicaciones web y es una opción popular entre desarrolladores y empresas debido a su rendimiento, confiabilidad y facilidad de uso. Algunas características importantes por las cuales se ha decidido utilizar MySQL incluyen:

1. Multiplataforma: MySQL es compatible con varias plataformas, como Linux, Windows y macOS, lo que facilita su implementación en una variedad de entornos.
2. Como se mencionó anteriormente la base de datos transaccional será poblada desde Magento y este funciona sobre un sistema gestor de bases de datos MySQL.



## DBeaver

Es una herramienta de administración y desarrollo de bases de datos que proporciona un entorno gráfico para interactuar con diversas bases de datos. Se trata de un software de código abierto y gratuito que admite múltiples sistemas de gestión de bases de datos, lo que significa que puede utilizarse para trabajar con diferentes tipos de bases de datos, como MySQL, PostgreSQL, Oracle, Microsoft SQL Server, SQLite y muchos otros. DBeaver es compatible con Windows, Linux y macOS, lo que permite a los desarrolladores y administradores de bases de datos utilizarla en diversos entornos.

Debido a la compatibilidad con Windows, el soporte a bases de datos MySQL y en vista que la base de datos transaccional está definida en MySQL se tomó ha bien hacer uso de esta herramienta. Abonado a esto el conocimiento previo de la herramienta, facilita el uso de la misma.



## Talend Open Studio

Es una suite de software de integración de datos de código abierto. Se utiliza para realizar tareas relacionadas con la extracción, transformación y carga de datos (ETL). El principal objetivo de Talend

Open Studio es facilitar el movimiento y la transformación de datos entre diferentes sistemas y plataformas.

Debido a la gratuidad de la herramienta y a las amplias funcionalidades que proporciona se ha hecho uso de esta para el diseño del ETL.



### **Amazon Web Services(AWS)**

Es una plataforma de servicios en la nube ofrecida por Amazon.com. Proporciona una amplia variedad de servicios de computación en la nube, incluyendo almacenamiento, potencia de cálculo, bases de datos, análisis, aprendizaje automático, redes, Internet de las cosas (IoT), seguridad y muchos otros servicios. Estos servicios se ofrecen bajo demanda, lo que significa que los usuarios pueden acceder y utilizar recursos de manera flexible según sus necesidades.

a principal razón por la que se hará uso de la herramienta es porque se ocuparan 2 de los servicios que esta plataforma ofrece como lo es S3 y Amazon Redshift, el servicio AWS es de pago, como ya se mencionó en el capítulo anterior, pero brinda una capa gratuita de dos meses en ciertos productos para tomarlo en cuenta.



### **Amazon S3**

Amazon S3, o Simple Storage Service, es un servicio de almacenamiento en la nube ofrecido por Amazon Web Services (AWS). Proporciona un sistema de almacenamiento escalable y duradero, diseñado para almacenar y recuperar grandes cantidades de datos desde prácticamente cualquier lugar de la web.

Se utilizará para guardar los archivos generados del proceso ETL y conectar estos mismos con redshift.

### **Amazon S3**



## Redshift

Es un servicio de almacenamiento de datos en la nube ofrecido por Amazon Web Services. Se trata de un almacén de datos totalmente gestionado y altamente escalable que se utiliza para realizar rápidos análisis de grandes conjuntos de datos. Está diseñado específicamente para el procesamiento y análisis de datos en grandes volúmenes, lo que lo convierte en una solución efectiva para empresas que requieren un almacenamiento de datos robusto y capacidades de análisis de datos.

Se utilizará para el procesamiento de los datos que se reciban de Amazon S3 (modelamiento de base de datos que se cargará a Power BI).



## Power BI

Es una suite de herramientas de inteligencia empresarial (BI) desarrollada por Microsoft. Esta suite está diseñada para permitir a las personas visualizar y compartir información de manera eficaz, convirtiendo datos en información significativa y toma de decisiones informadas. Proporciona una amplia gama de funciones, desde la conexión a diversas fuentes de datos hasta la creación de informes interactivos y paneles de control.

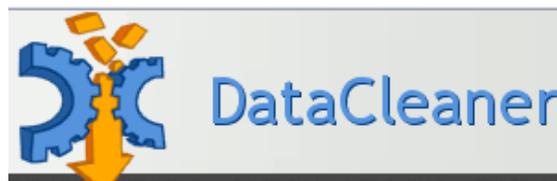
Se utilizará Power BI Desktop que es la herramienta gratuita de Power BI para el manejo de la información y obtener los reportes.



## DataCleaner

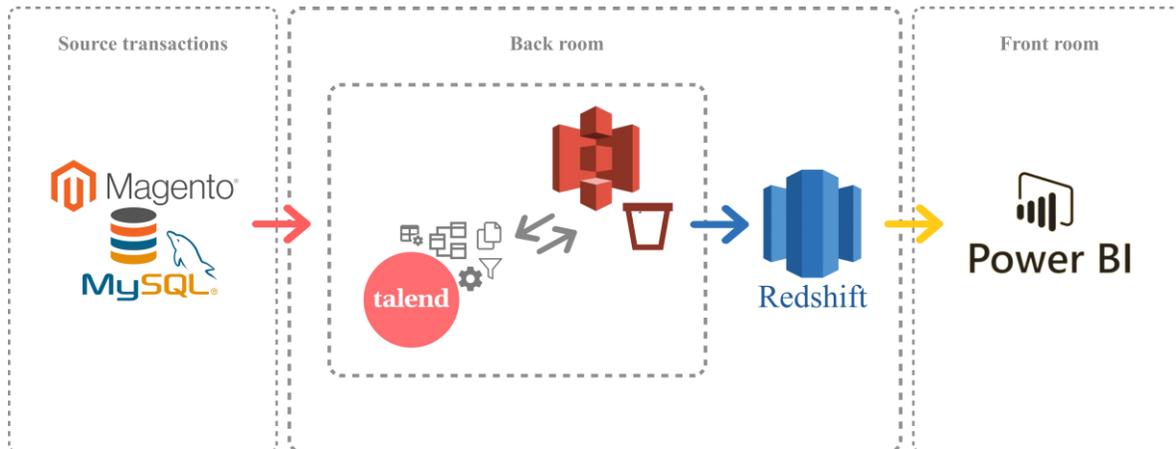
Es una herramienta Open Source (licencia LGPL) que mediante una amigable UI permite aplicar a los datos diversas técnicas referentes a la calidad de datos y data profiling.

Se utiliza para procesar los datos del modelo transaccional y brindar información agrupada de la misma para mejorar el análisis.



#### d. Diagrama arquitectónico de la solución

A continuación, se presenta el modelo arquitectónico de la solución propuesta, basado en el modelo de Ralph Kimball. En este se representan las tecnologías empleadas en cada una de las capas descritas en el modelo.



*Ilustración 8. Modelo de Ralph Kimball aplicado a la solución*

**Source transactions:** Primer capa del modelo de Ralph Kimball la cual representa el sistema origen de datos.

**Back room:** Segunda capa del modelo de Ralph Kimball la cual representa los elementos de extracción, transformación y carga de datos

**Front room:** Tercer capa del modelo de Ralph Kimball la cual representa el componente de visualización del modelo dimensional.

## e. Descripción de cada componente de la solución

### **Magento**

Software utilizado por Importaciones El Ángel para realizar su actividad comercial de manera electrónica, siendo el medio de entrada para los datos de la base de datos transaccional. En ella se almacenan los datos de la empresa y se obtiene su Base para trabajar bajo el modelo de ventas.

Magento cuenta con un total de 326 tablas, de las cuales para obtener las ventas de la empresa se hará uso de un total de 11 tablas para modelar las ventas, las tablas son:

- catalog\_product\_entity
- catalog\_product\_entity\_text
- catalog\_product\_entity\_varchar
- customer\_address\_entity
- customer\_entity
- sales\_order
- sales\_order\_address
- sales\_order\_item
- sales\_order\_status
- sales\_order\_payment
- Store

El modelo transaccional se almacena como base de datos con el nombre de **tienda**.

### **Talend Open Studio**

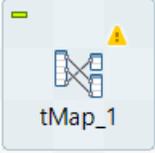
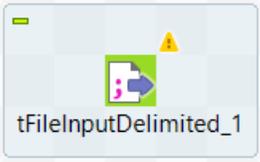
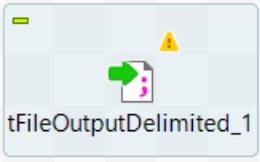
#### **Primera transformación de los datos**

Luego de obtener las tablas necesarias para el modelo de ventas de la empresa “Importaciones El Ángel”, se pasa al segundo componente (back room) que es la primera transformación de los datos mediante la herramienta de Talend Open Studio, esta transformación de los datos lleva a la creación de las dimensiones y la tabla de hechos, siguiendo el modelo de estrella (modelo que se verá más adelante) del cual salen 5 dimensiones y la tabla de hechos las cuales son:

- dim\_producto
- dim\_cliente
- dim\_orden
- dim\_fecha
- dim tienda

- fact\_ventas

A continuación, se describen los componentes de Talend Open Studio utilizados en los Jobs del ETL.

Componente	Descripción
	<p><b>tDBInput:</b> Suele utilizarse para leer datos desde una fuente de base de datos. Estos componentes permiten especificar la conexión a la base de datos, la consulta SQL o el procedimiento almacenado que se ejecutará y otros parámetros relevantes para extraer datos de la fuente de datos.</p>
	<p><b>tMap:</b> Se utiliza para realizar transformaciones de datos durante el proceso de integración. Es una herramienta gráfica que permite definir reglas de mapeo y transformación entre los datos de entrada y salida en un flujo de trabajo.</p>
	<p><b>tDBSCD:</b> Se utiliza para implementar procesos de Slowly Changing Dimensions (SCD) en el contexto de la integración de datos y almacenamiento en bases de datos.</p>
	<p><b>tRunJob:</b> Se utiliza para ejecutar subtrabajos (subjobs) dentro de un trabajo principal. Permite modularizar y reutilizar lógica al dividir un trabajo grande en tareas más pequeñas y manejables.</p>
	<p><b>tFileInputDelimited:</b> Se utiliza para leer datos desde archivos de texto delimitados, como archivos CSV (Comma-Separated Values) o archivos de texto con campos separados por un carácter específico, como una coma o un tabulador. Este componente es parte de la familia de componentes tFileInput, que se utiliza para leer datos desde diversas fuentes de archivos.</p>
	<p><b>tFileOutputDelimited:</b> Se utiliza para escribir datos en archivos de texto delimitados, como archivos CSV (Comma-Separated Values) u otros archivos de texto donde los campos están separados por un delimitador específico. Este componente es parte de la familia de componentes tFileOutput, que se utiliza para escribir datos en diversas fuentes de archivos.</p>
	<p><b>tPrejob:</b> Este no es un componente en sí mismo, sino una estructura especial dentro de un job que representa una sección de operaciones o configuraciones que se deben ejecutar antes de que el job principal comience. La idea del tPrejob es proporcionar una sección dedicada para</p>

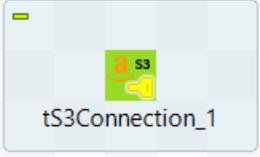
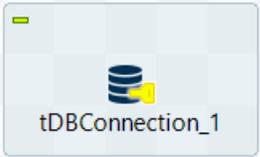
	la inicialización y preparación antes de ejecutar las operaciones principales del job.
	<p><b>tS3Connection:</b> Se utiliza para establecer una conexión con el servicio de almacenamiento en la nube Amazon Simple Storage Service (Amazon S3).</p> <p>es esencial cuando se necesita interactuar con el servicio de almacenamiento en la nube Amazon S3 en los flujos de trabajo de Talend. Facilita la gestión de la conexión y la autenticación, permitiendo acceder y manipular datos almacenados en Amazon S3 desde los Jobs de integración de datos.</p>
	<p><b>tDBConnection:</b> Se utiliza para establecer una conexión con una base de datos relacional. Este componente es fundamental en la mayoría de los jobs de Talend que involucran la extracción, transformación y carga (ETL) de datos entre sistemas de bases de datos. Este componente facilita la gestión de conexiones, mejorando la reutilización y la eficiencia en los flujos de trabajo de integración de datos.</p>
	<p><b>tS3Put:</b> Se utiliza para cargar archivos desde un sistema local o una fuente de datos a Amazon Simple Storage Service (Amazon S3). Además, facilita la carga de datos en la nube, lo que es especialmente útil en situaciones donde se necesita almacenar y procesar grandes volúmenes de datos de manera escalable y eficiente en Amazon S3.</p>
	<p><b>tDBClose:</b> Se utiliza para cerrar conexiones de bases de datos al finalizar las operaciones necesarias en un job de Talend. Este componente es esencial para asegurar una gestión adecuada de recursos y mantener la integridad de la conexión a lo largo del flujo de trabajo.</p>
	<p><b>tS3Close:</b> Al igual que tDBClose este componente se utiliza para cerrar conexiones de Amazon S3 al finalizar las operaciones necesarias en un job de Talend. Este componente es esencial para asegurar una gestión adecuada de recursos y mantener la integridad de la conexión a lo largo del flujo de trabajo.</p>

Tabla 6. Componentes en TOS

## MySQL

El primer componente en el modelo también representa la introducción del modelo DataWarehouse como base dentro de **MySQL**, para ello se crea la base de datos “**tiendadw**” el cual almacena los datos transformados ya en dimensiones. Este modelo ya está cargado con los datos para cada dimensión y por ende se ha cargado la fact.

