UNIVERSIDAD DE EL SALVADOR
FACULTAD DE INGENIERIA Y ARQUITECTURA
ESCUELA DE INGENIERIA DE SISTEMA INFORMATICOS

CURSO DE ESPECIALIZACIÓN DE INGENIERÍA DE DATOS

ANÁLISIS DE VENTAS DE LA ORGANIZACIÓN "ALMACENES PAPAGAYO" QUE UTILIZA UNA TIENDA DE COMERCIO ELECTRÓNICO MONTADA SOBRE NOPCOMMERCE.

PRESENTADO POR:

KEVIN ALEXANDER CHICAS NOLASCO

CHRISTIAN ROBERTO MONTERROSA SURIO

CARLOS ALBERTO MORENO AREVALO

PARA OPTAR AL TÍTULO DE:
INGENIERO DE SISTEMAS INFORMÁTICOS

CIUDAD UNIVERSITARIA, ENERO 2023

UNIVERSIDAD DE EL SALVADOR

RECTOR:

MSC. ROGER ARMANDO ARIAS ALVARADO

SECRETARIO GENERAL:

MSC. FRANCISCO ANTONIO ALARCÓN SANDOVAL
FACULTAD DE INGENIERÍA Y ARQUITECTURA

DECANO:

PHD. EDGAR ARMANDO PEÑA FIGUEROA

SECRETARIO:

ING. JULIO ALBERTO PORTILLO
ESCUELA DE INGENIERÍA DE SISTEMAS INFORMÁTICOS

DIRECTOR:

ING. RUDY WILFREDO CHICAS VILLEGA

UNIVERSIDAD DE EL SALVADOR
FACULTAD DE INGENIERIA Y ARQUITECTURA
ESCUELA DE INGENIERIA DE SISTEMA INFORMATICOS

Trabajo de Graduación previo a la opción al Grado de:

INGENIERO DE SISTEMAS INFORMÁTICOS

Título:

ANÁLISIS DE VENTAS DE LA ORGANIZACIÓN "ALMACENES PAPAGAYO" QUE UTILIZA UNA TIENDA DE COMERCIO ELECTRÓNICO MONTADA SOBRE NOPCOMMERCE.

Presentado por:

KEVIN ALEXANDER CHICAS NOLASCO

CHRISTIAN ROBERTO MONTERROSA SURIO

CARLOS ALBERTO MORENO AREVALO

Trabajo de Graduación Aprobado por:

Docente Asesor:

Ing. RENE FABRICIO QUINTANILLA GOMEZ

SAN SALVADOR, ENERO 2023

Trabajo de Graduación Aprobado por:

Docente Asesor:

Ing. Rene Fabricio Quintanilla Gómez

ÍNDICE

Agradecimientos.

Gracias a Dios y a todas esas personas que estuvieron incondicionalmente para bien o para mal, para los que están y los que lamentablemente ya no lo están; gracias a todos ellos somos lo que somos y el seguir viviendo al día de hoy, es meramente una demostración que nuestro camino aun no acaba y aún nos queda mucho por mejorar.

Kevin Alexander Chicas Nolasco.

A mi madre, Carmen Elena Surio, por llevarme en sus oraciones, porque nunca se rindió conmigo y siempre confió en mí.

A mis amigos Jessica Vides y Henry López, quienes, cuando yo quería rendirme, lograron animarme para luchar, para cumplir esta meta.

A todos los que conocí y me tendieron su mano.

Christian Roberto Monterrosa Surio.

Principalmente quiero agradecer a Dios por brindarme la oportunidad de llegar hasta este punto de mi vida que en algún momento resultaba difícil de visualizar por la complejidad de la carrera; a mis padres que de manera incondicional estuvieron ahí apoyándome en cada momento de este largo trayecto, a mi compañera de vida que siempre vio lo mejor de mí y me motivo a seguir adelante, aunque muchas veces deseaba tirar la toalla. He aprendido mucho durante este viaje y he puesto en práctica mucho de ello y deseo seguir aprendiendo más porque esta carrera tiene un sinfín de ramas que están en constante evolución. Por ultimo y no menos importante, agradecer al personal docente por concedernos parte de su sabiduría y ver un ingeniero dentro de nosotros. GRACIAS.

Carlos Alberto Moreno Arévalo.

Introducción

La ingeniería de datos es la disciplina que se encarga de recolectar, trasladar y validar datos para su explicación. Utiliza principios de ingeniería de software para construir modelos de datos como almacenes de datos (DataWarehouse) y lagos de datos (Data Lake) dando así una garantía a la disponibilidad, mantenibilidad, consistencia, seguridad y limpieza de datos para que estos sean utilizados de la manera más apropiada.

Con la gran aparición del Big Data en la era digital que actualmente se vive, se vuelve necesario contar con metodologías sofisticadas y complejas pero eficientes para el manejo de grandes volúmenes de datos. Las bases de datos comunes relacionales se vuelven más limitadas sobre todo cuando se presentan datos no estructurados por lo que surgen una innumerable cantidad de nuevas tecnologías para completar esas necesidades.

Cuando se habla de un ingeniero de datos éste se especializa con estas herramientas y hace uso y dominio del negocio o industria en la que se trabaje, para volver a los datos, unos de los activos más valiosos que la organización pueda tener.

Este análisis permite que los analistas de datos planteen y respondan a preguntas como “qué pasó”, “por qué pasó”, “qué pasará” y “qué se puede hacer con los resultados”.

Almacenes Papagayo actualmente no cuenta con algún método o prototipo de análisis de datos significativo para dar una mejor administración a futuro, relacionado con nuevas decisiones a tomar como ofertas, descuentos a clientes, productos más vendidos, épocas y lugares donde más se vende un producto. Sin embargo, está al alcance de la organización proporcionar los recursos necesarios para que se elabore el análisis respectivo para las especificaciones que se han solicitado con el fin de mejorar la toma de decisiones en un futuro relativamente constante.

En el presente documento se ha dejado plasmado el trabajo realizado en tres capítulos principales. En el primer capítulo se realiza una descripción de la situación actual y de la problemática que se presenta con toda la data de Almacenes Papagayo además del planteamiento del problema donde se involucran diferentes herramientas para cada etapa del diseño de la solución. También, se describen aspectos como los objetivos, alcances y justificación del proyecto a realizar.

En el segundo capitulo se encuentra reflejado el análisis y diseño de la solución, desde la metodología de trabajo utilizada hasta la descripción de la arquitectura que sostiene la solución propuesta.

En el tercer y ultimo capitulo encontramos la estrategia de implementación de la propuesta de la solución que contiene datos acerca de la preparación del ambiente donde se instalará la solución, las herramientas y procesos ETL utilizados, presentación de la información y análisis de los resultados obtenidos.

Capítulo I: Especificación de proyecto

a. Situación actual

i. Antecedentes

Almacén Papagayo es una despensa en línea con más de 5 años en el mercado que cuenta con una gran variedad de productos de electrónica, vestimenta, libros, joyería, entre otros.

La organización ha presenciado un alto incremento en sus ventas a lo largo del tiempo y la acumulación de datos de transacciones de diversos países, clientes y productos al momento de la realización de tanta venta. La organización se ha visto en la necesidad de conocer cuál es el comportamiento por el que atraviesan sus productos, así como los que se venden más, donde es que se vende más o incluso en que época hay más ventas para el almacén.

Esto ha complicado un poco la toma de decisiones para el gerente ya que la cantidad de datos ha sido considerablemente grande como para poder dar una solución pronta a dicha necesidad.

ii. Descripción del problema

En la actualidad, el alto incremento de datos suele generar una serie de inconvenientes para las empresas al momento de buscar algún dato o conjunto de datos que compartan alguna peculiaridad en común, al intentar detectar alguna anomalía o algo que este fuera de lo relativamente normal o realizar algún tipo de análisis en específico. Este tipo de inconvenientes tienen mayor presencia en las empresas que realizan a diario un considerablemente alto volumen de ventas y/o cuyos negocios llevan mucho tiempo en el mercado acumulando una gran cantidad de datos como consecuencia de las actividades del negocio y por lo que tomaría mucho tiempo consultar un dato en específico. Teniendo en cuenta que muchas empresas están empezando a evolucionar involucrando la actividad de sus negocios a la era digital y que los datos todas sus ventas han sido migradas de forma digital dejando atrás el papel; a pesar de esto, la búsqueda de información que se necesiten de forma casi inmediata o con alguna característica en común dentro de una inmensa cantidad de datos, representa una dificultad.

Imagine una situación en la que un negocio sabe que hay productos que se venden con mayor frecuencia, pero no se sabe dónde se venden más, en que épocas se vende más o si algún producto en particular genera más ingresos monetarios o es más frecuente su venta independientemente de su precio; Para la gerencia sería de gran interés tener la capacidad de hacer una gestión para dichos productos, haciendo que la actividad del negocio adopte un carácter más estratégico y minucioso. De esta manera se aprovecha la gran oportunidad que representa la información de la organización respecto a las ventas ha sido simplemente "almacenada" pero nunca analizada, convirtiéndolas así en un activo.

Con el avance de la tecnología han aparecido una diversidad de aplicaciones digitales con las cuales una organización puede gestionar una gran cantidad de información en cuestión de minutos y mostrar reportes y/o análisis claros para una necesidad en concreto sobre un proceso de negocios.

iii. Planteamiento del problema

Cuando se trata de Ingeniería de Datos y se busca la manera de desarrollarla correctamente, hoy en día la tecnología brinda innumerables posibilidades de efectuar procesos de modelados de datos ya que no siempre tienden a estar en una estructura estándar como lo son las bases de datos relacionales, por ejemplo, y vienen representadas simplemente como un Data Lake (lago de datos) donde los datos no tienen una organización o estructura específica, simplemente están ahí "existiendo"; así también para realizar procesos de extracción, transformación y carga de datos que para muchos nos podría resultar complicado y desconocido hacer una depuración de datos y posteriormente realizar una representación de ellos a partir del resultado de esa misma depuración para lograr identificar alguna característica en particular según los criterios y necesidades que se hayan tenido antes de realizar dichos procesos, aún más cuando la información es demasiada y variada, sin tomar en cuenta la falta de estructuración de los datos antes mencionada.

A pesar de la gran cantidad de herramientas que existen en la era digital para el trabajo dentro de la Ingeniería de Datos se resulta complejo realizar un manejo limpio y completo de los datos, lo cual da origen a la necesidad de identificar qué herramienta tecnológica se acopla más adecuadamente a la necesidad que se presente ya que algunas tienen mayor capacidad de procesamiento analítico que otras y ventajas convenientes para la organización. Sin embargo, no solo se debe de considerar la capacidad analítica y efectiva con la que estas herramientas trabajen, sino también el tiempo que estas requieran para realizar los procesos, ya que sería contraproducente hacer uso de una herramienta "poderosa" para el análisis pero que consuma mucho tiempo y recursos (hardware, software adicional como extensiones de paga, entre otros) para llevarlo a cabo. Otra dificultad que se suele manifestar, cuando se habla de trabajar con software desconocido, es lo fácil y practica que sea la interacción con el usuario ya que muchas herramientas suelen cumplir su función, pero suele resultar complicado para muchos lograr comprender su correcto y optimo funcionamiento.

Tomando en cuenta las circunstancias que se han mencionado, se ha llegado a seleccionar ciertas herramientas para ciertos procesos específicos entre las cuales podemos mencionar Talend Open Studio como herramienta ETL donde sus principales ventajas pueden mencionarse: un ***all-in-one***, lo que permite prescindir y reducir el número de herramientas y, por lo tanto, de configuraciones adicionales; Documentación y comunidad relativamente longevas y bien estructuradas; El uso de componentes genéricos permite a Talend conectarse

con prácticamente cualquier plataforma para la extracción de datos; Existe un componente para casi cualquier acción y muchas otras cosas que puede brindar esta herramienta.

Otra herramienta a considerar dentro de este proyecto ha sido Power BI que se utiliza para convertir datos sin procesar en información significativa mediante el uso de visualizaciones y tablas intuitivas. Gracias a esta herramienta, se puede analizar los datos fácilmente y tomar decisiones comerciales importantes basadas en ellos; y que, al igual que la herramienta de ETL, Power BI también posee muchas ventajas para el cómodo manejo con el usuario.

Cabe mencionar que requiere tiempo para lograr entender el buen funcionamiento de la Ingeniera de Datos sobre una gran cantidad de información, al igual recursos tanto materiales como digitales que, a través de éstos, se logra realizar un buen proceso de Datawarehouse, ETL y presentación de datos con los cuales un ingeniero de datos se caracteriza dentro de circunstancias similares dentro de una organización, independientemente del rubro o funciones que realicen. Además, se necesita que la persona o grupo de personas encargados de gestionar estos procesos esté óptimamente capacitados para del uso de las herramientas y materiales que se necesiten.