## Segunda transformación de los datos

Se vuelve al segundo componente con la herramienta de **Talend Open Studio**, en este componente se cargan los datos en la estructura de carpetas que representa el modelo, dicha estructura es:

- **Raw:** contiene las tablas extraídas de la base original de Magento (transaccional), las cuales se utilizan para el modelo.
- **Staging:** contiene las tablas de la primera transformación de los datos, datos con los que aún se puede realizar pruebas para llegar al modelo de DataWarehouse ideal para la solución.
- **Presentation:** contiene las tablas que ya están listas para pasar a la representación de los datos y obtener las métricas que ayuden al negocio.

## AWS (S3 y Redshift)

Dentro del segundo componente además de Talend Open Studio, Se realiza una interacción entre dos herramientas en las que se trabaja en la nube mediante **Amazon Web Service (AWS)**.

La primera siendo **S3**, en ella se crea la estructura antes mencionada a la cual se cargan los datos. Para ello se crea primeramente un Bucket(espacios con los que trabaja S3 para almacenar una gran cantidad de información) en el que se crearan las carpetas para posteriormente ser cargadas desde Talend. El bucket creado tiene el nombre de “importaciones-el-angel-dw” y dentro de él se crean las carpetas Raw, Staging y Presentation, cada una se completa con las tablas en formato csv.

Para la carpeta Presentation los archivos que se suben desde Talend Open Studio (parte de este componente) son:

- dim\_fecha.csv
- dim\_producto.csv
- dim\_orden.csv
- dim\_cliente.csv
- dim\_tienda.csv
- fact\_venta.csv

Luego la segunda herramienta de la que se hace uso en AWS es **Redshift**, herramienta que sirve para el modelado de base de datos relacionales y no relacionales. Se crea un Clúster (el cual es un espacio

donde se puede almacenar información en forma de base de datos o tabular), para ello se crea el Clúster con el nombre de “p3-2023-cluster” y se configura para tener conexión con **S3**, luego se crea una estructura de base de datos que soporte el modelo dimensional, una vez creado se carga la base con las dimensiones que se encuentran en **S3**.

### **Power BI**

Se llega a lo que es el último componente front room, para ello se hará uso de **Power BI Desktop**, que es la herramienta de Power BI para trabajar de forma local los datos que se encuentran en la nube.

De esta manera se hace la conexión con RedShift, pasando todas las tablas de la base y se carga en Power BI Desktop, con esta carga ya se puede trabajar en la obtención de las métricas descritas anteriormente, para lo cual se crean informes que presenten las soluciones de estas de manera visual.

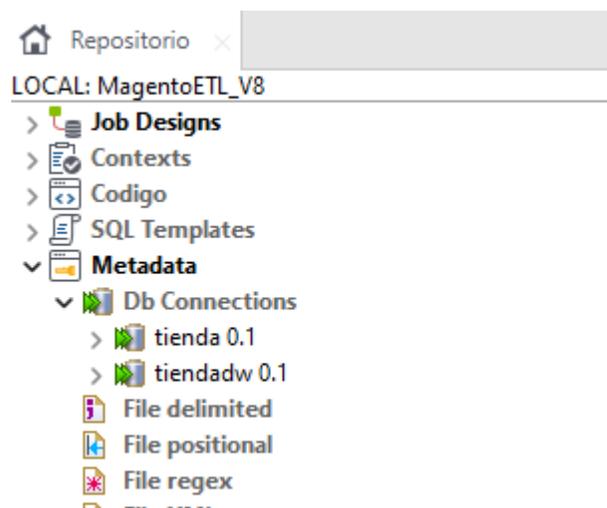
## CAPÍTULO III: ESTRATEGIA DE IMPLEMENTACIÓN DE PROPUESTA DE SOLUCIÓN

### a. Estrategia de implementación

Para satisfacer las necesidades especificadas a través de las métricas planteadas por Importaciones El Ángel y con el fin de convertir la información recolectada por su sistema transaccional de comercio electrónico en un activo de la empresa, se ha propuesto la integración de un sistema informático de Business Intelligence, que comprende de un DataWarehouse y un componente de procesamiento para la extracción de datos de la fuente y transformaciones de datos en información; un componente de almacenamiento en la nube, así como un elemento de visualización grafica de la información. Todo esto, diseñado para su integración con el sistema transaccional de Importaciones El Ángel, que opera bajo el software Magento Commerce.

### • Configuración de acceso a base de datos transaccional

Desde la herramienta de transformación y carga, TalendOpenStudio, es indispensable realizar la configuración para otorgar a la herramienta el acceso a la base de datos transaccional. Para establecer la conexión con la base de datos, se ha creado una especificación de acceso en la sección de “Metadata” dentro de TalendOpenStudio.



*Ilustración 9 Metadata, conexión a bases de datos*

Dentro del apartado de “Metadata”, dentro de “Db Connections”, se han creado los accesos a dos bases de datos importantes. La primera es “tienda”, que corresponde a la base de datos transaccional utilizada por Magento Commerce para persistir los datos de las operaciones de

Importaciones El Ángel”. La segunda es “tiendadw” que corresponde a la base de datos del DataWarehouse que será poblado con la información del modelo dimensional diseñado.

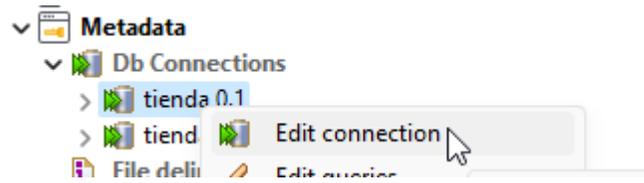


Ilustración 10. Editar conexión

Para modificar la información de la conexión a la base de datos, se hace clic derecho sobre el respectivo objeto de conexión y se da clic en la opción “Edit connection”, como se muestra en la imagen anterior. Esto abrirá la pantalla con la información del objeto de conexión. Al dar clic en el botón “Next >”, se accede a la configuración de la conexión, donde se deberá especificar las credenciales de conexión:

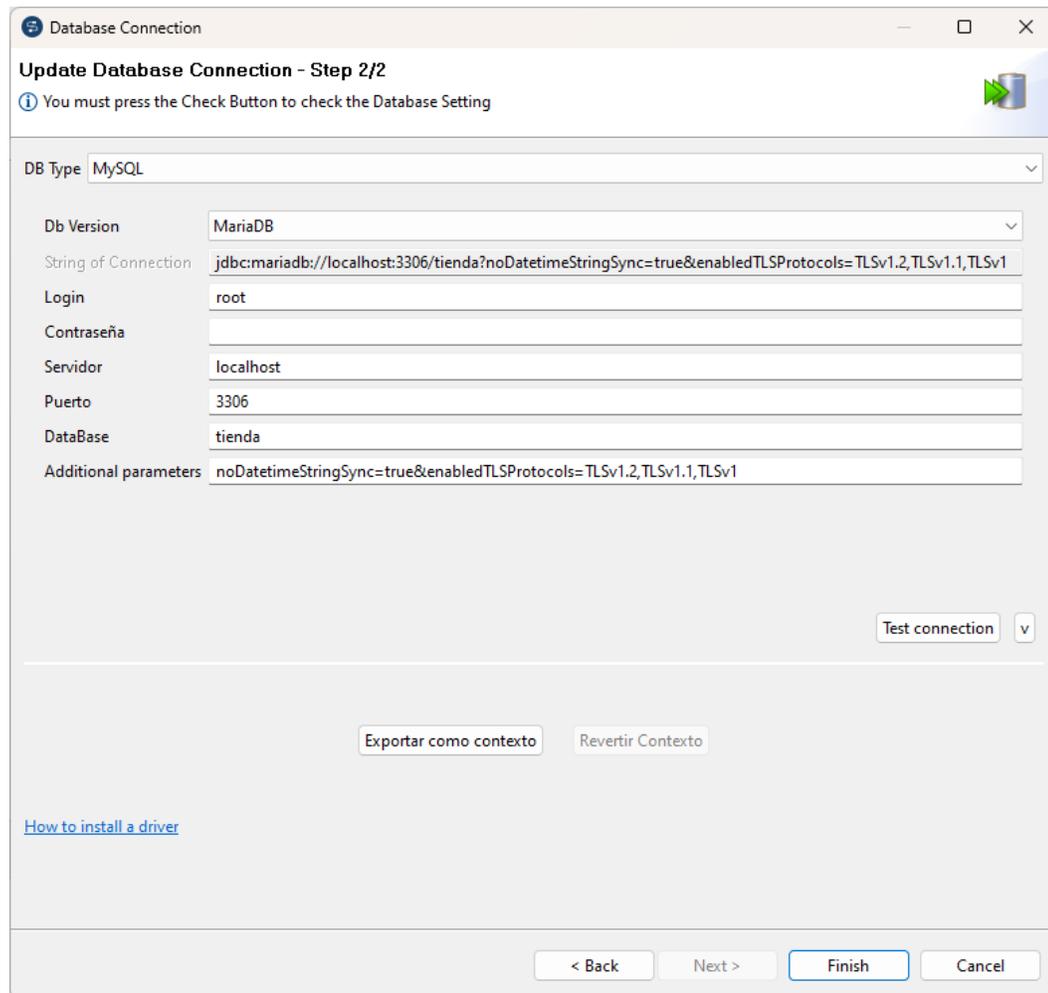
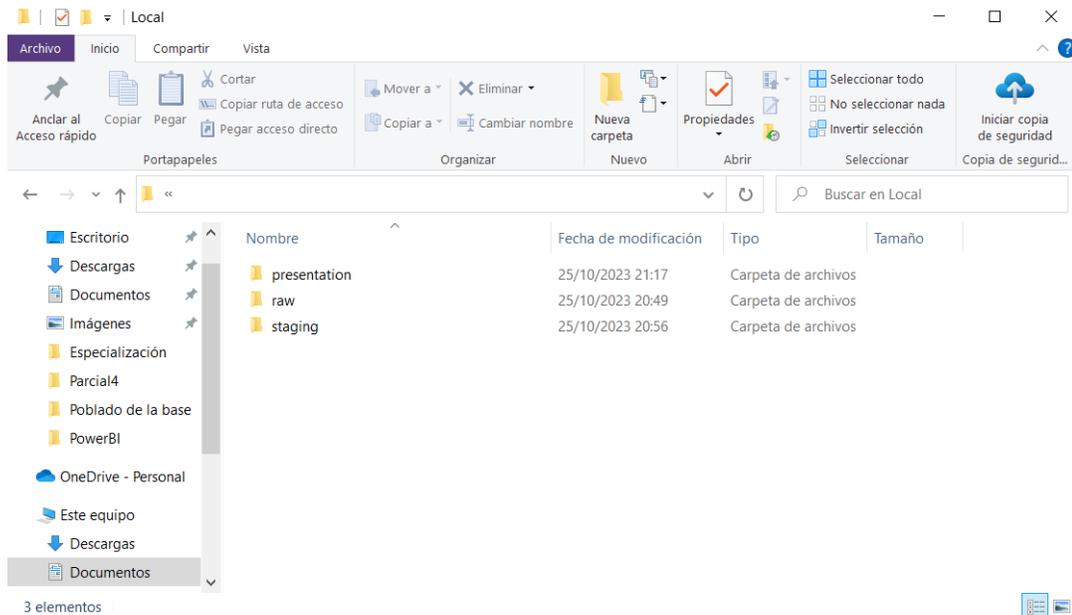


Ilustración 11. Configurar credenciales de acceso

De la misma forma, se deberá hacer la configuración para la conexión con el DataWarehouse una vez se haya creado.

- **Configuración local para estructura de capas de datos**

La solución requiere de una estructura de carpetas en capas, siguiendo el modelo de Ralph Kimball, el cual se recomienda establecerse un entorno local bajo el sistema operativo Windows 10.