En la organización de Almacenes Papagayo se logran identificar estas circunstancias antes mencionadas por lo que se ha tomado a bien solventar dichas dificultades para la finalidad principal de este proyecto lo cual involucraría la obtención e instalación de ciertas herramientas digitales para realizar la Ingeniería de Datos con las ventas de ésta organización y darle así un mejor enfoque a la lógica de negocios inclinada a representación de datos específicos según la necesidad que se tenga para la mejor toma de decisiones en beneficio del negocio.

b. Objetivos

Objetivo general

Construir un prototipo para la implementación de un DataWarehouse sobre el análisis de datos de las ventas para la empresa ALMACENES PAPAGAYO que permita depurar, transformar y presentar información que apoye a la toma de decisiones, utilizando diversas herramientas tecnológicas.

Objetivos específicos

- Aplicar las diferentes metodologías y técnicas para la extracción, transformación y carga de datos para detallar las especificaciones requeridas para el desarrollo del análisis de datos.
- Identificar la información relevante proveniente de la base de datos transaccional que conformarán el lago de datos a utilizar para la transformación y construcción del Datawarehouse de Almacenes Papagayo.
- Realizar un diseño de modelo dimensional que represente la arquitectura del prototipo de Datawarehouse de Almacenes Papagayo
- Utilizar la herramienta Talend Open Estudio, para desarrollar el prototipo que realizará los procesos de Extracción, Transformación y Carga (ETL) de datos que conformarán el Datawarehouse de Almacenes Papagayo.
- Utilizar la herramienta de Microsoft Power BI para realizar, a partir de los datos procesados producto del ETL diseñado, la presentación de los informes de datos que faciliten el análisis y toma de decisiones para los gerentes de la organización, con base a las necesidades y los requerimientos solicitados.

c. Alcances

- Lograr que personal administrativo de la organización obtenga un mejor modelo de análisis de los datos generados con sus ventas para mejor toma de decisiones.

- Se espera que el personal administrativo logre usar la Ingeniería de Datos y la tecnología para un mejor enfoque sobre el negocio analítico de la organización.

- Al efectuar pruebas delante del personal administrativo de la organización, se espera que se logre convencer acerca de las ventajas que se posee al utilizar este tipo de metodologías analíticas con la Ingeniería de Datos.

- Se mostrarán las herramientas al personal administrativo de la organización esperando que sean utilizadas en ella, siendo previamente convencidos y conscientes de las ventajas que se poseen en su utilización y que sean las herramientas mas apropiadas para la mejoría pronunciada de una venta común contra un enfoque analítico de las ventas y productos relacionados a esta para el beneficio del negocio.

- Se pretende facilitar a la organización la toma de decisiones por medio de las herramientas mencionadas en este documento, diferenciando un enfoque tradicional de ventas contra las ventajas de usar la Ingeniería de Datos y las herramientas adecuadas según sea la necesidad o circunstancia.

d. Justificación

La problemática de esta organización es la dificultad que se tiene para hacer análisis de datos para tomas de decisiones específicas, ya que su lógica de negocio ha sido nada más que solo vender sin ningún otro enfoque favorable para el negocio, así de igual manera la demostración de la información almacenada de manera resumida y útil no ha sido la adecuad conforme la necesidad que se presenta para mejorar el negocio de la organización, adicionalmente se necesita el software necesario para llevar a cabo toda esta metodología de análisis para futuras ocasiones, también existe la necesidad de la depuración de los datos ya que la cantidad de información es grande y suelen existir datos que no sean necesarios según lo que se necesite analizar y posteriormente presentar gráficamente de alguna manera. Como necesidad complementaria, pero no menos importante, si la información no puede ser filtrada ni depurada, tampoco se podrá presentar de manera amigable al personal administrativo de la organización ya que puede ser muy confusa con el gran volumen de datos que esta puede generar con el pasar del tiempo a lo cual se necesita una manera de mostrar al personal una manera resumida de los datos para su análisis.

e. Cronograma de actividades

En el cronograma de actividades se muestran las tareas necesarias para llevar a cabo el presente proyecto.

Para visualizar el cronograma de actividades, dirigirse al anexo 3 del presente documento.

f. Presupuesto

Este presupuesto que se presenta a continuación contiene los gastos realizados en el proceso de desarrollo del proyecto.

Para visualizar el cronograma de actividades, dirigirse al anexo 4 del presente documento.

Capitulo II: Análisis y diseño de la propuesta de solución

a. Metodología de trabajo

El proceso de construcción de la solución se ha llevado a cabo en cuatro grandes etapas, definiendo primeramente el estudio del origen de los datos, seguido de la construcción de un modelo dimensional; el diseño de los procesos ETL para el tratamiento de la información en conjunto con herramientas de Amazon Web Services (en adelante AWS) para soportar el Data Lake y el almacenamiento de los datos; finalmente el diseño de dashboards para la presentación de los datos a través de la herramienta de Microsoft Power BI.

Estudio del origen de los datos.

Actualmente el negocio utiliza la plataforma de código abierto NopCommerce desarrollada en ASP.NET Con una base de datos de MS SQL Server 2008. Para poder instalar la plataforma y ver la estructura de la base de datos es necesario crear una base de datos previamente; posteriormente utilizar una de las dos opciones que se ofrecen en la Web oficial de la plataforma para su instalación.

Cuando se ha logrado instalar correctamente la plataforma, se crea automáticamente toda la estructura de la base de datos, la cual posee 126 tablas por defecto. La gran mayoría de las tablas contienen muchos campos que pertenecen a las configuraciones que se pueden realizar en la plataforma al gusto del cliente, sin embargo, únicamente se hará uso de aquellos campos que son de interés para el desarrollo del modelo dimensional.

Para determinar los datos necesarios que se necesitan en el desarrollo de la solución, es imperativo agrupar y enfocarse en aquellas tablas donde se almacenan los datos de interés; especialmente se realiza un enfoque en las tablas que contienen información acerca de los clientes, productos, direcciones y las que contienen información sobre las ventas.

A continuación, se muestran de forma resumida los esquemas de la base de datos correspondientes a los aspectos mencionados anteriormente (ver imágenes 1, 2, 3 y 4)

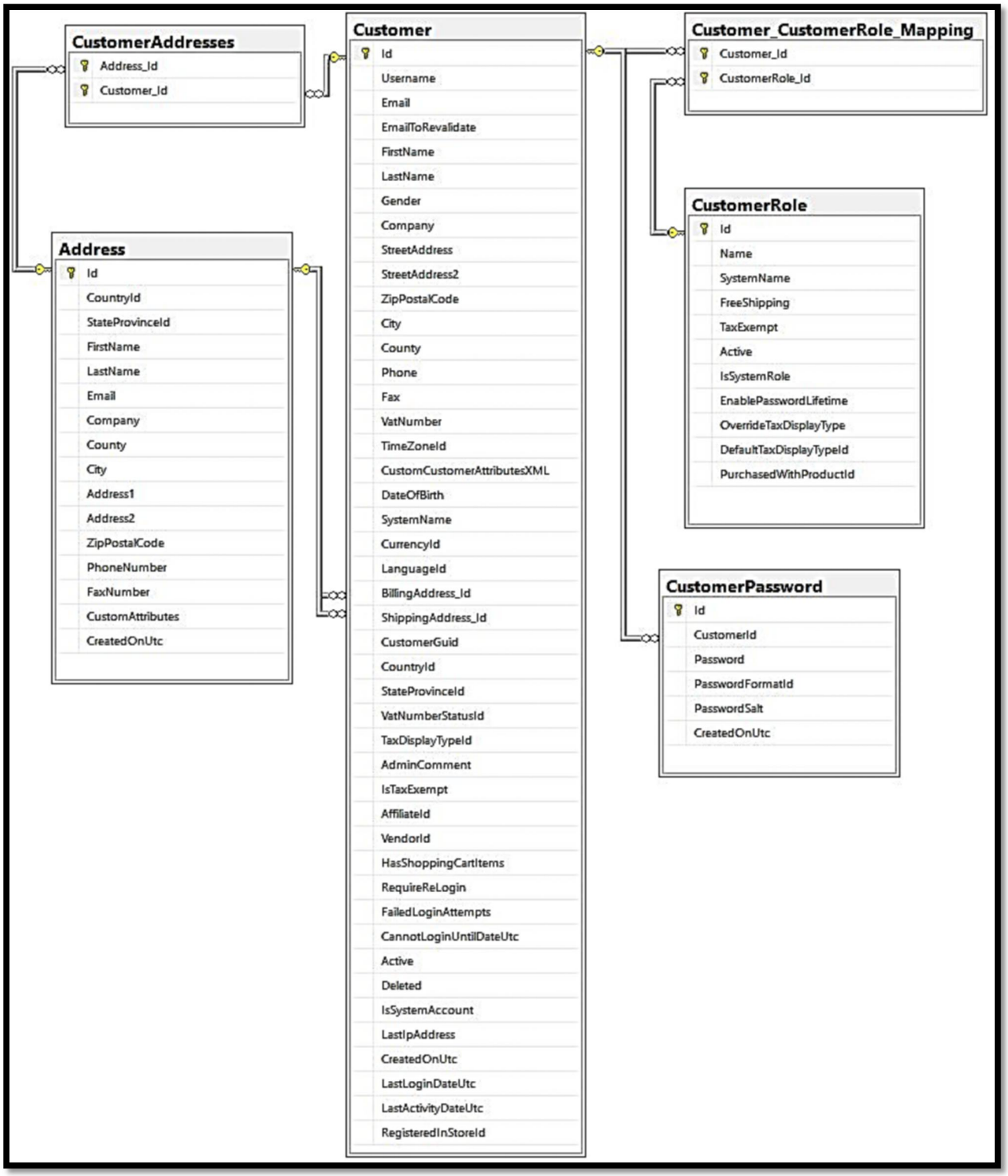


Imagen 1: Información relacionada a los Clientes

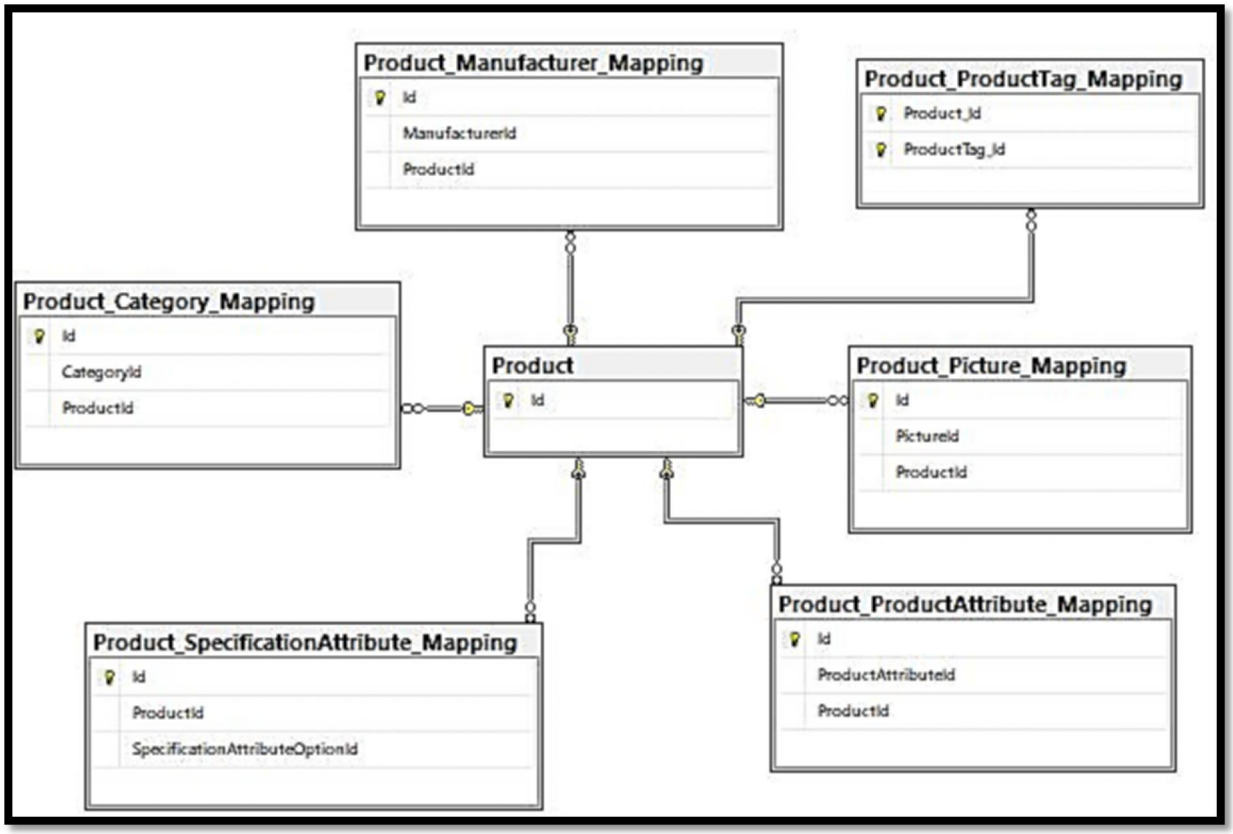


Imagen 2: Información relacionada a los Productos:

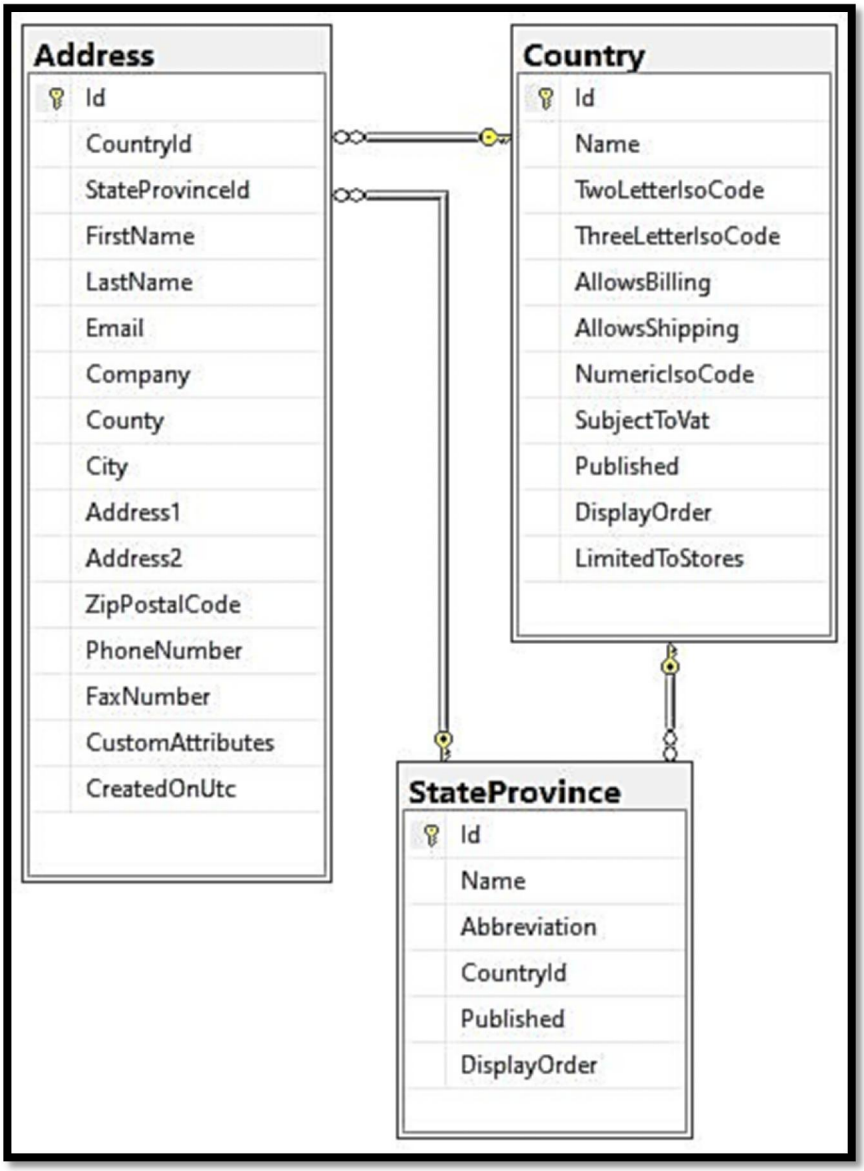


Imagen 3: Información relacionada a las direcciones de envío

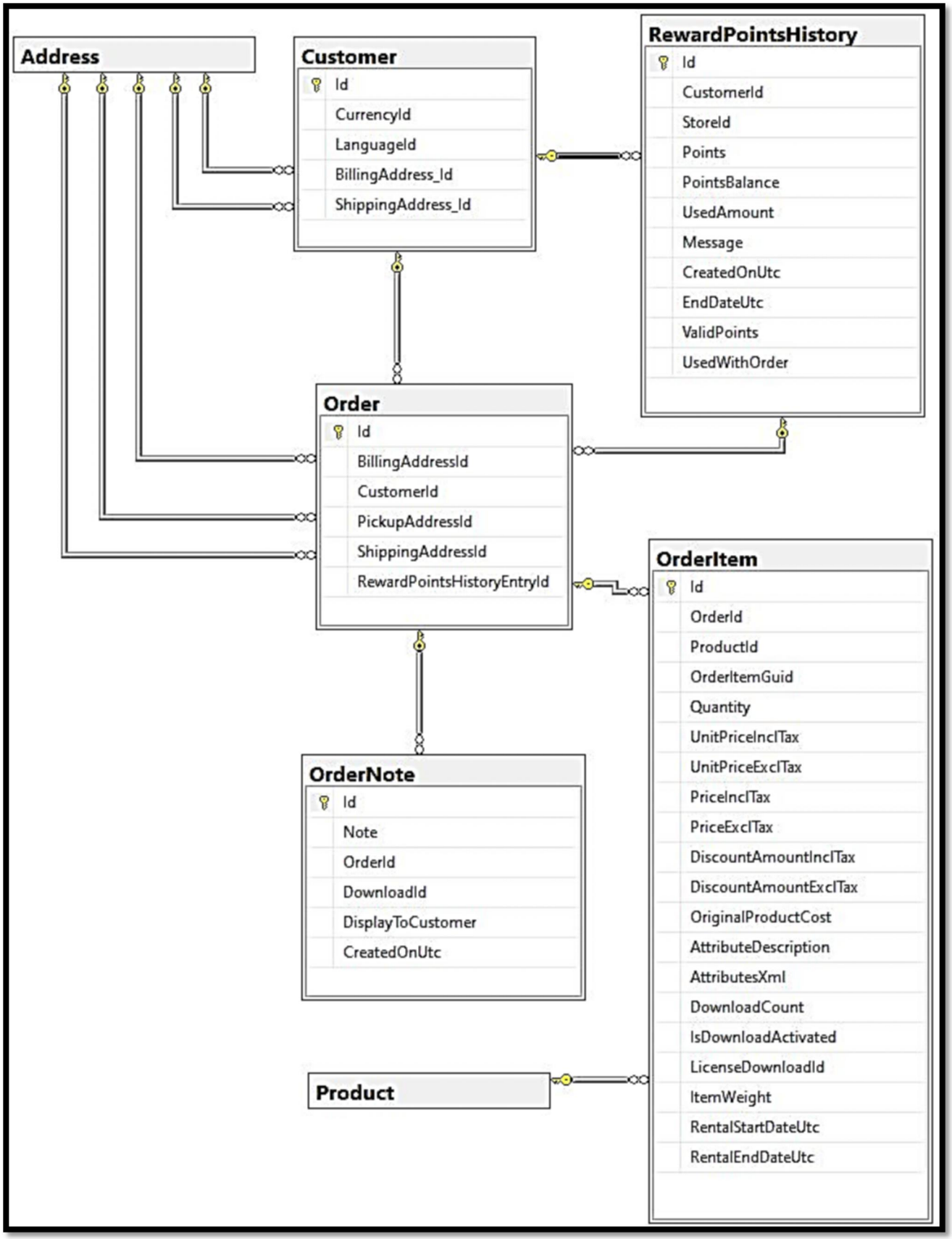
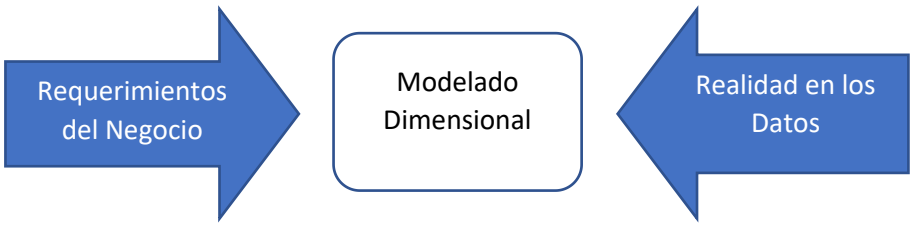


Imagen 4: Información relacionada a las ventas

Creación del modelo dimensional.

La metodología utilizada para llegar a una solución al problema presentado comienza con la propuesta de Ralph Kimball para el modelado dimensional de los datos, [1] especialmente el proceso de diseño dimensional en cuatro pasos, con el objetivo de obtener información acerca de las necesidades del negocio y la capacidad de los datos existentes para cubrir esas necesidades.



El proceso de diseño dimensional en cuatro pasos incluye:

1. Seleccionar el proceso de negocio.
2. Definir la granularidad.
3. Identificar las dimensiones.
4. Identificar las métricas.

Seleccionar el proceso de negocio consiste en escoger una de las actividades que se llevan a cabo en la organización como bien pueden ser ventas, registros de tickets, vuelos, reservaciones, etc. Los datos concernientes al proceso de negocio seleccionado posteriormente pasaran a la tabla de hechos (Fact table en inglés).

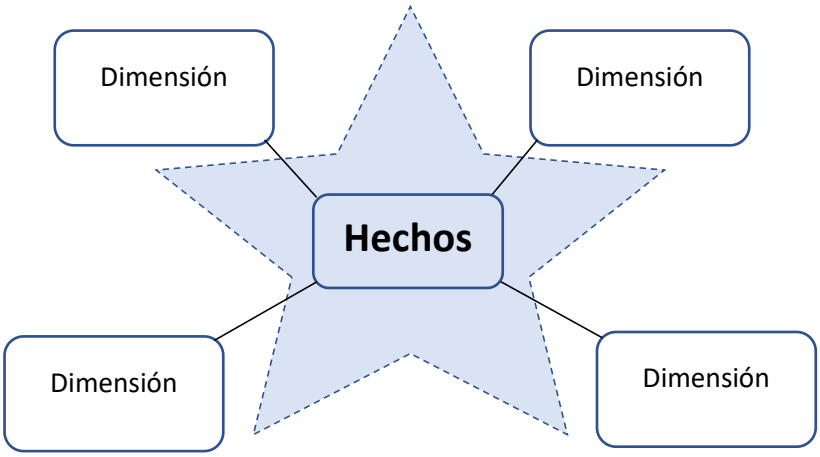
La tabla de hechos únicamente debe contener información concerniente al hecho que se quiere medir, por ejemplo, el único dato concerniente a las ventas es el monto de una venta lo cual es un dato numérico y aditivo. El resto de información como el nombre del producto que se ha vendido, el nombre de la tienda, la categoría y otros datos de interés son almacenados en las dimensiones.

Una vez seleccionado el proceso del negocio a analizar, se debe definir qué nivel de detalle se posee en el sistema fuente en el que se encuentran los datos, a esto se le conoce como granularidad en los datos; este nivel de detalle se debe acordar con el usuario de negocio que hará uso del DataWarehouse y se verá reflejado en cada una de las filas presentes en la tabla de hechos.

El siguiente paso es definir las dimensiones que conforman el modelo dimensional y sus atributos de interés, los cuales describen el contexto en el que sucede el evento que se necesita medir. Las dimensiones dotan de información descriptiva sobre los hechos, además, ayudan a filtrar y agrupar los hechos que se quieren medir. Normalmente las dimensiones responden a preguntas como: ¿Qué?, ¿Cuándo?, ¿Donde?, ¿Quién?, ¿Cómo? Y ¿Por qué? De los hechos.

Una vez definidas las dimensiones, se debe declarar las métricas que necesita el usuario de negocio. Usualmente el usuario es quien define qué es lo que quiere medir en función de un atributo numérico presente en la tabla de hechos.

Como resultado del proceso del diseño del modelo dimensional se tiene un esquema estrella, donde al centro se ubica la tabla de hechos y en cada una de las puntas, las dimensiones que dan el contexto de los hechos:



Es necesario mencionar que además de los atributos numéricos contenidos en la tabla de hechos, aquí se almacenan también las referencias a cada una de las dimensiones que conforman el modelo estrella.

Procesos ETL.

Para poder llevar toda la información necesaria desde el sistema transaccional hasta el Data Warehouse, se deben diseñar los procesos encargados de la extracción, transformación, y carga de datos, los cuales se pueden construir utilizando diversas herramientas especializadas en procesos ETL.

Como primer paso para la construcción de los procesos de extracción de datos, se debe leer y comprender el origen de los datos. En este caso, los datos se encuentran en una base de datos en SQL Server y previamente se han definido las tablas en la base de datos que poseen información relevante para la construcción del DW. Una vez extraídos los datos de interés de la base de datos, se almacenan en el Data Lake (previamente construido e identificado con tres zonas: Raw, Staging y Presentation) específicamente en la zona Raw.

A partir de la extracción, los datos quedan disponibles para realizar con ellos cualquier operación o manipulación, pudiendo ser: limpieza de datos, cambios en los formatos de origen, tratamiento de datos nulos, etc. A esta etapa se le conoce como la etapa de Transformación, en la cual los datos se adecuan al modelo dimensional que ha sido creado previamente y se guardan en la zona de Staging en el Data Lake. En esta etapa se pueden reconocer aspectos que pueden mejorar en los procesos del negocio y en la manera en cómo se capturan los datos en el sistema transaccional.

La fase de Carga de datos se realiza en la zona Presentation del Data Lake. En esta zona los datos son resguardados hasta que se envían al sistema capaz de soportar la estructura dimensional, que para este caso se utiliza el servicio Redshift de AWS.

Presentación de los datos.

La presentación de los datos se ha realizado a través de tableros diseñados con la aplicación Power BI. Los tableros contienen información obtenida del DW a través de la conexión entre Power Bi y Redshift; de esta forma Power BI es capaz de presentar información de forma interactiva y sencilla al usuario de negocio final con el propósito de brindar una guía para la toma de decisiones en el negocio.

b. Descripción de la propuesta de solución.

Como parte de la propuesta de solución se propone el siguiente análisis para el modelado dimensional:

1. Proceso de negocios

La actividad seleccionada para su análisis es el proceso de ventas.

2. Nivel de granularidad

    a. ¿Qué detalle requiere el usuario del negocio?

El usuario de negocio necesita ver las ventas en función de productos, marca, ventas por año/mes/día, países.

    b. ¿Qué detalle es efectivamente posible con los datos?

Se pueden ver las ventas en función de todo lo que el usuario de negocio requiere ya que actualmente el sistema transaccional cuenta con los datos necesarios para poder presentar ese tipo de información.

3. Identificar las dimensiones

Las dimensiones que se proponen son: Product, Date, Customer, y Address.

4. Identificar las métricas: Usualmente el usuario final decide que es lo que quiere medir

- Necesito saber el producto más vendido cada 3 meses.
- Necesito saber el lugar donde se vende más ese producto por cada trimestre.
- Necesito saber qué día de la semana se vende más.
- Necesito saber que cliente es el que más compra para hacerle un descuento.
- Necesito saber de qué marca se ha vendido más por cada 3 meses en el último año.
- Necesito saber cuál es el producto menos vendido cada 3 meses.

Diagrama propuesto para el modelo dimensional:

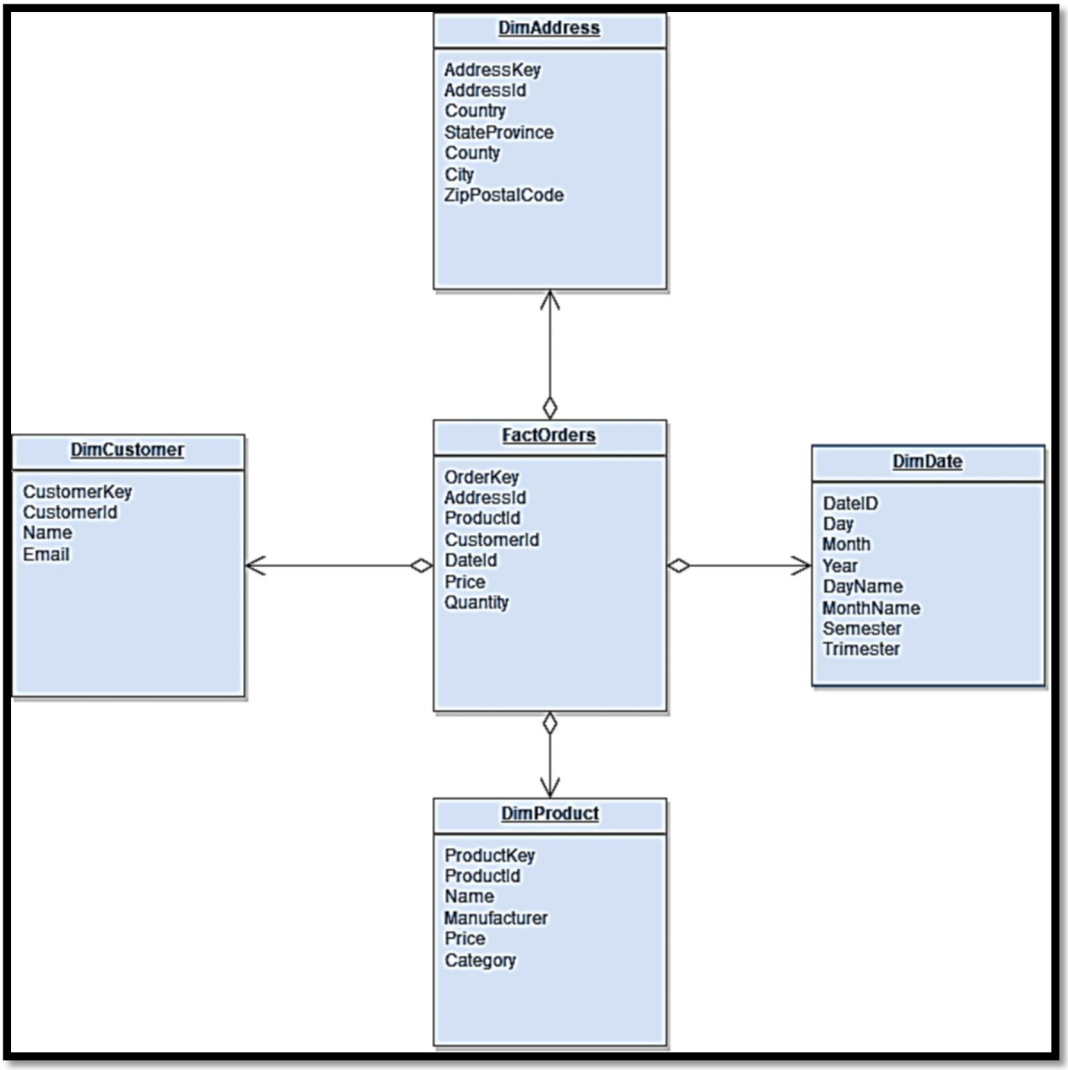


Imagen 5: Diagrama propuesto para el modelo dimensional

Al tener definido la información a recopilar y almacenar en el DW, se procede a definir los procesos ETL con la herramienta Talend Open Studio para adecuar los datos y llevarlos a la plataforma donde quedara almacenado el DW y a partir de ahí los datos se podrán consultar y generar las visualizaciones a través de Power BI.

Se han definido procesos de extracción de datos a través de una conexión a la base de datos del sistema transaccional; una vez extraídos son llevados a la zona raw del Data Lake. El Data Lake se ha elaborado en un bucket (contenedor de objetos) con el servicio S3 de AWS en el cual se han dejado carpetas identificadas cada una con el nombre de las tres zonas utilizadas (Raw, Staging y Presentation).

Una vez extraídos los datos son alojados en la carpeta raw en formato de archivos separados por coma CSV; posteriormente, los datos son transformados de acuerdo con las necesidades descritas por el usuario del negocio. Una vez los datos han sido transformados son llevados a la zona de presentación y cargados en el servicio de Amazon Redshift donde será almacenado el DW.

Para la visualización de la información que brindará apoyo a la toma de decisiones, se han elaborado visualizaciones interactivas con Power BI, donde se establece la conexión con Amazon Redshift para obtener toda la información contenida en el DW.

c. Descripción de la tecnología a utilizar

Se han utilizado diferentes herramientas tecnológicas para el desarrollo de la solución. Aquí se describen las herramientas utilizadas desde el sistema transaccional hasta la presentación de los datos que es el producto final donde el usuario de negocio interactúa con la información solicitada.

NopCommerce

NopCommerce es una plataforma dirigida al comercio electrónico de código abierto con una comunidad muy activa entre usuarios y también desarrolladores [2]; además ofrece una extensa documentación disponible y gratuita para la personalización completa de la herramienta, tanto para la interfaz como para el desarrollo de nuevas funcionalidades.

Esta plataforma se ofrece bajo dos modalidades, una totalmente gratuita donde se debe descargar el código fuente y montar el proyecto utilizando en una base de datos de Microsoft SQL Server y otra donde se ofrece el servicio de hosting lo cual es de pago dependiendo el tiempo que se desee utilizar.

La plataforma posee dos interfaces, una para usuarios finales y otra de administración donde se hace un seguimiento de todas las transacciones, permisos y configuraciones relacionadas a la funcionalidad de la tienda en línea.

Talend Open Studio

Talend Open Studio es una suite que ofrece diversas herramientas para la creación de procesos ETL que son de mucha utilidad en el ámbito de Ciencia de Datos.[3] Esta herramienta posee dos versiones, una versión community edition que posee toda la funcionalidad disponible para el desarrollo de esta solución, y una versión de pago que ofrecen distintas herramientas para la integración de datos de diferentes fuentes.

Talend Open Studio está disponible para poder utilizarse en distintas familias de sistemas operativos (Linux, Microsoft, Mac y Amazon Workspace), aunque idealmente se recomienda su uso en sistemas con Windows 10, Catalina 10.15, Mojave 10.14, o High Sierra 10.13. Los requerimientos para la instalación y uso de Talend Open Studio recomendados son: 8 GB de memoria RAM, 20 GB o más de almacenamiento disponibles.

Esta herramienta posee la capacidad de integrarse con múltiples fuentes de datos y tiene una gran variedad de componentes disponibles para la creación de procesos ETL y la conexión con servicios de AWS que son los que se utilizan en este proyecto.

Amazon Web Services.

Principalmente se han utilizado dos de los servicios que ofrece AWS para el desarrollo de la solución. Los servicios utilizados son Amazon Simple Storage Service abreviado como S3 y Amazon Redshift, a continuación, se describe cada uno de ellos.

Amazon Simple Storage Service (S3).

S3 es uno de los servicios web ofrecidos por Amazon desde 2006 para el almacenamiento de objetos, entre sus principales puntos destacables S3 ofrece una solución para el almacenamiento de objetos segura, duradera y escalable a un bajo costo debido a que es un servicio bajo demanda,[4] es decir, se paga conforme lo que se utiliza. Amazon tiene en cuenta seis componentes de costos para S3 los cuales son: el precio de almacenamiento, el precio de solicitud y de recuperación de datos, el precio de transferencia y de aceleración de transferencia de datos, el precio de administración y análisis de datos, precio de replicación y el precio de procesamiento de datos con S3 Object Lambda.

Amazon S3 permite almacenar cualquier tipo de objetos de datos en estructuras llamadas cubos (buckets en inglés), lo que permite utilizarlo como un lago de datos (Data Lake en inglés); al utilizar este servicio se pretende obtener escalabilidad, alta disponibilidad y baja latencia con alta durabilidad además de otros beneficios como un costo considerablemente más bajo en comparación a otros servicios similares.

Hay que señalar también que el precio dependerá también de la clase de almacenamiento que se utiliza, estas clases son: S3 Standard, S3 Intelligent-Tiering, S3 Standard-Infrequent Access, S3 One Zone-Infrequent Access, S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval (antes S3 Glacier) y S3 Glacier Deep Archive. Para la demostración de este proyecto se utiliza una clase de almacenamiento estándar.

Amazon Redshift.

Es un servicio de AWS utilizado para el almacenamiento masivo de datos llegando hasta el orden de los petabytes. Amazon Redshift está orientado al almacenamiento de datos en la nube tanto para datos estructurados como semi estructurados.

Amazon Redshift soporta el uso de SQL para la consulta y recuperación de datos de manera eficiente ya que monitorea continuamente la carga de trabajo de los usuarios y utiliza métodos sofisticados para determinar mejoras que pueden aplicarse al diseño físico de los datos con el fin de optimizar la velocidad de las consultas.

Un almacén de datos en Amazon Redshift se compone de nodos que se organizan en un grupo llamado clúster y cada uno de estos clústeres dispone de un motor de Amazon Redshift; cada clúster se conforma de un nodo principal y uno o más nodos de computación. El nodo principal es el encargado de analizar las consultas y crear planes de ejecución que posteriormente son ejecutados de forma paralela por los nodos de computación.

Gracias a la capacidad de Amazon Redshift llamada elastic resize, es posible contar con un servicio escalable y en el momento que se necesite un mayor rendimiento se pueden agregar más nodos a un clúster o mayor capacidad de almacenamiento. Se puede visualizar una ilustración de esta característica en la imagen 6.

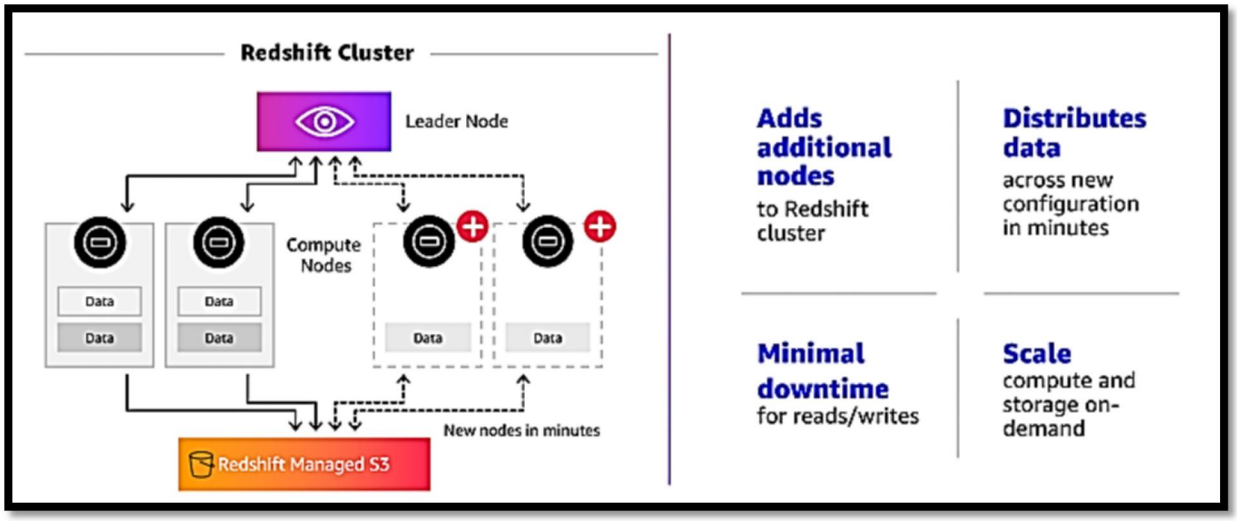


Imagen 6: nodos y clúster en AWS Redshift

Power BI

Power BI es un conjunto de herramientas de Microsoft,[6] utilizada para el análisis de datos capaz de crear visualizaciones interactivas para el usuario de negocio. Power BI es capaz de conectarse a diferentes fuentes de datos, analizarlos y presentarlos por medio de gráficas, informes y paneles interactivos.