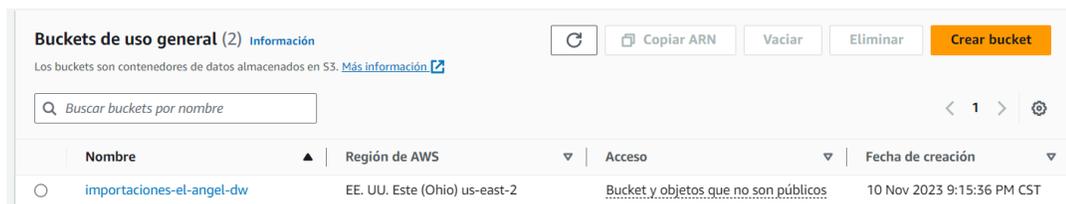
*Ilustración 12. Capas del modelo local*

Tal como se muestra en la imagen anterior, se crea una estructura de tres carpetas correspondientes a las 3 capas que sugiere la teoría sobre Business Intelligence.

- **Configuración de S3 para estructura de capas de datos**

Con una cuenta de Amazon, será necesario la utilización del servicio Amazon S3, para la persistencia en la nube del proceso de transformación del DataWarehouse.

Desde la barra de búsqueda, seleccionar el servicio de S3 dentro del cual se deberá crear un nuevo “Bucket” con el nombre "importaciones-el-angel-dw" en la región “US East (Ohio)”.



*Ilustración 13. Bucket de S3*

Una vez creado, al igual que con la estructura local de capas según Kimball, deben crearse 3 carpetas, una para cada capa con los nombres

- “01-raw”
- “02-staging”
- “03-presentation”

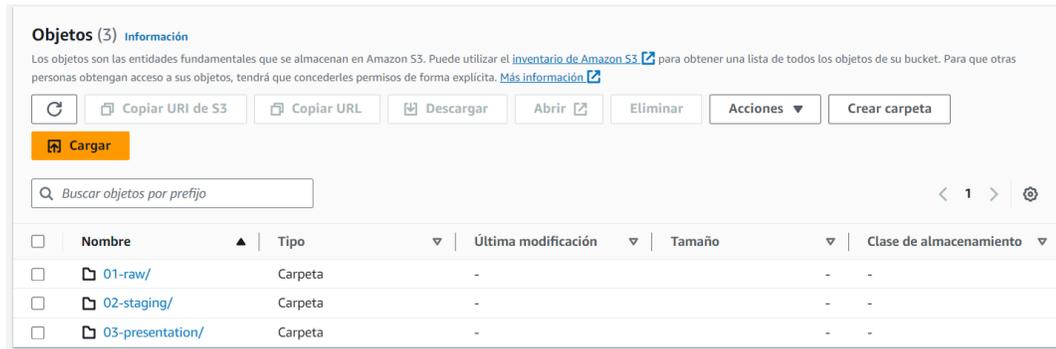


Ilustración 14. Capas del modelo en S3

- **Construcción de DataWarehouse Local**

La solución propuesta, requiere de una base de datos en la cual se almacena de manera local el DataWarehouse que resultará del proceso de extracción y transformación de datos. La base de datos vacía, debe crearse con el nombre “tiendadw”. Una vez creada la base de datos, la creación de las tablas puede realizarse de dos maneras diferentes:

- **Mediante la ejecución de scripts**

Se deben crear las tablas correspondientes a las dimensiones mediante la ejecución de los siguientes scripts de MariaDB:

Dimensión cliente

```
CREATE TABLE `dim_cliente` (  
  `cliente_key` int(11) NOT NULL AUTO_INCREMENT,  
  `cliente_id` int(10) NOT NULL,  
  `nombre` varchar(510) DEFAULT NULL,  
  `departamento` varchar(255) DEFAULT NULL,  
  `municipio` varchar(255) DEFAULT NULL,  
  `fecha_de_nacimiento` datetime DEFAULT NULL,  
  `genero` varchar(15) DEFAULT NULL,  
  `scd_start` date DEFAULT NULL,  
  `scd_end` date DEFAULT NULL,  
  `scd_active` int(11) DEFAULT NULL,
```

```

`previous_departamento` varchar(255) DEFAULT NULL,
`previous_municipio` varchar(255) DEFAULT NULL,
PRIMARY KEY (`cliente_key`)
) ENGINE=InnoDB AUTO_INCREMENT=102 DEFAULT CHARSET=utf8mb4;

```

#### Dimensión fecha

```

CREATE TABLE `dim_fecha` (
  `fecha_id` int(11) NOT NULL,
  `fecha_completa` date DEFAULT NULL,
  `anio` int(11) DEFAULT NULL,
  `trimestre` int(11) DEFAULT NULL,
  `mes` int(11) DEFAULT NULL,
  `nombre_mes` varchar(20) DEFAULT NULL,
  `dia` int(11) DEFAULT NULL,
  `dia_semana` int(11) DEFAULT NULL,
  `nombre_dia` varchar(20) DEFAULT NULL,
  PRIMARY KEY (`fecha_id`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4;

```

#### Dimensión orden

```

CREATE TABLE `dim_orden` (
  `orden_key` int(11) NOT NULL AUTO_INCREMENT,
  `orden_id` int(10) NOT NULL,
  `estado` varchar(32) DEFAULT NULL,
  `metodo` varchar(128) DEFAULT NULL,
  `gran_total` float(12,4) DEFAULT NULL,
  `departamento` varchar(255) DEFAULT NULL,
  `municipio` varchar(255) DEFAULT NULL,
  `scd_start` datetime DEFAULT NULL,
  `scd_end` datetime DEFAULT NULL,
  `scd_active` int(11) DEFAULT NULL,
  `previous_metodo` varchar(128) DEFAULT NULL,
  `previous_departamento` varchar(255) DEFAULT NULL,
  `previous_municipio` varchar(255) DEFAULT NULL,
  PRIMARY KEY (`orden_key`)

```

```
) ENGINE=InnoDB AUTO_INCREMENT=1503 DEFAULT CHARSET=utf8mb4;
```

#### Dimensión producto

```
CREATE TABLE `dim_producto` (  
  `producto_key` int(11) NOT NULL AUTO_INCREMENT,  
  `producto_id` int(10) NOT NULL,  
  `sku` varchar(64) DEFAULT NULL,  
  `nombre` varchar(510) DEFAULT NULL,  
  `descripcion` text DEFAULT NULL,  
  `scd_start` datetime DEFAULT NULL,  
  `scd_end` datetime DEFAULT NULL,  
  `scd_active` int(11) DEFAULT NULL,  
  PRIMARY KEY (`producto_key`)  
) ENGINE=InnoDB AUTO_INCREMENT=2047 DEFAULT CHARSET=utf8mb4;
```

#### Dimensión tienda

```
CREATE TABLE `dim_tienda` (  
  `tienda_key` int(11) NOT NULL AUTO_INCREMENT,  
  `tienda_id` int(10) NOT NULL,  
  `codigo` varchar(32) DEFAULT NULL,  
  `nombre` varchar(510) DEFAULT NULL,  
  PRIMARY KEY (`tienda_key`)  
) ENGINE=InnoDB AUTO_INCREMENT=14 DEFAULT CHARSET=utf8mb4;
```

Al finalizar de ejecutar estos scripts, se debe ejecutar el siguiente, el cual corresponde a la tabla de hechos:

```
CREATE TABLE `fact_venta` (  
  `venta_key` int(11) NOT NULL AUTO_INCREMENT,  
  `venta_id` int(10) NOT NULL,  
  `estado` varchar(32) DEFAULT NULL,  
  `cantidad_ordenada` double(12,4) DEFAULT NULL,  
  `precio_producto` double(12,4) DEFAULT NULL,  
  `gran_total` double(12,4) DEFAULT NULL,  
  `monto_base_descuento` double(12,4) DEFAULT NULL,
```

```

`base_linea_total` double(12,4) DEFAULT NULL,
`monto_impuesto_base` double(12,4) DEFAULT NULL,
`precio_base_incluyendo_impuestos` double(12,4) DEFAULT NULL,
`total_base_linea_incluyendo_impuestos` double(12,4) DEFAULT
NULL,
`cliente_key` int(10) DEFAULT NULL,
`tienda_key` int(10) DEFAULT NULL,
`orden_key` int(10) DEFAULT NULL,
`producto_key` int(10) DEFAULT NULL,
`fecha_de_creacion` int(10) DEFAULT NULL,
`scd_start` datetime DEFAULT NULL,
`scd_end` datetime DEFAULT NULL,
`scd_active` int(11) DEFAULT NULL,
`previous_cantidad_ordenada` double(12,4) DEFAULT NULL,
PRIMARY KEY (`venta_key`),
KEY `fk_fact_dim_tienda` (`tienda_key`),
KEY `fk_fact_dim_cliente` (`cliente_key`),
KEY `fk_fact_dim_orden` (`orden_key`),
KEY `fk_fact_dim_producto` (`producto_key`),
KEY `fk_fact_dim_fecha` (`fecha_de_creacion`),
CONSTRAINT `fk_fact_dim_cliente` FOREIGN KEY (`cliente_key`)
REFERENCES `dim_cliente` (`cliente_key`),
CONSTRAINT `fk_fact_dim_fecha` FOREIGN KEY
(`fecha_de_creacion`) REFERENCES `dim_fecha` (`fecha_id`),
CONSTRAINT `fk_fact_dim_orden` FOREIGN KEY (`orden_key`)
REFERENCES `dim_orden` (`orden_key`),
CONSTRAINT `fk_fact_dim_producto` FOREIGN KEY (`producto_key`)
REFERENCES `dim_producto` (`producto_key`),
CONSTRAINT `fk_fact_dim_tienda` FOREIGN KEY (`tienda_key`)
REFERENCES `dim_tienda` (`tienda_key`)
) ENGINE=InnoDB AUTO_INCREMENT=4503 DEFAULT CHARSET=utf8mb4;

```

Script para poblar la dimension fecha

```

INSERT INTO dim_fecha (fecha_id, fecha_completa, anio,
trimestre, mes, nombre_mes, dia, dia_semana, nombre_dia)

```

```

SELECT
DATE_FORMAT(fecha_secuencia, '%Y%m%d') AS fecha_id,
fecha_secuencia AS fecha_completa,
YEAR(fecha_secuencia) AS anio,
QUARTER(fecha_secuencia) AS trimestre,
MONTH(fecha_secuencia) AS mes,
MONTHNAME(fecha_secuencia) AS nombre_mes,
DAY(fecha_secuencia) AS dia,
DAYOFWEEK(fecha_secuencia) AS dia_semana,
DAYNAME(fecha_secuencia) AS nombre_dia
FROM (
SELECT DATE_ADD('2020-01-01', INTERVAL n DAY) AS fecha_secuencia
-- AQUÍ
FROM (
SELECT a.N + b.N * 10 + c.N * 100 + d.N * 1000 AS n
FROM (
SELECT 0 AS N UNION SELECT 1 UNION SELECT 2 UNION SELECT 3
UNION SELECT 4 UNION
SELECT 5 UNION SELECT 6 UNION SELECT 7 UNION SELECT 8 UNION
SELECT 9
) AS a
CROSS JOIN (
SELECT 0 AS N UNION SELECT 1 UNION SELECT 2 UNION SELECT 3
UNION SELECT 4 UNION
SELECT 5 UNION SELECT 6 UNION SELECT 7 UNION SELECT 8 UNION
SELECT 9
) AS b
CROSS JOIN (
SELECT 0 AS N UNION SELECT 1 UNION SELECT 2 UNION SELECT 3
UNION SELECT 4 UNION
SELECT 5 UNION SELECT 6 UNION SELECT 7 UNION SELECT 8 UNION
SELECT 9
) AS c
CROSS JOIN (

```

```

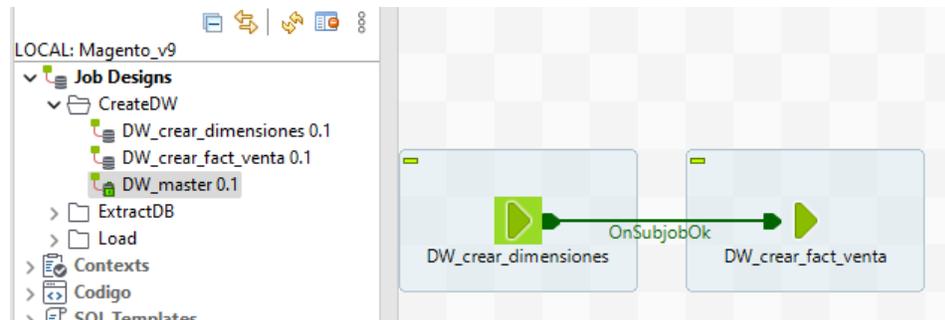
        SELECT 0 AS N UNION SELECT 1 UNION SELECT 2 UNION SELECT 3
UNION SELECT 4 UNION
        SELECT 5 UNION SELECT 6 UNION SELECT 7 UNION SELECT 8 UNION
SELECT 9
    ) AS d
) AS numeros
WHERE DATE_ADD('2020-01-01', INTERVAL n DAY) <= '2023-12-31' --
AQUI 2
) AS secuencia_fechas;
-- Actualizar la columna nombre_mes a español
UPDATE dim_fecha
SET nombre_mes =
CASE nombre_mes
WHEN 'January' THEN 'Enero'
WHEN 'February' THEN 'Febrero'
WHEN 'March' THEN 'Marzo'
WHEN 'April' THEN 'Abril'
WHEN 'May' THEN 'Mayo'
WHEN 'June' THEN 'Junio'
WHEN 'July' THEN 'Julio'
WHEN 'August' THEN 'Agosto'
WHEN 'September' THEN 'Septiembre'
WHEN 'October' THEN 'Octubre'
WHEN 'November' THEN 'Noviembre'
WHEN 'December' THEN 'Diciembre'
ELSE nombre_mes -- Mantener cualquier otro valor tal como está
END;
-- Actualizar la columna nombre_dia a español
UPDATE dim_fecha
SET nombre_dia =
CASE nombre_dia
WHEN 'Monday' THEN 'Lunes'
WHEN 'Tuesday' THEN 'Martes'
WHEN 'Wednesday' THEN 'Miércoles'
WHEN 'Thursday' THEN 'Jueves'

```

```
WHEN 'Friday' THEN 'Viernes'  
WHEN 'Saturday' THEN 'Sábado'  
WHEN 'Sunday' THEN 'Domingo'  
ELSE nombre_dia -- Mantener cualquier otro valor tal como está  
END;
```

- **Mediante Jobs de Talend**

Una forma automatizada para la creación de la estructura del DataWarehouse es utilizando la herramienta Talend Open Studio. Se ha incluido un conjunto de Jobs que se encargan de la ejecución de los scripts necesarios para la creación de las tablas correspondientes a las dimensiones y tabla de hechos del DataWarehouse, además de poblar la tabla dimensional “dim\_date”.