Con Power BI es posible conectarse a diferentes orígenes como servicios en la nube de AWS, Azure de Microsoft, archivos locales como hojas de Excel, archivos en texto plano entre otros. Power BI cuenta con tres componentes para su utilización los cuales son:

Power BI Desktop: Es la aplicación de escritorio gratuita que Microsoft pone a la disposición para poder visualizar y crear informes de los datos.

Power BI Service: Este es un servicio de pago con la misma funcionalidad de la versión Desktop con el agregado de poder configurar actualizaciones de los datos automáticamente, de esta forma los usuarios tendrán acceso a información actualizada.

Power BI Mobile: Esta es una aplicación disponible para las plataformas Android, IOS y Windows para la visualización de los datos a través de los smartphones.

Para la elaboración de las visualizaciones en este proyecto se hace uso de Power BI Desktop.

d. Diagrama arquitectónico de la solución

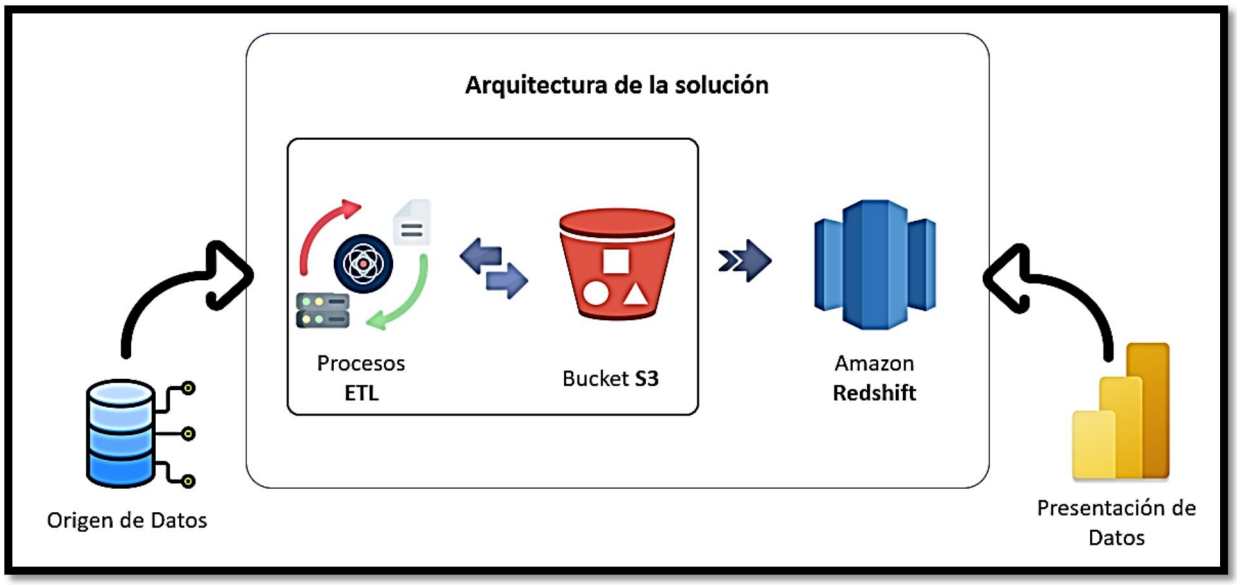


Imagen 7: diagrama arquitectónico de la solución.

e. Descripción de cada componente de la solución.

Origen de Datos: La base de datos del sistema transaccional contiene toda la información relacionada a las ventas, clientes, productos, direcciones y fechas que se realizan a través de la tienda en línea montada en NopCommerce. Esta base de datos de Microsoft SQL Server sirve como punto de partida para determinar la forma en como son almacenados los datos y realizar la estrategia del diseño de los procesos ETL.

Procesos ETL: Estos procesos son los encargados de extraer toda la información necesaria de la base de datos transaccional y llevarla al lago de datos utilizando la suite de Talend Open Studio (TOS). Una vez extraídos los datos, los procesos de transformación se encargan de realizar todas aquella modificaciones, sustituciones, reemplazo y limpieza de datos. Finalmente, los datos se cargan a través de procesos para ese fin en la zona de presentación del bucket en S3.

Bucket S3: El lago de datos montado en S3 es utilizado continuamente durante todas las fases del proceso de extracción, transformación y carga de los datos por medio de los procesos ETL construidos en TOS. Los datos extraídos son llevados a la zona Raw tal y como están en la base de datos. Al realizar algún proceso de transformación de los datos, estos son cargados a

una zona intermedia llamada Staging. Finalmente, los datos se adecuan al modelo dimensional propuesto y son llevados a la zona Presentation del lago de datos.

Amazon Redshift: Este componente que aparece en el diagrama de la arquitectura de solución es utilizado para alojar toda la data proveniente de la zona Presentation del lago de datos. La estructura de estos datos ya se encuentra acorde al modelo dimensional y listo para ser almacenado en el Clúster de Amazon Redshift. Una vez alojados los datos en el clúster, se puede aprovechar todas las características que ofrece Redshift para la consulta de grandes cantidades de datos.

Presentación de Datos: El ultimo componente que interviene en el diagrama arquitectónico de la solución es Power BI, que es la herramienta de inteligencia de negocios destinada a utilizar para la creación de visualizaciones interactivas de los datos disponibles en Amazon Redshift y que servirán de base para la toma de decisiones en el negocio.

Capitulo III: Estrategia de implementación de propuesta de solución

a. Estrategia de implementación

La solución diseñada para satisfacer la solicitud de Almacenes Papagayo sobre su necesidad de información de datos en forma de reporte para la toma de decisiones gerenciales, representa un producto por integración pues, esta solución trabaja en conjunto con el sistema transaccional de Almacenes Papagayo que es el principal insumo de la solución al ser la fuente de recolección de dato, más específicamente, con su base de datos. La solución está compuesta por una serie de módulos que trabajan en conjunto para lograr el objetivo de extracción, transformación, carga y presentación de datos.

- **Sistema transaccional NopCommerce y Microsoft SQL Server.**

  El sistema de la organización, consiste a grandes rasgos, de dos partes;

  La aplicación web NopCommerce: Esta es una aplicación web que provee interfaces de usuarios para los dueños del comercio electrónico de modo que puedan ingresar sus productos, precios, categorías, etc. Y de este modo poder modelar fielmente el catálogo de productos dentro de NopCommerce. La contraparte, es la interfaz para los usuarios clientes del comercio, donde pueden realizar funciones de registro de usuario, consulta, realización de pedidos, etc.

  Base de datos Microsoft SQL Server: Toda la información referente al comercio, sus clientes y sus compras se almacena en esta base de datos. El diseño de la base de datos está pensado para almacenar tanto, información sobre clientes, como nombre, correo, dirección; como para almacenar una serie de datos sobre información parametrizables referente a la configuración del comercio que ha sido customizado para adaptarse lo mejor posible a las necesidades del usuario dueño del comercio. Por lo cual, muchas de las tablas de la base de datos están destinadas a guardar las preferencias de configuración del comercio.

- **Carpeta dentro de sistema Operativo Windows**

  Para almacenar archivos de datos, se debe crear una carpeta con la siguiente estructura:

  NombreCarpeta

  Raw/

  Stagging/

  Dentro, se almacenarán los datos para ser procesados y subidos al almacenamiento en la nube S3.

- **AWS S3**

  Para la integración de servicios en la nube, se optó por la utilización de S3 (Simple Storage Service), un servicio de almacenamiento de Amazon web services. Debido a su seguridad y disponibilidad, pero principalmente por su costo, el cual es gratuito hasta cierto límite, dentro de la capa gratuita de AWS.

  Para la creación de un bucket es necesario seguir los pasos

  1. Debe iniciar sesión y dentro del panel de administración de AWS, abrir la consola de Amazon S3 en https://console.aws.amazon.com/s3/.
  2. Elija Create bucket (Crear bucket). Se abrirá el asistente Crear bucket (Crear bucket).
  3. En Bucket name (Nombre del bucket), escriba un nombre compatible con DNS para el bucket.
  4. En Region (Región), elija la Región de AWS en la que desea que se encuentre el bucket.
  5. En Object Ownership (Propiedad de objetos), para desactivar o habilitar las ACL y controlar la propiedad de los objetos cargados en el bucket.
  6. Elija la opción: Crear bucket.

  Posteriormente a la creación del bucket, para poder hacer uso de la solución propuesta se necesita de la siguiente estructura de carpetas dentro del bucket S3:

  BucketName

  01 Raw/

  03 Presentation/

- **IAM**

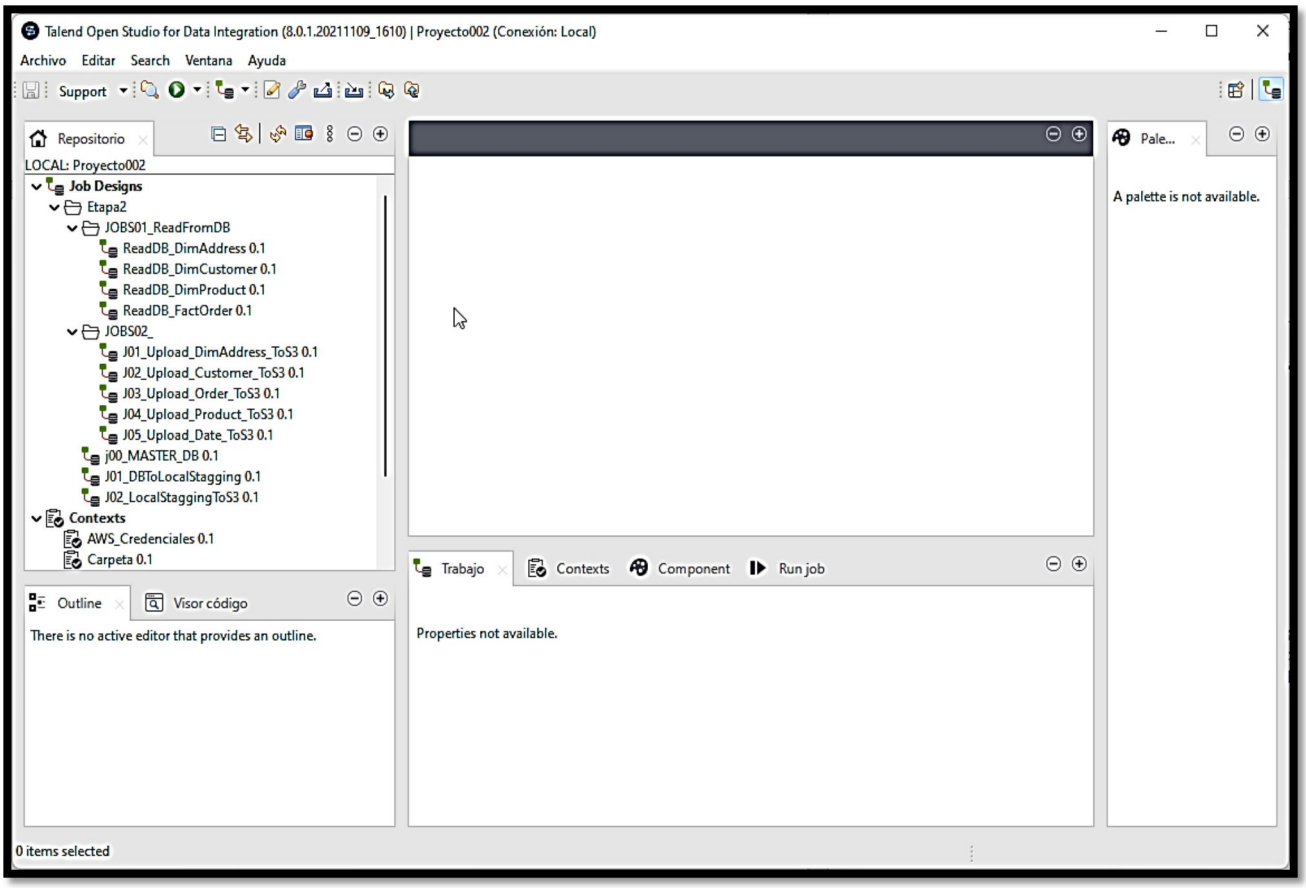
  Una vez creado el Bucket con su estructura de carpetas, es necesaria la creación de un usuario IAM y otorgar permisos de acceso al bucket. Si el usuario de IAM y el bucket de S3 pertenecen a la misma cuenta de AWS, puede conceder al usuario acceso a una carpeta de bucket específica mediante una política de IAM. Siempre que la política de bucket no deniegue explícitamente al usuario el acceso a la carpeta, no es necesario actualizar la política de bucket si la política de IAM concede acceso. Puede añadir la política de IAM a usuarios de IAM individuales o puede asociar la política de IAM a un rol de IAM al que puedan cambiar varios usuarios [2].