*Ilustración 15. Job master crear DW*

Para ejecutar este Job, es necesario tener una base de datos con el nombre “tiendadw” y haber realizado la configuración de conexión a esta base de datos, desde el apartado de “Metadata” dentro de Talend Open Studio. Una vez realizado, al ejecutar el Job “DW\_master”, se creará la estructura de tablas del modelo dimensional listo para el proceso de transformación.

- **Ejecución de Jobs de transformación en Talend Open Studio**

Ahora, con todos los elementos anteriormente mencionados, se debe proceder a la ejecución de los Jobs que han sido diseñados para el proceso de transformación. En primer lugar, se deben ejecutar los Jobs de la carpeta “ExtractDB”. Cada Job realiza la extracción de los datos y las transformaciones necesarias para construir una tabla del modelo dimensional. Hay un Job para cada dimensión, así como un job para la tabla de hechos:

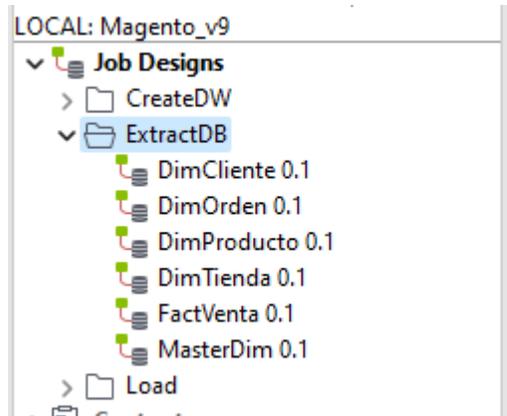


Ilustración 16. Jobs de extracción y transformación

Para la ejecución correcta y ordenada de todos los Jobs en esta carpeta, se cuenta con el Job “MasterDim”, el cual se encarga de llamar a cada Job en el orden correcto para evitar errores debido a conflictos referenciales. Gracias a esto, solo es necesario ejecutar el Job “MasterDim” para iniciar el volcado de datos transformados dentro del DataWarehouse local.

Al finalizar la ejecución del Job anterior, se tendrá el DataWarehouse local listo para su uso. Sin embargo, la solución también requiere de una manera de persistir el DataWarehouse en la nube. Esto se realiza haciendo una carga de los datos que se encuentran en las dimensiones y la tabla de hechos, extraerlos en archivos .CSV dentro de la carpeta Staging local y posteriormente, realizar la carga de datos al bucket en S3. Este proceso se realiza ejecutando los Jobs que se encuentran en la carpeta

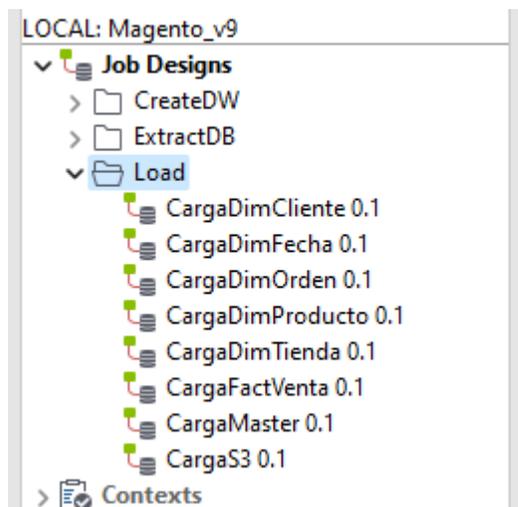


Ilustración 17. Jobs de carga

Se ha construido un Job por cada dimensión a ser extraída, un job para extraer la tabla de hechos y un job para realizar la carga hacia el bucket de S3. Al igual que los Jobs de transformación, se ha construido un Job para la ejecución secuencial de estos Jobs. Así,

solamente es necesario ejecutar el Job “CargaMaster” para persistir los archivos .csv del modelo dimensional diseñado en el bucket de S3.

- **Construcción de estructura de DataWarehouse en Amazon Redshift**

Mediante una cuenta de Amazon, se hará uso del servicio de Amazon Redshift. Es necesaria la creación de un clúster para albergar el modelo dimensional, así como la creación de un rol IAM para gestionar la seguridad y acceso al clúster.

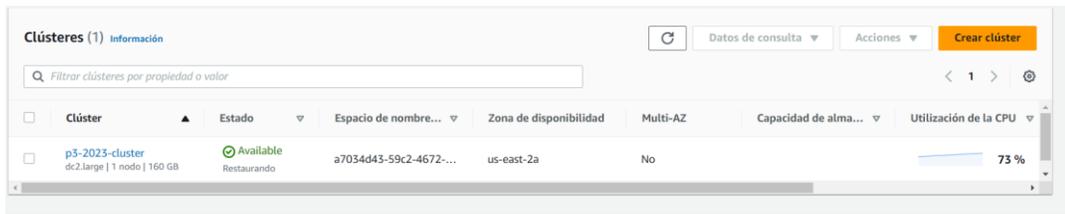


Ilustración 18. Clúster Amazon Redshift

Una vez creado el clúster, es necesario iniciarlo (tomar en cuenta que el cobro de este servicio se realiza solo cuando el clúster está encendido). Se ingresa dando clic en el nombre del clúster en la lista de clústeres. Una vez dentro, se debe buscar el botón “Datos de consulta” y seleccionar la opción “Consultas en el editor de consultas v2”.

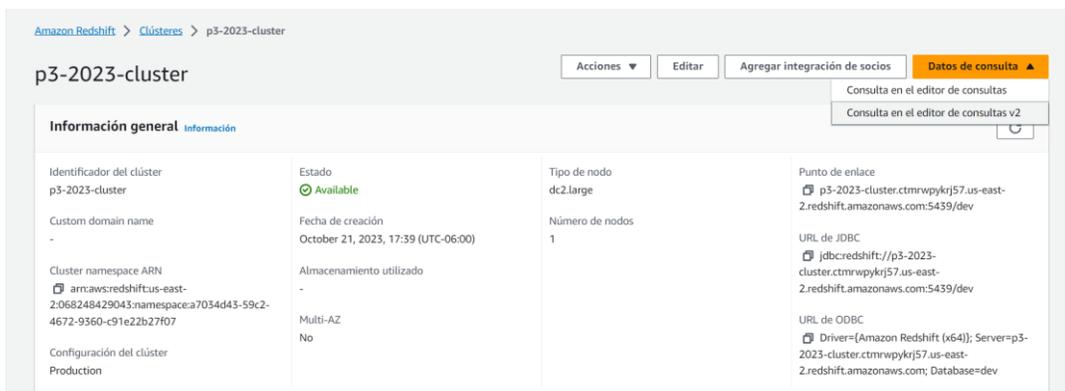


Ilustración 19. Información de clúster en Amazon Redshift

Esto abrirá el editor de consultas, desde el cual se debe proceder a la creación de la estructura del modelo dimensional, tal y como se realizó con el DataWarehouse local en MaríaDB. Dentro del editor de consultas, se debe crear dentro del clúster una base de datos en la cual albergar el modelo dimensional. Dentro de esta base, deberán ejecutarse los scripts siguientes:

```
CREATE TABLE dim_cliente(  
    cliente_key INTEGER NOT NULL IDENTITY(1,1),  
    cliente_id INTEGER,
```

```

nombre VARCHAR(510),
departamento VARCHAR(255),
municipio VARCHAR(255),
fecha_de_nacimiento DATETIME,
genero VARCHAR(15),
scd_start DATETIME,
scd_end DATETIME,
scd_active INTEGER,
previous_departamento VARCHAR(255),
previous_municipio VARCHAR(255),
PRIMARY KEY (cliente_key)
)DISTSTYLE AUTO;

```

```

CREATE TABLE dim_orden(
orden_key INTEGER NOT NULL IDENTITY(1,1),
orden_id INTEGER,
estado VARCHAR(32),
metodo VARCHAR(128),
gran_total NUMERIC(12,4),
departamento VARCHAR(255),
municipio VARCHAR(255),
scd_start DATETIME,
scd_end DATETIME,
scd_active INTEGER,
previous_departamento VARCHAR(255),
previous_municipio VARCHAR(255),
PRIMARY KEY (orden_key)
)DISTSTYLE AUTO;

```

```

CREATE TABLE dim_tienda(
tienda_key INTEGER NOT NULL IDENTITY(1,1),
tienda_id INTEGER,
codigo VARCHAR(32),
nombre VARCHAR(510),
PRIMARY KEY (tienda_key)

```

```
)DISTSTYLE AUTO;
```

```
CREATE TABLE dim_fecha(  
    fecha_id INTEGER NOT NULL IDENTITY(1,1),  
    fecha_completa DATETIME,  
    anio INTEGER,  
    trimestre INTEGER,  
    mes INTEGER,  
    nombre_mes VARCHAR(20),  
    dia INTEGER,  
    dia_semana INTEGER,  
    nombre_dia VARCHAR(20),  
    PRIMARY KEY (fecha_id)  
)DISTSTYLE AUTO;
```

Una vez creadas las tablas de dimensiones, se procede a la creación de la tabla de hechos, así como sus relaciones a través de la especificación de llaves foráneas.