- **Talend Open Studio 8**

  Talend es un software desarrollado para la integración de datos, lo que lo hace una herramienta ideal para realizar procesos de extracción, transformación y carga (ETL) de grandes volúmenes de datos. Talend Open Studio 8 es la versión comunitaria que no implica ningún costo de licencia o plan de pago.

  Para la ejecución de la solución creada con Talend, es necesario usar el kit de desarrollo de Java (JDK) en su versión 11 o superior.

  Para hacer uso de la solución creada en Talend, al ejecutar Talend Open Studio 8, se debe seleccionar la opción "importa an existing Project" para crear un nuevo proyecto a partir del proyecto existente que comprende la solución creada; asignar un nombre adecuado; en el campo "Select root directory" se debe seleccionar la ruta de la carpeta donde se encuentra el proyecto solución; dar clic en el botón "Finish". Posteriormente, para abrir el proyecto, este debe seleccionarse dentro de la sección "Select an existing project" y buscar el proyecto con el nombre que fue asignado al importarlo y dar clic al botón "Finish", ver imagen 8 para interpretar



mejor la descripción.

Imagen 8: Pantalla principal de Talend Open Studio

En primer lugar, se debe realizar la configuración para establecer conexión con la base de datos. Para cada uno de los Jobs mostrados dentro de la carpeta **JOBS01_ReadFromDB**, se debe buscar el objeto de conexión y establecer dentro de la pestaña "Component" los parámetros necesarios para la conexión (ver imagen 9).

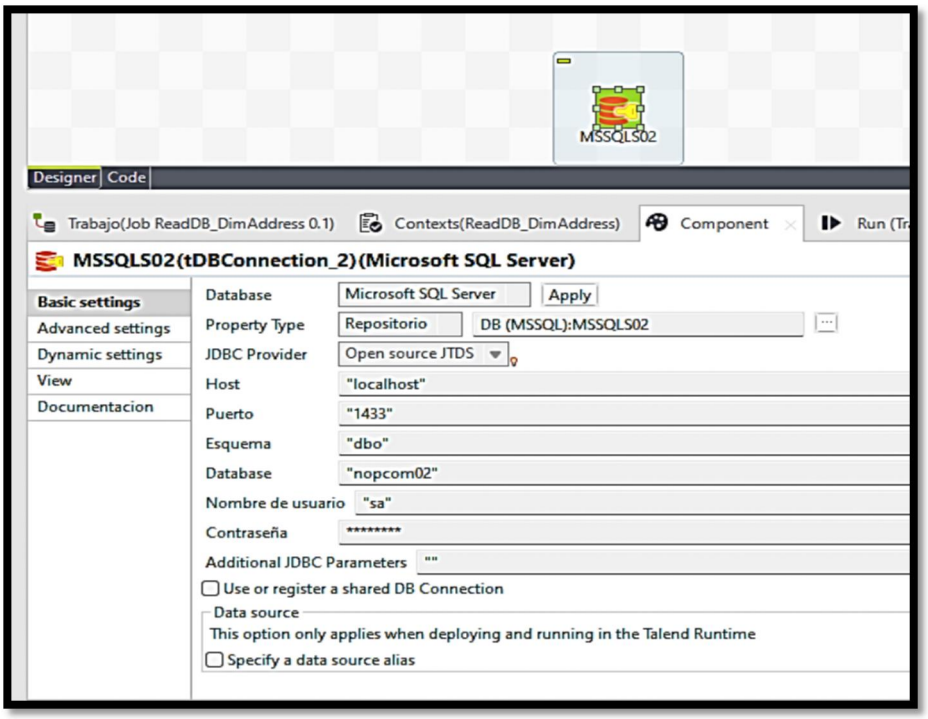


Imagen 9: pestaña de componentes (Component)

Cada uno de los cuatro Jobs anteriormente tratados, finaliza su flujo almacenando los datos procesados en un archivo delimitado por comas (ver imagen 10).

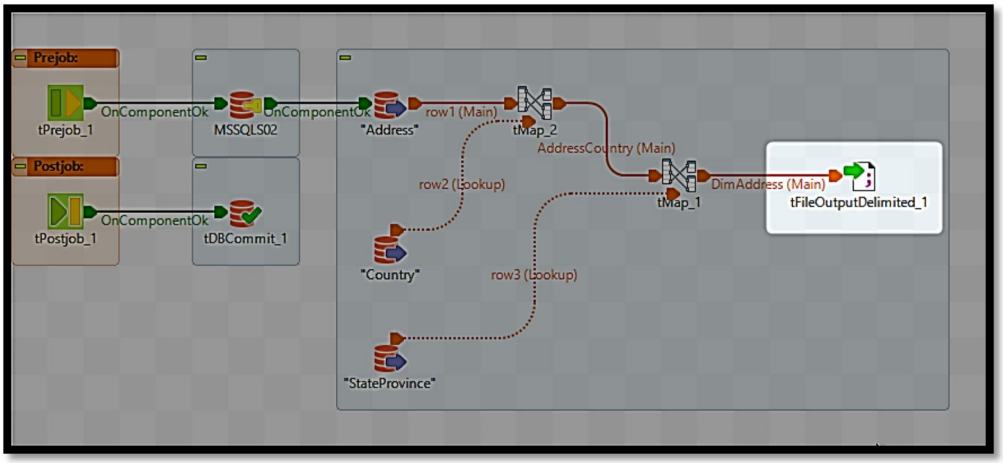


Imagen 10: flujo de almacenamiento

Este archivo será almacenado localmente. Por esto, es necesario haber realizado la configuración mencionada en el apartado anterior: **Carpeta dentro de sistema Operativo Windows** donde se dio nombre a las carpetas a utilizar para almacenar los archivos de datos. Deberá colocar la ruta de la carpeta local de almacenamiento en las variables de contexto respectivas, como se muestra en la imagen 11.

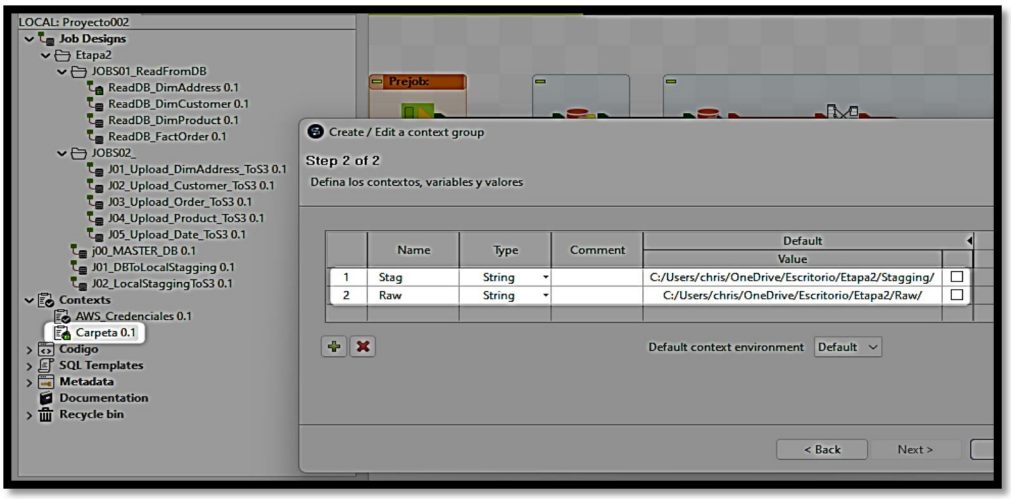


Imagen 11: especificación de ruta de carpeta local de almacenamiento.

La solución usa también otras variables de contexto con las cuales se especifica el acceso al bucket de AWS S3.

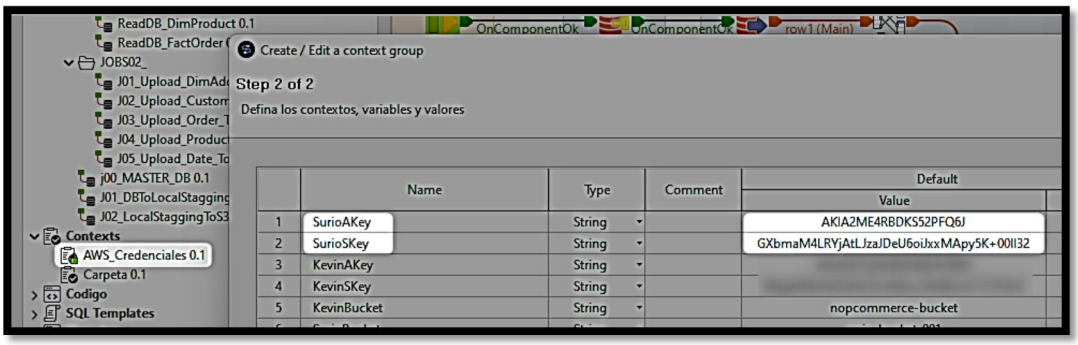


Imagen 12: variables para tener acceso al bucket de AWS S3.

En la imagen anterior debe editarse el valor de las dos variables de contexto:

- SurioAKey: Debe reemplazarse con la Access Key del usuario IAM que posee permisos en el bucket de S3
- SurioSKey: Debe reemplazarse con la Secret Key del usuario IAM respectivo que posee permisos en el bucket de S3.

Además, para acceder a una carpeta específica dentro del bucket de s3, deben especificarse los respectivos valores en las variables de contexto:

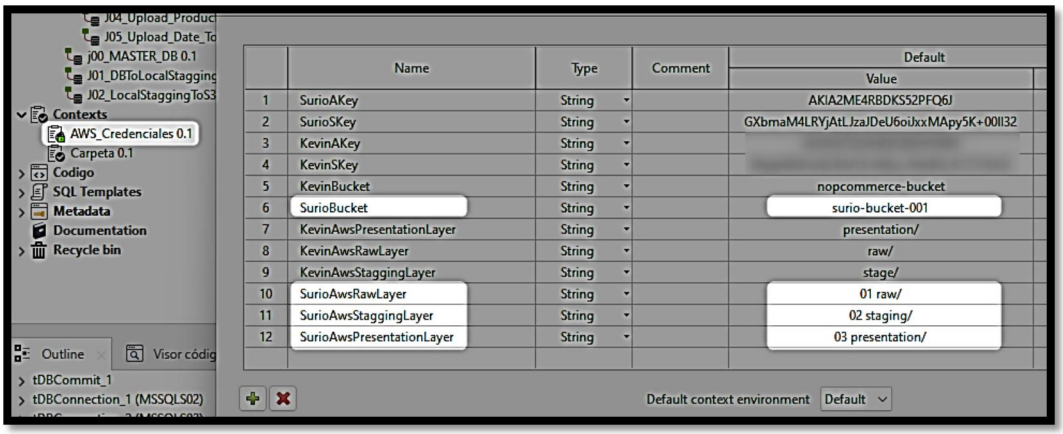


Imagen 13: Valores de variables para acceder al bucket.

En la imagen anterior debe editarse el valor de las cuatro variables de contexto:

- **SurioBucket:** Debe reemplazarse con el nombre del bucket de S3
- **SurioAwsRawLayer:** Debe reemplazarse con el nombre de la carpeta que será destinada a la capa de datos crudos (sin procesar) dentro del bucket de S3, seguido de una pleca.
- **SurioAwsStaggingLayer:** Debe reemplazarse con el nombre de la carpeta que será destinada a la capa Satgging dentro del bucket de S3, seguido de una pleca.
- **SurioAwsPresentationLayer:** Debe reemplazarse con el nombre de la carpeta que será destinada a la capa de presentación dentro del bucket de S3, seguido de una pleca.

- **AWS Redshift**

  Para montar el datawarehouse sobre la plataforma de Amazon, se debe acceder a Amazon Redshift y realizar la creación de un nuevo cluster de 2 cpu's virtuales y 160 GB de almacenamiento por nodo.

  Para la creación de las tablas de RedShift, debe accederse mediante el editor de consultas de RedShift. Se debe conectar al cluster que se ha creado. Dentro se debe crear una base de datos para alojar las tablas del modelo dimensional que comprende el Datawarehouse.

Para la creación de las tablas del Datawarehouse debe ejecutarse el script “creadb” que se encuentra en el anexo 1.

Las tablas del datawarehouse, deben poblarse con datos a partir de los archivos CSV que se encuentran en S3 y que son el resultado del proceso de transformación hecha con Talend, Para esto debe ejecutarse el script que se encuentra en el anexo 2, en el que debe sustituirse los parámetros de las sentencias SQL por la información concerniente a cada dimensión y tabla de hechos que conforman el modelo dimensional del datawarehouse.

El formato es el siguiente:

COPY ***nombreDeLaTabla***
FROM 's3://***nombreDelBucket***/03 presentation/***NombreDeArchivo***.csv'
credentials
'aws_access_key_id=***ValorDeAccessKey***; aws_secret_access_key=***ValorDeSecretKey***'
ignoreheader 1
CSV;

- **Power BI**

Para poder hacer uso del reporte, en Power BI, debe seleccionar la opción “Abrir informe” y dar clic en el botón “Examinar informes”; se deberá navegar entre los archivos, hasta ubicar el archivo con extensión .pbix correspondiente a la solución elaborada.

Para actualizar el origen de los datos, se debe ingresar al menú inicio > Transformar datos > Configuración de origen de datos. Ver imagen 14.

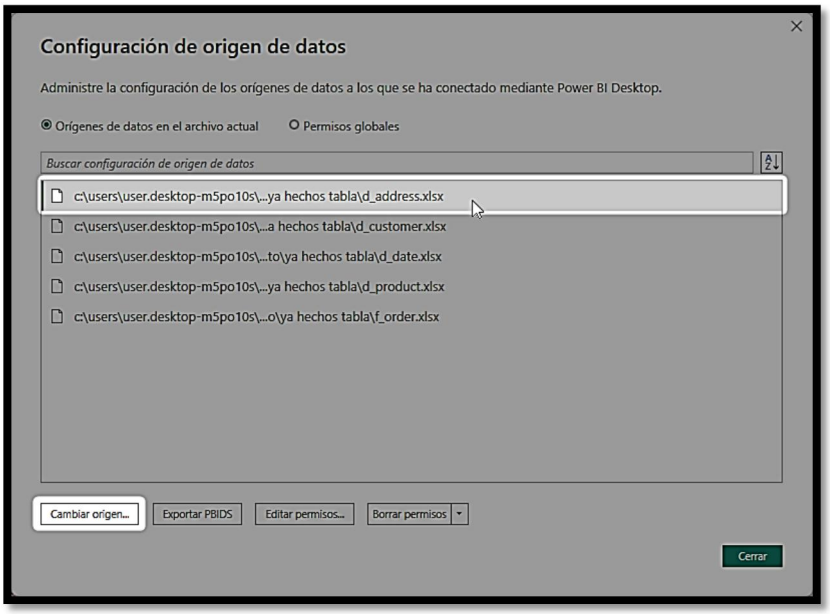


Imagen 14: configuración de origen de datos.

Se debe seleccionar uno de los archivos y dar clic al botón “Cambiar origen..”. Seleccionar la opción “Libro Excel” en el campo “Abrir archivo como”; luego, en el campo “Ruta de acceso de archivo, dar clic en el botón “Examinar” y navegar hasta el archivo correspondiente dentro de la carpeta de *datos para Power BI*. Se debe realizar esto para cada uno de los 5 archivos. Luego cerrar la ventana de configuración de origen de datos y aplicar los cambios.

b. Presupuesto de implementación

Actualmente, la organización Almacén Papagayo cuenta con una infraestructura capaz de almacenar una cantidad considerablemente grande de datos. Estos equipos corresponden al sistema transaccional que comprende el servidor donde se ejecuta el sistema de comercio en línea, NopCommerce, y un servidor de base de datos que persiste los datos del sistema transaccional. Sin embargo, No es conveniente interferir en el rendimiento del sistema transaccional, asignándole la carga del procesamiento de datos para la generación de reportes.

Recurso humano

Se calcula que la integración de la solución con el sistema informático de la organización Almacenes Papagayo, se realice a lo largo de un periodo de dos días, realizado por una sola persona. Para el cálculo del costo del recurso humano, se tomará el salario base de un analista programador del sector público, el cual es de $1000.00 mensuales (luego de descuentos equivale a un salario líquido de $837.00 mensuales). A partir de esta cifra, se calcula el salario diario del recurso humano en $33.33 por día o $4.167 por hora. Ver tabla 1.

Cantidad de recursos humanos	Cantidad de horas requeridas para la integración	Precio por hora de trabajo de un recurso humano	Costo total en recursos humanos
1	16 horas	$4.167	$66.67

Tabla 1: recurso humano estimado

Recurso tecnológico

Debido a que, en primer lugar, para el diseño de la solución se ha optado por utilizar herramientas de software que pueden funcionar de manera gratuita; y, en segundo lugar, la organización ya cuenta con licencias para el sistema operativo utilizado, además de contar con equipo de hardware suficiente para la instalación de la solución; No se incurre en gasto tecnológico para realizar la implementación de la solución.

Servicios básicos

La organización Almacenes Papagayo, ya cuenta con servicios de red de conexión a internet. Además, para la creación de la cuenta necesaria para la integración de los servicios de AWS no necesita más que el registro. Por lo que, en cuanto a servicios básicos, para la implementación de la solución, no se incurre en gastos.

c. Análisis de resultados

Conexión de Talend con la base de datos.

Los objetos que retornan las tablas necesarias para realizar la transformación de datos se encuentran dentro de la carpeta Metadata > Db Connections > MSSQLS02 0.1 > Table schemas.

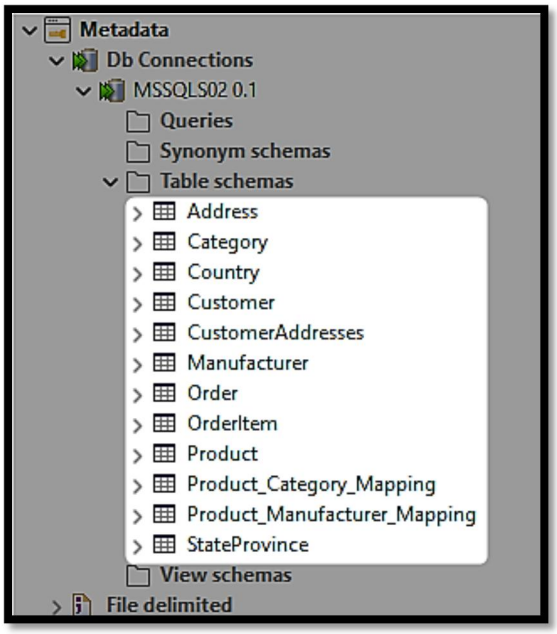


Imagen 15: diagrama de árbol de la carpeta Table schemas

A partir de estas tablas del esquema, se diseña la estructura de los Jobs que transforman dichos esquemas en las dimensiones y la tabla de hechos del modelo dimensional planteado.

Imagen 16: estructura de los Jobs de lectura.

Jobs de transformación de datos.

Para transformar los esquemas de Direcciones, País y Provincia, en la dimensión Address

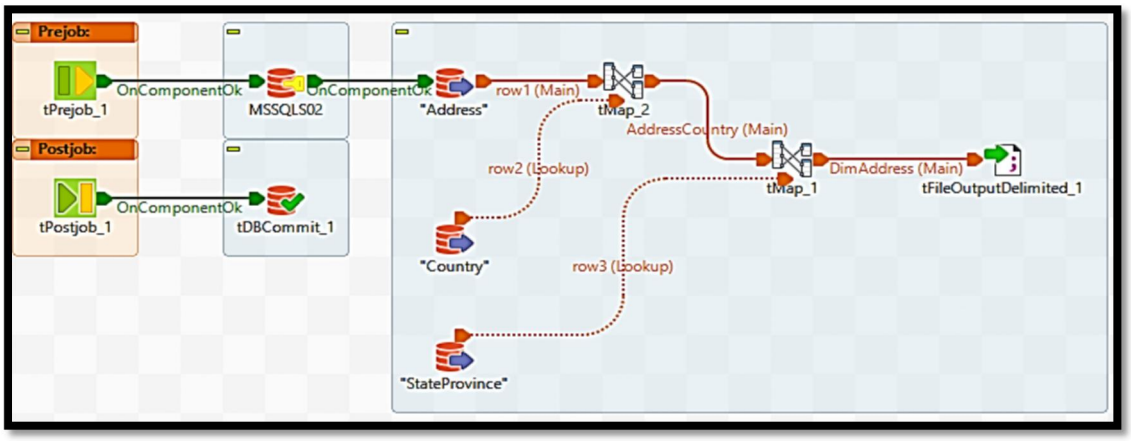


Imagen 17: diagrama de Jobs de transformación a la dimensión Address.

Para transformar los esquemas de Clientes, Direcciones de clientes y Dirección, en la dimensión Customer. Ver imagen 18.

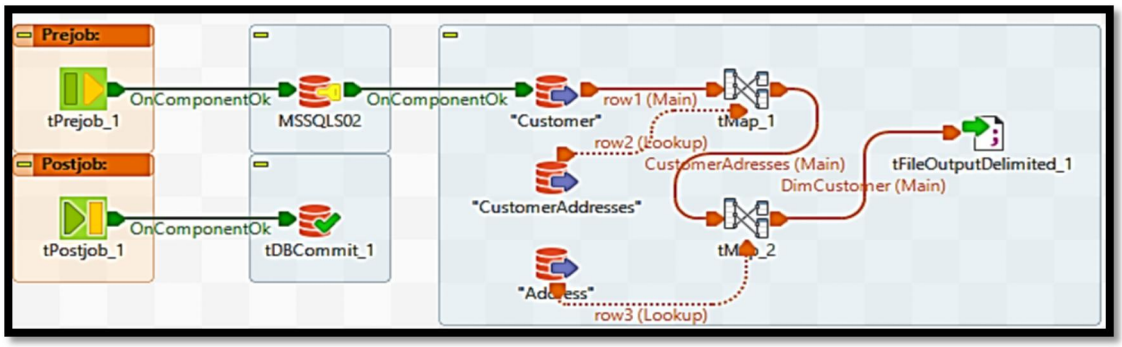


Imagen 18: diagrama de Jobs para crear dimensión Customer.

Para transformar los esquemas referentes a las líneas de compra y órdenes. Ver imagen 19.

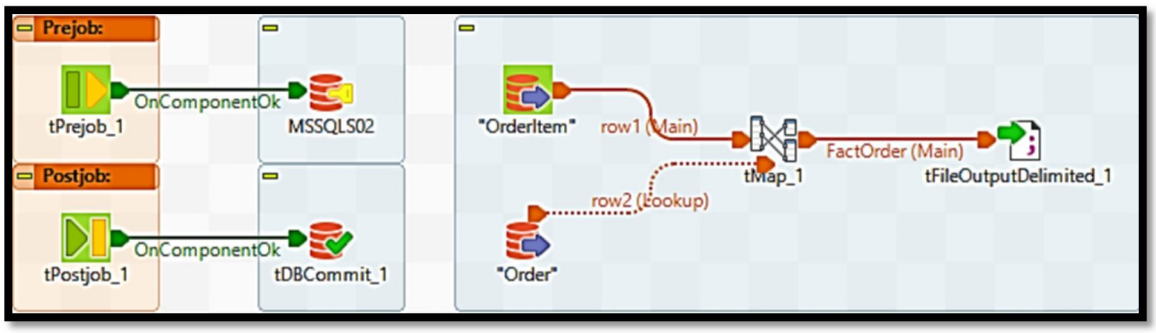


Imagen 19: diagrama de Jobs compra y órdenes.

Para transformar los esquemas de Productos, Producto_Categoría, Categorías, Producto_Fabricante y Fabricantes, en la dimensión Product. Ver imagen 20.