Con la estructura de tablas creada, se debe proceder a la ejecución de los archivos de volcado. Estos scripts se encargan de indicar a Redshift el bucket y el nombre del archivo de datos que debe buscar; características para la lectura del archivo fuente, como el carácter delimitador, formatos de fechas, etc.; y la tabla destino.

```
COPY "importadora-el-angel".public.dim_cliente FROM  
's3://importaciones-el-angel-dw/03-presentation/dim_cliente.csv'  
IAM_ROLE 'arn:aws:iam::068248429043:role/LecturaS3' FORMAT AS CSV  
DELIMITER ';' QUOTE '' IGNOREHEADER 1 DATEFORMAT 'dd-MM-yyyy' REGION  
AS 'us-east-2'
```

```
COPY "importadora-el-angel".public.dim_orden FROM 's3://importaciones-  
el-angel-dw/03-presentation/dim_orden.csv' IAM_ROLE  
'arn:aws:iam::068248429043:role/LecturaS3' FORMAT AS CSV DELIMITER ';' QUOTE ''  
IGNOREHEADER 1 DATEFORMAT 'dd-MM-yyyy' REGION AS 'us-east-2'
```

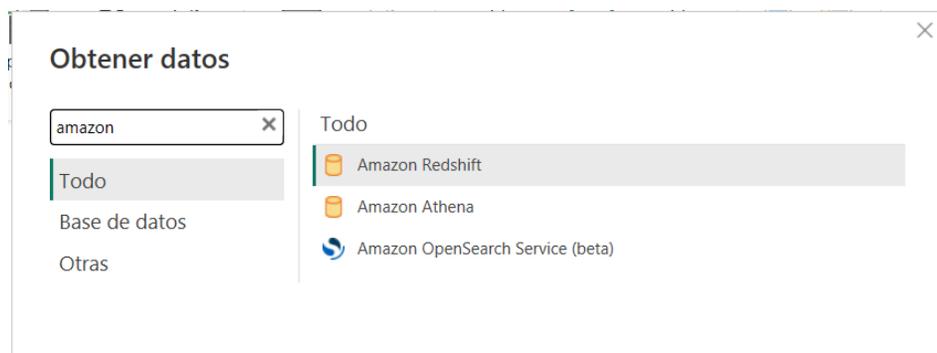
```
COPY "importadora-el-angel".public.dim_fecha FROM 's3://importaciones-
el-angel-dw/03-presentation/dim_fecha.csv' IAM_ROLE
'arn:aws:iam::068248429043:role/LecturaS3' FORMAT AS CSV DELIMITER ';'
QUOTE '"' IGNOREHEADER 1 DATEFORMAT 'dd-MM-yyyy' REGION AS 'us-east-2'
```

```
COPY "importadora-el-angel".public.dim_producto FROM
's3://importaciones-el-angel-dw/03-presentation/dim_producto.csv'
IAM_ROLE 'arn:aws:iam::068248429043:role/LecturaS3' FORMAT AS CSV
DELIMITER ';' QUOTE '"' IGNOREHEADER 1 DATEFORMAT 'dd-MM-yyyy' REGION
AS 'us-east-2'
```

```
COPY "importadora-el-angel".public.dim_tienda FROM
's3://importaciones-el-angel-dw/03-presentation/dim_tienda.csv'
IAM_ROLE 'arn:aws:iam::068248429043:role/LecturaS3' FORMAT AS CSV
DELIMITER ';' QUOTE '"' IGNOREHEADER 1 REGION AS 'us-east-2'
```

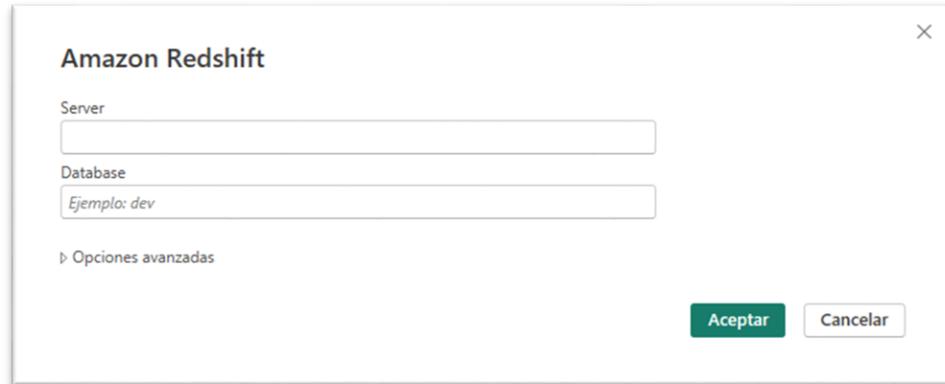
```
COPY "importadora-el-angel".public.fact_ventas FROM
's3://importaciones-el-angel-dw/03-presentation/fact_ventas.csv'
IAM_ROLE 'arn:aws:iam::068248429043:role/LecturaS3' FORMAT AS CSV
DELIMITER ';' QUOTE '"' IGNOREHEADER 1 DATEFORMAT 'dd-MM-yyyy' REGION
AS 'us-east-2'
```

- Configuración de acceso a base de datos Redshift desde Power BI  
Dentro de la herramienta de Power BI, se debe seleccionar el origen de datos para Amazon Redshift.



*Ilustración 20. Selección de origen de datos en PowerBI*

Al acceder se solicitará el servidor del clúster y el nombre de la base de datos a la que se desea tener acceso. Esta información se obtiene al ingresar a Amazon Redshift y hacer clic sobre el nombre del clúster, en la subsección “Punto de enlace”.

A screenshot of a dialog box titled "Amazon Redshift". It contains two input fields: "Server" and "Database". The "Database" field has the placeholder text "Ejemplo: dev". Below the fields is a link "Opciones avanzadas" with a right-pointing arrow. At the bottom right are two buttons: "Aceptar" (green) and "Cancelar" (white).

*Ilustración 21. Especificación de servidor y base de datos para la conexión con Amazon Redshift*

Una vez establecido el servidor y base de datos a conectarse, se solicitará el usuario y contraseña especificados al momento de crear el clúster.

A screenshot of a dialog box titled "AmazonRedshift". It shows a list on the left with "Amazon Redshift" and "Cuenta de Microsoft". The main area displays a cluster name "p3-2023-cluster.ctmrwpykrj57.us-east-2.redshift.am..." and two input fields: "Nombre de usuario" with the value "awsuser-p3" and "Contraseña" with masked characters. At the bottom right are two buttons: "Guardar" (green) and "Cancelar" (white).

*Ilustración 22. Especificación de credenciales de usuario.*

Al guardar, se iniciará la conexión con la base de datos del modelo dimensional que se cargó a Redshift:

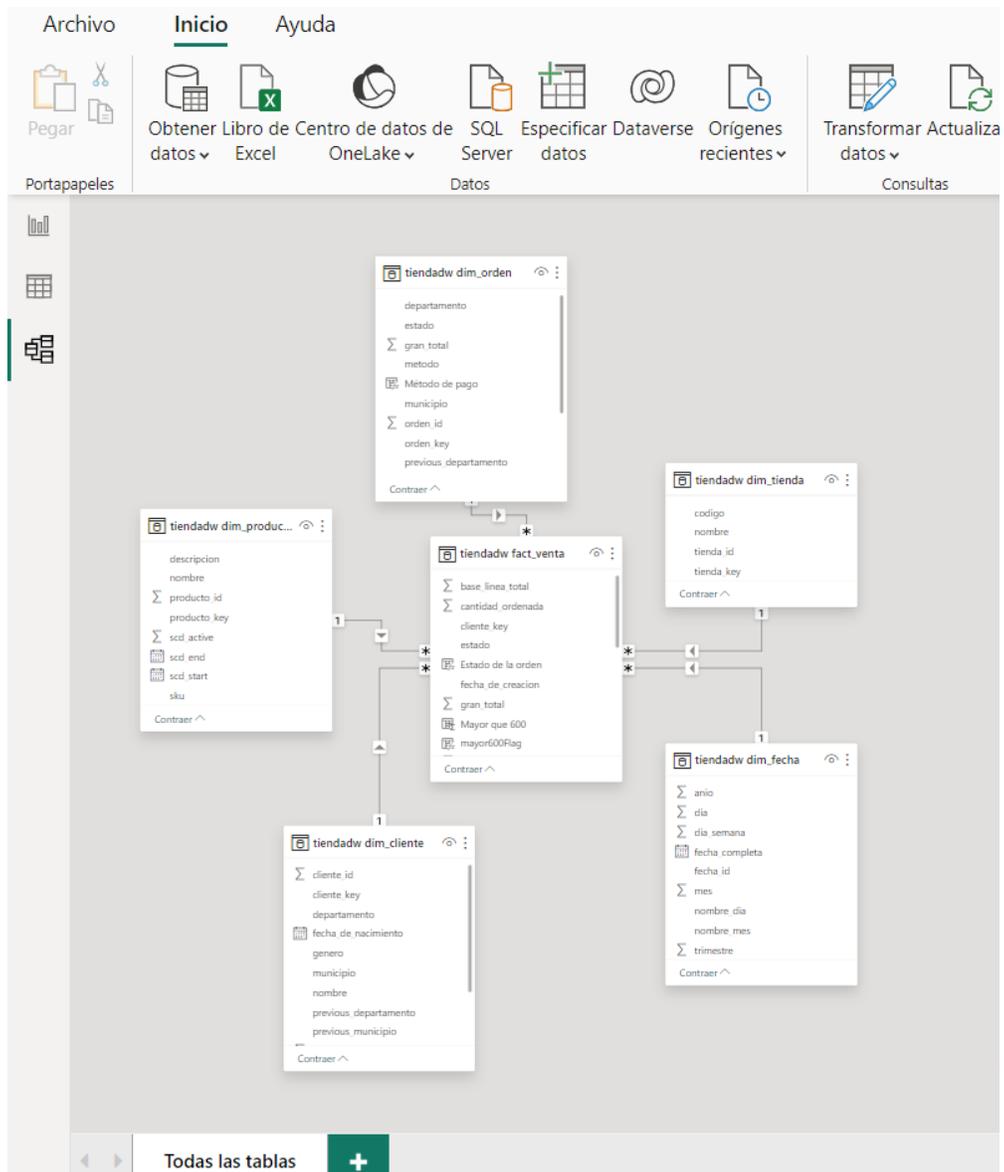


Ilustración 23. Modelo de estrella cargado en PowerBI

## b. Presupuesto de implementación

- **Recurso humano**

<b>Recurso humano especializado</b>				
Cargo	Salario por hora (USD)	Horas	Cantidad de recurso	Costo sub total (USD)
Ing. de software	\$7.00	24	1	\$168.00
Ing. de datos	\$9.00	16	1	\$144.00
Capacitador	\$6.00	16	1	\$96.00
<b>Costo total</b>				<b>\$408.00</b>

*Tabla 7. Costo de implementación para RRHH*

$$\text{Costo Total}_{RRHH_T} = (\text{Sueldo por hora en USD}) * (\text{Horas}) * (\text{Recurso})$$

$$\text{Costo SubTotal1}_{RRHH_T} = 7 * 24 * 1 = 168 \text{ USD}$$

$$\text{Costo SubTotal2}_{RRHH_T} = 9 * 16 * 1 = 144 \text{ USD}$$

$$\text{Costo SubTotal3}_{RRHH_T} = 6 * 16 * 1 = 96 \text{ USD}$$

$$\therefore \text{Costo Total}_{RRHH_T} = \sum_{n=3}^{n=1} \text{Costo SubTotal}_n$$

$$\therefore \text{Costo Total}_{RRHH_T} = 408 \text{ USD}$$

- **Hardware y software**

El software utilizado para la implementación del proyecto es de código abierto, por tanto, no se incurre en gastos de licencias por el uso de estos.

- **Servicios de terceros**

<b>Servicios de terceros</b>			
Servicio	Tarifa por mes	Meses	Costo sub total (USD)
Amazon Web Services	\$3,504.00	1	\$3,504.00
S3 Bucket	\$0.60	1	\$0.60
<b>Costo total</b>			<b>\$3,504.60</b>

*Tabla 8. Costos de servicios de terceros*

Para el cálculo de los servicios a terceros se hizo uso de la calculadora que provee AWS indicada en la bibliografía.

### c. Análisis de resultados

- **Conexión a la base transaccional**

Una vez establecida la conexión con la base de datos transaccional, se logró extraer los esquemas de datos necesarios para la extracción de datos y posterior transformación:

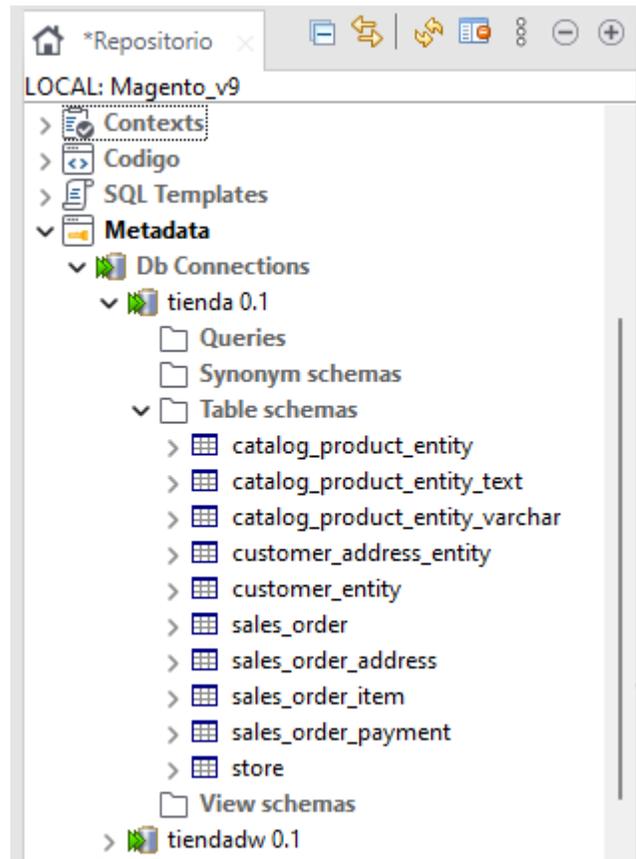


Ilustración 24. Esquemas extraídos del modelo transaccional

- **Disposición final de Jobs de transformación**

## JOBS DE TRANSFORMACIÓN

### DimCliente 0.1:

Tablas de entrada:

- customer\_entity
- customer\_address\_entity

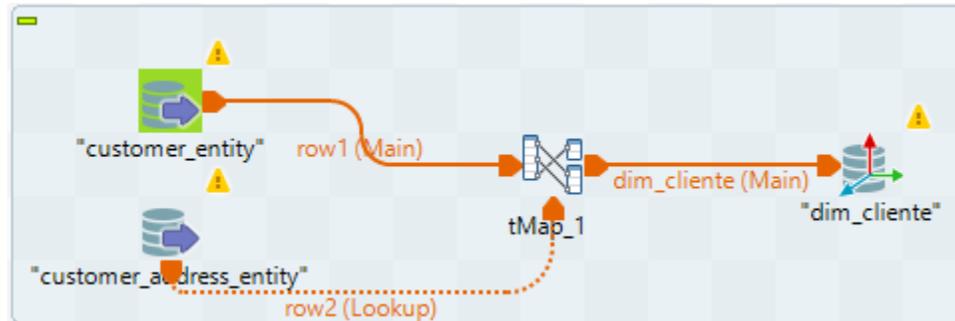


Ilustración 25. Job DimCliente

### DimOrden 0.1:

Tablas de entrada:

- sales\_order
- sales\_order\_address
- sales\_order\_payment

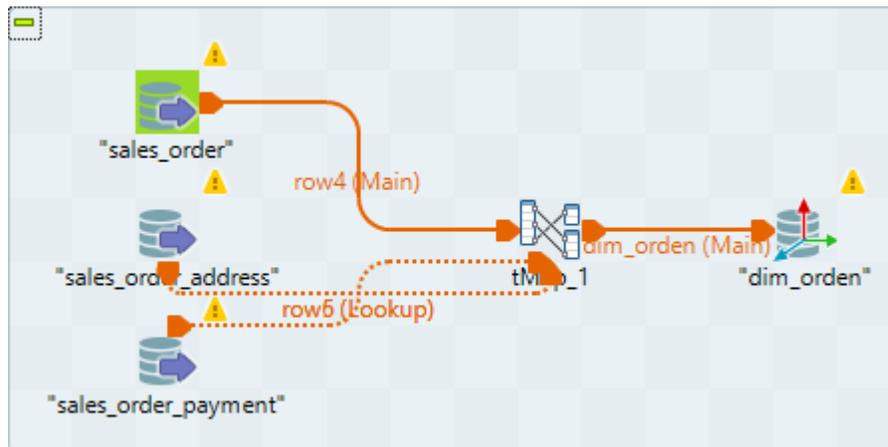


Ilustración 26. DimOrden

### DimProducto 0.1:

Tablas de entrada:

- catalog\_product\_entity

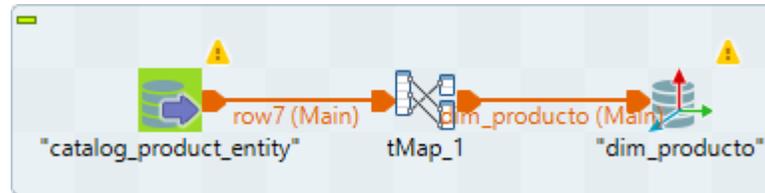


Ilustración 27. DimProducto

### DimTienda 0.1:

Tablas de entrada:

- store

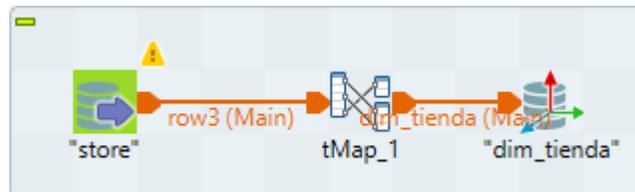


Ilustración 28. DimTienda

### FactVenta 0.1:

Tablas de entrada:

- dim\_cliente
- dim\_orden
- sales\_order\_item
- dim\_producto
- dim\_tienda

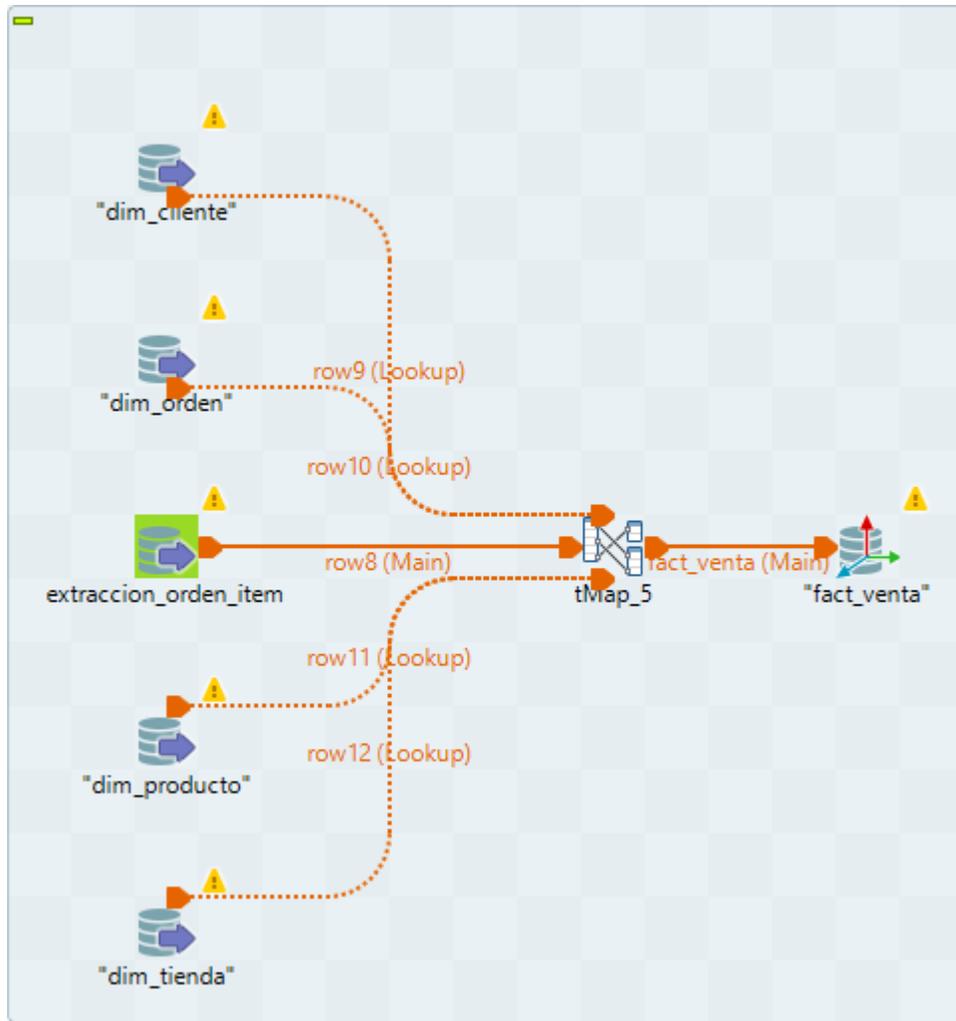


Ilustración 29. FactVentas

## JOBS DE CARGA DE DATOS

dim\_cliente a .CSV

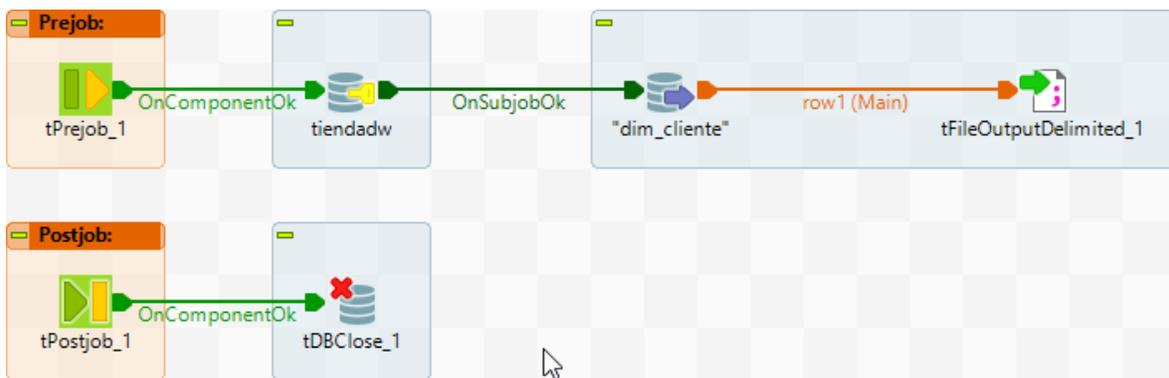


Ilustración 30. Job exportar cliente a .csv

dim\_fecha a .CSV

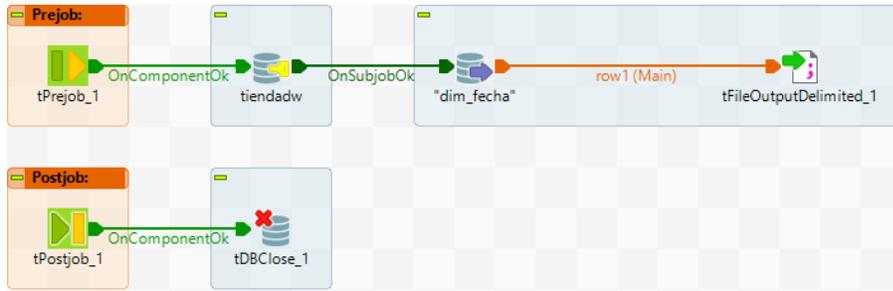


Ilustración 31. Job exportar fecha a .csv

dim\_orden a .CSV

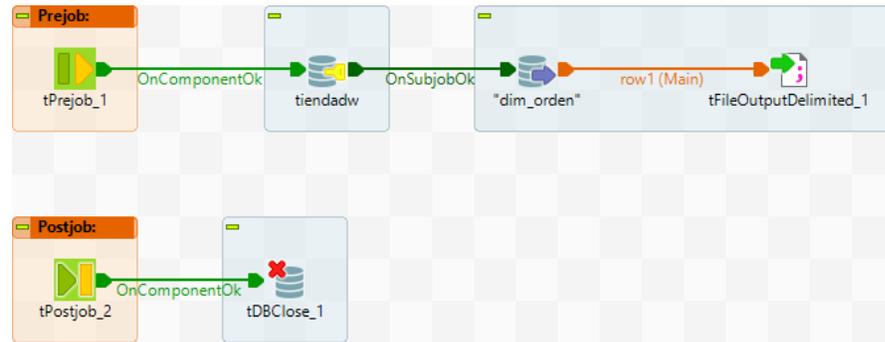


Ilustración 32. Job exportar orden a .csv

dim\_producto a .CSV

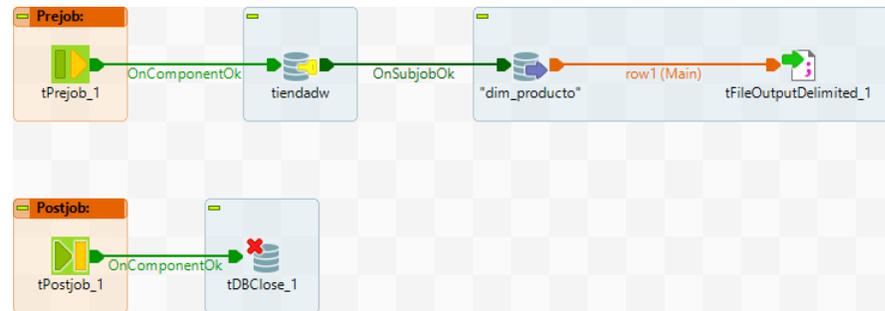


Ilustración 33. Job exportar producto a .csv

dim\_tienda a .CSV

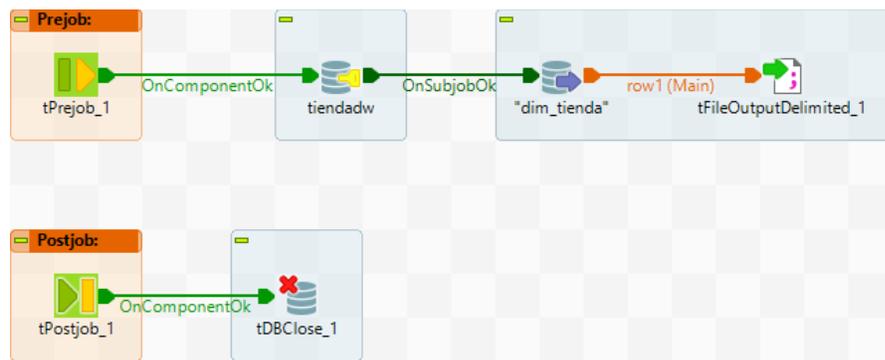


Ilustración 34. Job exportar tienda a .csv

### fact\_venta a .CSV

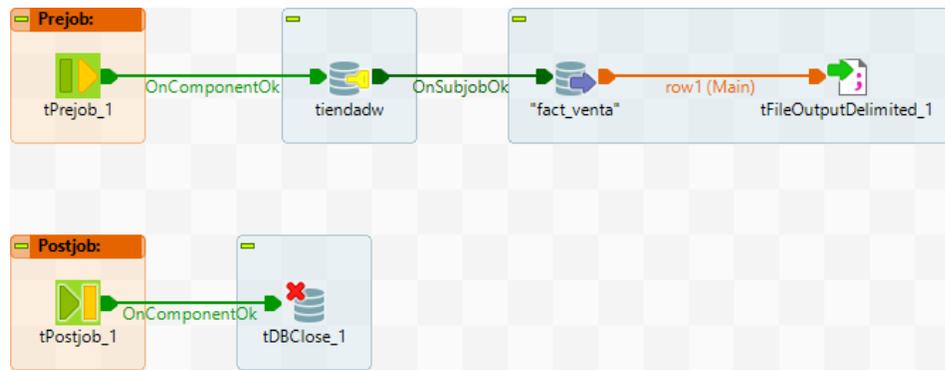


Ilustración 35. Job exportar venta a .csv

### Carga de Modelo a S3

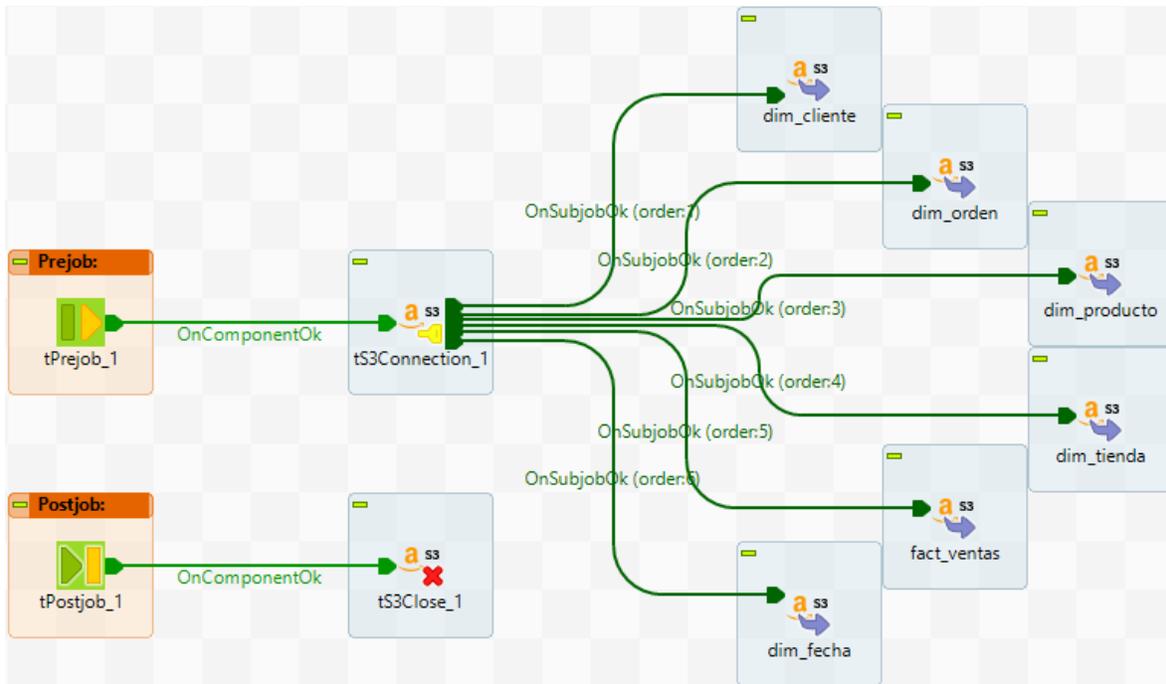


Ilustración 36. Job de carga a Amazon S3

- Archivos .csv resultado del proceso de transformación
- Una vez ejecutado los Jobs de extracción y transformación de datos; y posteriormente los Jobs previos a la carga a S3; se generan los archivos de datos en formato .csv dentro de la capa Staging local.

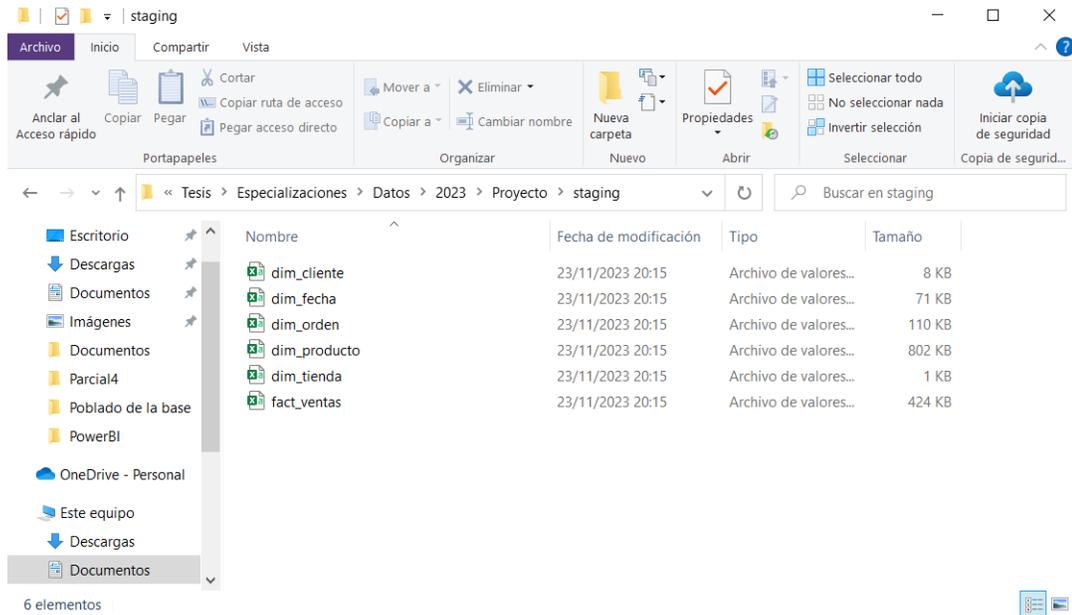


Ilustración 37. Capa de datos staging local

- **Archivos en Amazon S3**

Al ejecutar los Jobs de carga a S3, los archivos correspondientes a cada una de las dimensiones del modelo de estrella, serán resguardados en el correspondiente bucket de S3

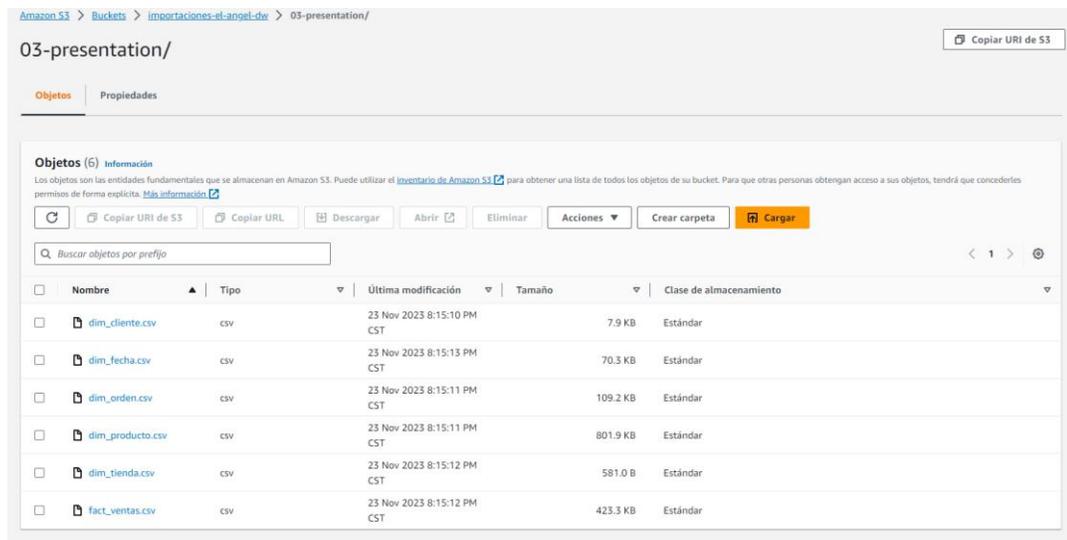


Ilustración 38. Capa de datos presentation Amazon S3

- **Base de datos volcada en Amazon Redshift**

Al haber ejecutado los scripts de volcado de datos desde el editor de consultas de Amazon Redshift, los datos son traídos desde los archivos en la capa de presentación del Bucket de S3 y cargados en sus respectivas tablas del modelo dimensional:

The screenshot shows the Amazon Redshift console interface. At the top, there are controls for running a query, including a 'Run' button, 'Limit 100', 'Explain', 'Isolated session', and a session identifier 'p3-2023-cluster'. The SQL query being executed is:

```

1 SELECT
2 *
3 FROM
4 "importadora-el-angel"."public"."dim_cliente";

```

Below the query, the results are displayed in a table with 101 rows. The first 10 rows are visible:

cliente_key	cliente_id	nombre	departamento	municipio	fecha_d
1	1	Veronica Costello	La Unión	Polorós	1973-12
2	2	Kevin Arevalo	Ahuachapán	Guatajagua	2023-11
3	3	Karen Bolaños	La Paz	Suchitoto	2023-11
4	4	Katherine Ventura	Morazán	Chirilagua	2023-11
5	5	Fernando Lopez	San Vicente	San Carlos	2023-11
6	6	Mónica Guzman	Chalatenango	San Salvador	NULL
7	7	Diana Martínez	Cuscatlán	San Jorge	NULL
8	8	Javier	Usulután	Atiquizaya	NULL
9	9	Elisa Parada	La Unión	Cuyulitán	2023-10
10	10	Fatima Rivera	La Libertad	San Vicente	2023-10

At the bottom of the console, it indicates 'Elapsed time: 6841 ms' and 'Total rows: 101'.

*Ilustración 39. Dimensión cliente Amazon Redshift*

The screenshot shows the Amazon Redshift console interface. At the top, there are controls for running a query, including a 'Run' button, 'Limit 100', 'Explain', 'Isolated session', and a session identifier 'p3-2023-cluster'. The SQL query being executed is:

```

1 SELECT
2 *
3 FROM
4 "importadora-el-angel"."public"."dim_fecha";

```

Below the query, the results are displayed in a table with 100 rows. The first 7 rows are visible:

fecha_id	fecha_completa	anio	trimestre	mes	nombre
20200101	2020-01-01	2020	1	1	Enero
20200102	2020-01-02	2020	1	1	Enero
20200103	2020-01-03	2020	1	1	Enero
20200104	2020-01-04	2020	1	1	Enero
20200105	2020-01-05	2020	1	1	Enero
20200106	2020-01-06	2020	1	1	Enero
20200107	2020-01-07	2020	1	1	Enero

At the bottom of the console, it indicates 'Elapsed time: 6841 ms' and 'Total rows: 101'.

*Ilustración 40. Dimensión fecha Amazon Redshift*

The screenshot shows a SQL query in the Amazon Redshift console. The query is:
 

```

1 SELECT
2   *
3 FROM
4   "importadora-el-angel"."public"."dim_orden";
  
```

 The results table, titled 'Result 1 (100)', contains the following data:

orden_key	orden_id	estado	metodo	gran_total	departam
1	1	processing	checkmo	36	Cabaña
2	2	complete	checkmo	39	Cabaña
3	3	complete	paypal	740	Sonson
4	4	complete	paypal	490	Cuscatl
5	5	complete	credit_card	333	Morazá
6	6	complete	paypal	404	Cabaña
7	7	complete	paypal	216	San Vic

*Ilustración 41. Dimensión orden Amazon Redshift*

The screenshot shows a SQL query in the Amazon Redshift console. The query is:
 

```

1 SELECT
2   *
3 FROM
4   "importadora-el-angel"."public"."dim_producto";
  
```

 The results table, titled 'Result 1 (100)', contains the following data:

producto_key	producto_id	sku	nombre	descripcion	scd_star
2	2	24-MB04	Strive Shoulder Pack	Convenience is next to no...	2023-11
4	4	24-MB05	Wayfarer Messenger Bag	Perfect for class, work or t...	2023-11
5	5	24-MB06	Rival Field Messenger	The Rival Field Messenge...	2023-11
7	7	24-UB02	Impulse Duffie	Good for beach trips, trac...	2023-11
9	9	24-WB02	Compete Track Tote	The Compete Track Tote ...	2023-11
10	10	24-WB05	Savvy Shoulder Tote	Powerwalking to the gym ...	2023-11
15	15	24-UG06	Affirm Water Bottle	You'll stay hydrated with e...	2023-11

*Ilustración 42. Dimensión producto Amazon Redshift*

The screenshot shows a SQL query in the Amazon Redshift console. The query is a simple SELECT statement that retrieves all columns from the 'dim\_tienda' table in the 'public' schema of the 'importadora-el-angel' database. The results are displayed in a table with 7 rows and 4 columns: tienda\_key, tienda\_id, código, and nombre.

tienda_key	tienda_id	código	nombre
1	0	admin	Admin
2	1	default	Default Store View
3	2	SS001	Importaciones El Angel ...
4	3	CH001	Importaciones El Angel ...
5	4	SA001	Importaciones El Angel ...
6	5	SO001	Importaciones El Angel ...
7	6	LL001	Importaciones El Angel ...

Ilustración 43. Dimensión tienda Amazon Redshift

The screenshot shows a SQL query in the Amazon Redshift console. The query is a simple SELECT statement that retrieves all columns from the 'fact\_ventas' table in the 'public' schema of the 'importadora-el-angel' database. The results are displayed in a table with 7 rows and 7 columns: venta\_key, venta\_id, estado, cantidad\_ordenada, precio\_producto, and gran\_tot.

venta_key	venta_id	estado	cantidad_ordenada	precio_producto	gran_tot
1	1	processing	1	0	36
2	2	complete	1	0	39
3	3	complete	4	39	740
4	4	complete	5	62	740
5	5	complete	1	99	740
6	6	complete	1	45	490
7	7	complete	3	32	490

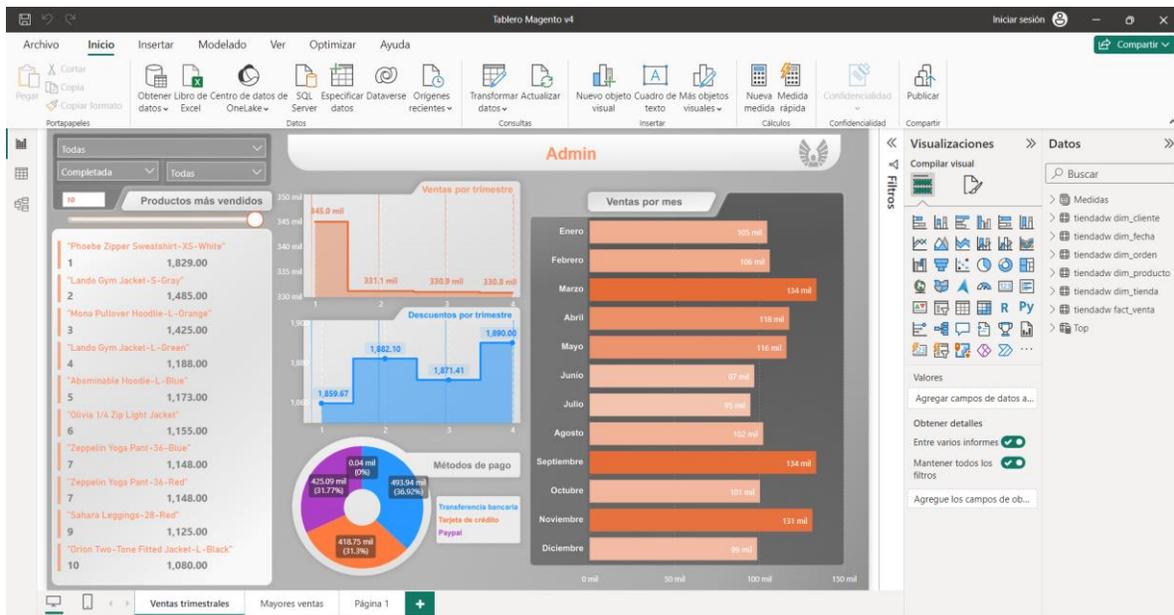
Ilustración 44. Fact ventas Amazon Redshift

- **Tableros PowerBI**

Al realizar con éxito la conexión desde PowerBI hacia la base de datos correspondiente al DataWarehouse que se encuentra alojado en el clúster de Amazon Redshift; El modelo dimensional y los datos de este, se cargan a PowerBI y los tableros diseñados reflejan los datos correspondientes a las ventas visualizando las métricas establecidas.

Las métricas que pueden visualizarse son las siguientes:

- Volumen de ventas totales cada trimestre por tienda.
- Producto con mayores ventas durante los últimos 3 meses por tienda.
- Producto con mayores ventas durante los últimos 3 meses. (Acumulado de todas las tiendas).
- Servicio de pago más usado en los últimos 3 meses.
- Meses con mayores ventas al año por tienda.
- Total, de descuentos aplicados en los últimos meses por tienda.



*Ilustración 45. Primer tablero en PowerBI*

Las métricas para el segundo tablero son las siguientes:

- Porcentaje de clientes a los que se les completa una venta mayor a \$600.
- Monto total de ventas mensuales por cliente.
- Ventas totales por cliente de un mismo departamento.

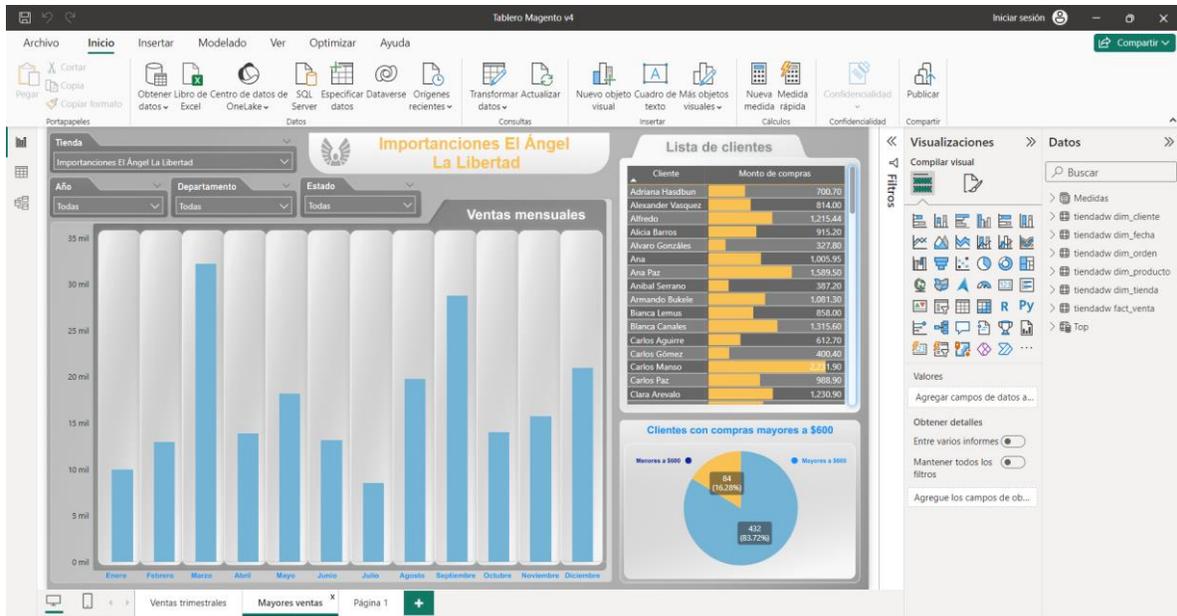


Ilustración 46. Segundo tablero en PowerBI

## CONCLUSIONES Y RECOMENDACIONES

### Conclusiones

- Las necesidades analíticas encontradas en la empresa “Importaciones El Ángel” son: volumen de ventas totales cada trimestre por tienda.; producto con mayores ventas durante los últimos 3 meses por tienda; producto con mayores ventas durante los últimos 3 meses. (Acumulado de todas las tiendas); servicio de pago más usado en los últimos 3 meses; meses con mayores ventas al año por tienda.; porcentaje de clientes a los que se les completa una venta mayor a \$600; monto total de ventas mensuales por cliente; ventas totales por cliente de un mismo departamento; total de descuentos aplicados en los últimos meses por tienda.
- Se creó un modelo de estrella conformado por las dimensiones; tienda, producto, orden, cliente, fecha y la tabla de hechos factventas, con una granularidad a nivel de línea de venta.
- A través de la herramienta de software Talend Open Studio se creó una rutina que, mediante un conjunto de Jobs, automatiza el proceso de extracción, transformación y carga de datos desde la base de datos transaccional de Magento Commerce hasta el DataWarehouse que ha sido diseñado.
- Se realizó el diseño y elaboración de dos tableros de datos utilizando Power BI, que permiten la visualización de las métricas identificadas a través del análisis.

### Recomendaciones

- La solución desarrollada requiere de un equipo con requisitos mínimos especificados en el desarrollo de este documento. El uso de un equipo con capacidades superiores incrementará la velocidad de procesamiento de los datos. Sin embargo, el rendimiento de la solución bajo los requisitos especificados funcionará de manera adecuada. Se recomienda utilizar un equipo aislado del equipo que aloja la base de datos transaccional para mantener, en la medida de lo posible, independencia en cuanto al rendimiento.
- La ejecución de la solución implica una consulta masiva sobre la base de datos transaccional que puede afectar a este durante un breve lapso de tiempo, causando que el servidor demore más tiempo en atender las peticiones del lado del cliente. Para minimizar el impacto sobre el sistema transaccional, se recomienda ejecutar la solución durante un momento dentro de las horas con menor actividad transaccional; estas suelen ser horas de la madrugada cuando los clientes son menos activos.
- Power BI ofrece flexibilidad de conexión a múltiples orígenes de datos. Esto hace posible que los tableros diseñados funcionen independientemente de la base de datos a la que se esté conectando; por este motivo, es importante conocer el procedimiento de cambio de origen de datos en PowerBI.

- La solución involucra los servicios de AWS como medio de maximizar la disponibilidad de los datos. No obstante, es posible no incurrir en el pago de los servicios de AWS y persistir los datos de manera local. Esto, como es de esperar, involucra una caída drástica en la disponibilidad de los datos al no contar con la disponibilidad de los datos.

## BIBLIOGRAFÍA

- Amazon Web Services. (s.f.). *aws pricing calculator*. Recuperado el 28 de 11 de 2023, de [https://calculator.aws/#/?refid=ap\\_card](https://calculator.aws/#/?refid=ap_card)
- Asiel. (2011). *Base de datos Avanzadas*. Obtenido de <https://asiel-bda.webnode.es/trabajos/tarea-1/ciclo-de-vida-dimENSIONAL-del-negocio/>
- ayudaley. (s.f.). *ayudaley*. Recuperado el 30 de 11 de 2023, de <https://ayudaleyprotecciondatos.es/bases-de-datos/transaccionales/>
- Calanca, P. (s.f.). *Data Science*. Obtenido de Ingeniería de datos: qué es y para qué sirve: <https://www.aluracursos.com/blog/ingenieria-de-datos>
- Computrabajo . (09 de 12 de 2023). *Computrabajo*. Obtenido de <https://sv.computrabajo.com/ofertas-de-trabajo/oferta-de-trabajo-de-programador-net-junior-plaza-presencial-en-antiguo-cuscatlan-8B49290C9E604B2261373E686DCF3405#lc=ListOffers-Score4-11>
- Espinosa, R. (19 de 4 de 2010). *El Rincon del BI*. Obtenido de Descubriendo el Business Intelligence...: <https://churriwifi.wordpress.com/2010/04/19/15-2-ampliacion-conceptos-del-modelado-dimENSIONAL/>
- Gonzalez), D. A. (06 de 07 de 2021). *Blog sobre Analytics, Desarrollo de Software Ágil, Data Science, Business Intelligence y Visualización de Datos*. Obtenido de <https://explodat.cl/Analytics/business-intelligence/la-metodologia-kimball-para-data-warehouses-y-bi-exitosos/>
- IBM. (s.f.). *Documentación*. Obtenido de InfoSphere Data Architect: <https://www.ibm.com/docs/es/ida/9.1.2?topic=concepts-dimensional-models>
- Kimball, R. (2013). *The Data Warehouse Toolkit* (Tercera ed.). Recuperado el 24 de 11 de 2023, de [https://github.com/ms2ag16/Books/blob/master/Kimball\\_The-Data-Warehouse-Toolkit-3rd-Edition.pdf](https://github.com/ms2ag16/Books/blob/master/Kimball_The-Data-Warehouse-Toolkit-3rd-Edition.pdf)
- Oracle. (s.f.). Obtenido de ¿Qué es big data?: <https://www.oracle.com/es/big-data/what-is-big-data/>
- Oracle. (s.f.). *Base de datos*. Obtenido de ¿Qué es OLTP?: <https://www.oracle.com/ar/database/what-is-oltp/#:~:text=OLTP%20o%20procesamiento%20de%20transacciones,mensajes%20de%20exto%2C%20por%20ejemplo>
- Talend a Qlik Company. (s.f.). *Talend Community*. Recuperado el 20 de noviembre de 2023, de <https://help.talend.com/>