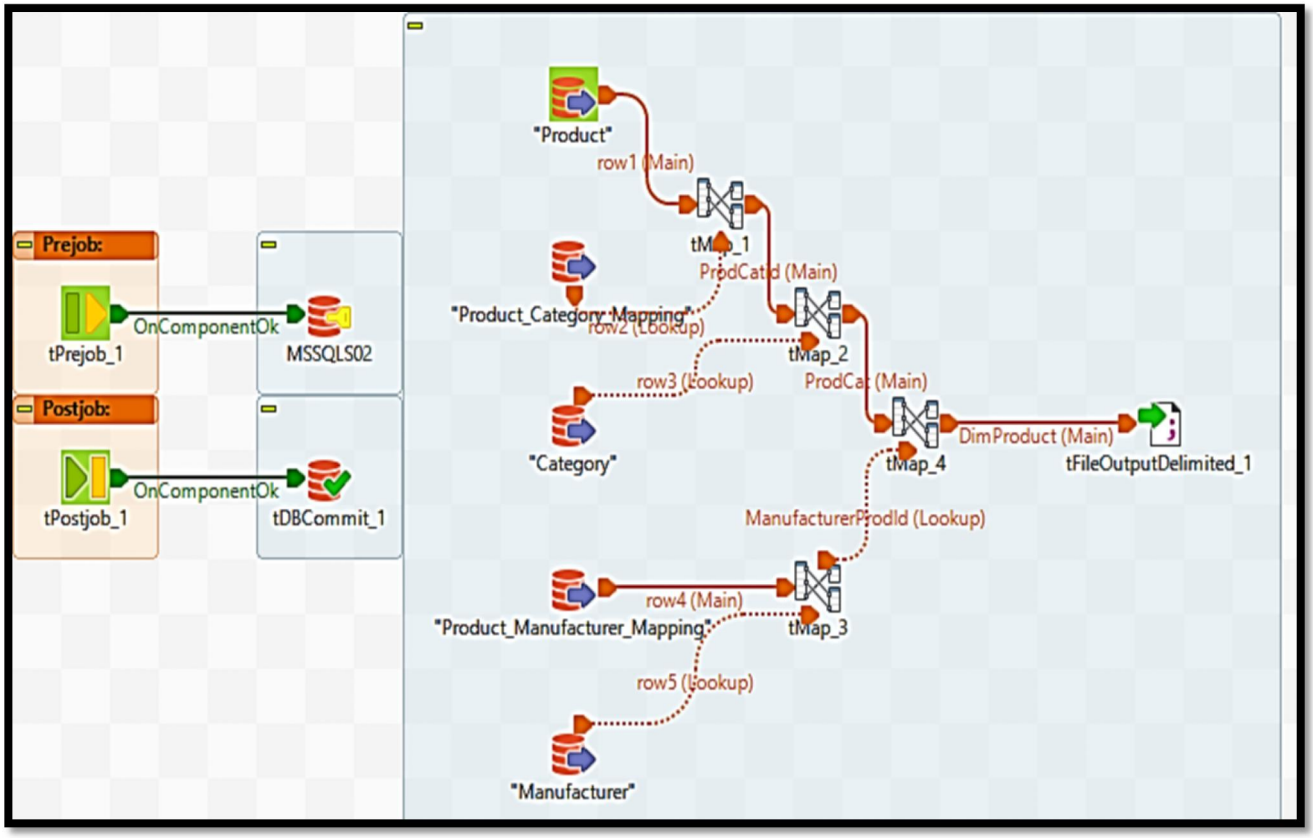


Imagen 20: diagrama de Jobs para dimensión Product.

Para ejecutar los anteriores Jobs a través de un Job Padre. Ver imagen 21.

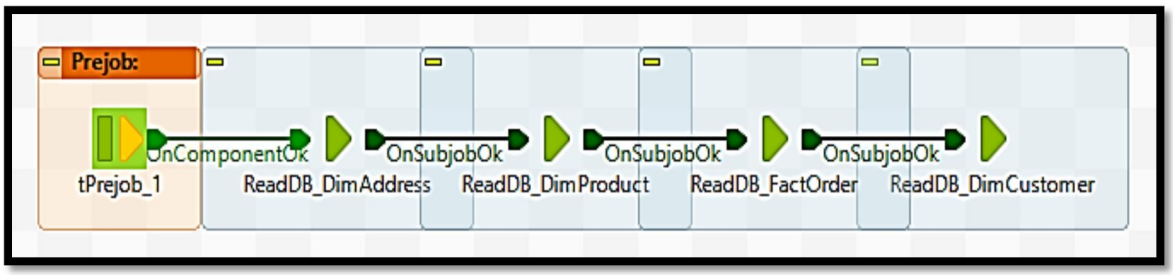


Imagen 21: diagrama de Jobs para ejecutar Job Padre.

La tabla de hechos y sus respectivas dimensiones, a excepción de la dimensión Date, son transformadas a partir de la base de datos transaccional y almacenadas en la carpeta local Staging:

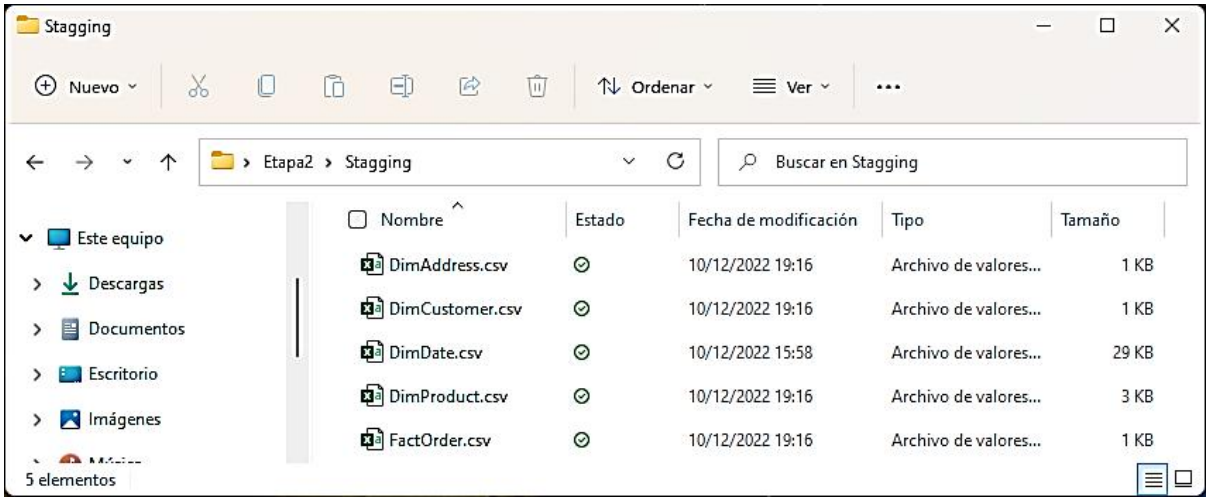


Imagen 22: Carpeta local Stagging

Desde esta carpeta local, los archivos CSV resultantes del proceso de transformación, son subidos a S3 mediante otro conjunto de Jobs:

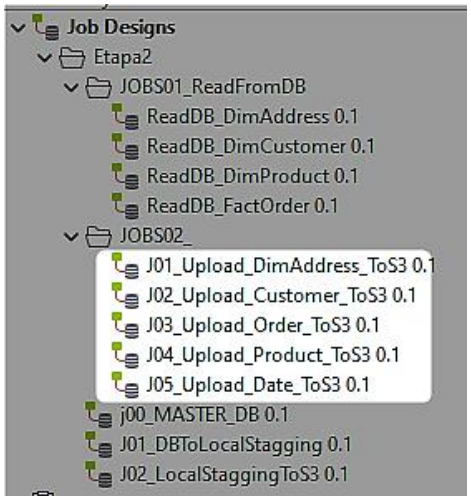


Imagen 23: Carpeta Talend, Jobs de carga de archivos a S3

Cada Job se realiza la conexión con el bucket de S3 y sube el respectivo archivo csv a la carpeta de presentación de S3:

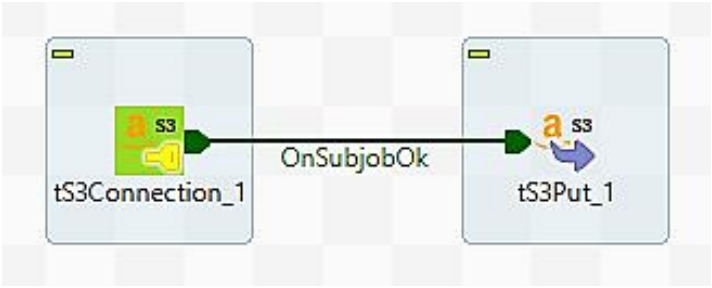


Imagen 24: Elementos del Job de carga a S3

Cada archivo es subido individualmente por un Job y cada Job es orquestado por un Job padre:

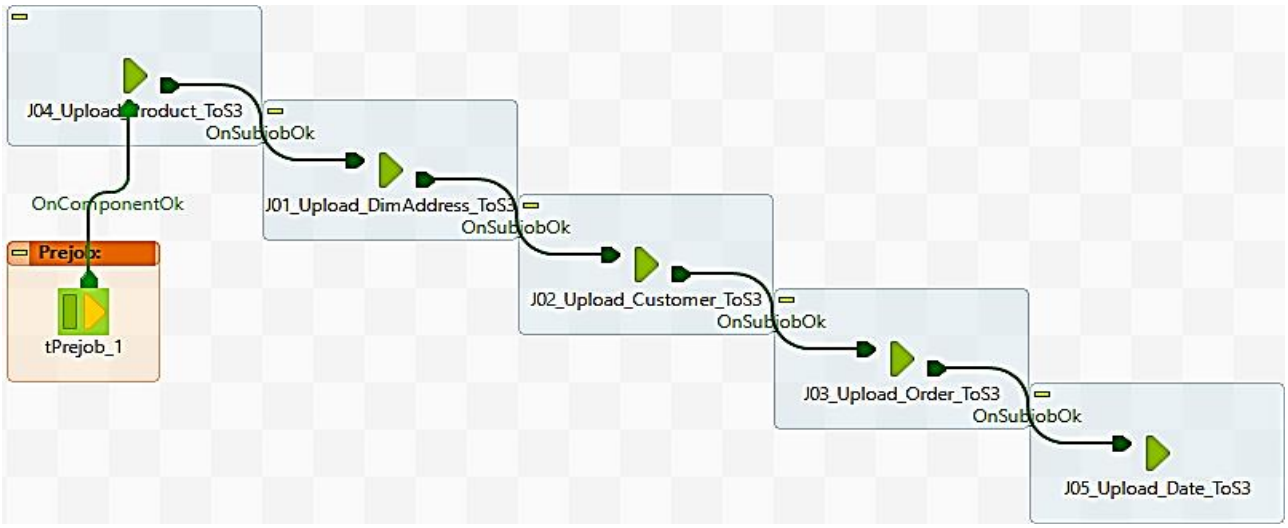


Imagen 25: Job Padre, secuencia de ejecución de Jobs de carga a S3

Una vez almacenados los archivos dentro de la respectiva carpeta de S3, a partir de los scripts de los Anexos 1 y 2, se logra estructurar y poblar el datawarehouse creado en Amazon RedShift.

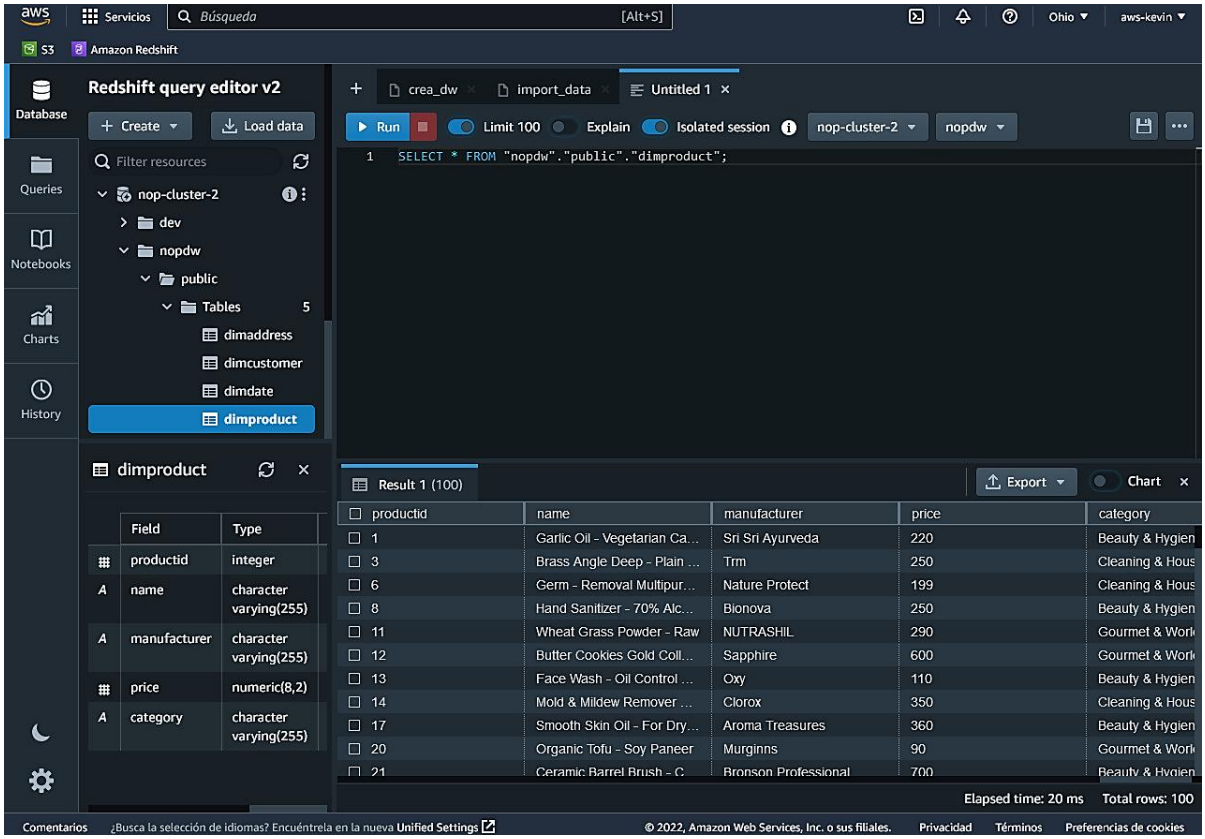


Imagen 26: Vista previa de DB en Redshif

Los datos procesados correspondientes al modelo dimensional son el insumo para la visualización de reportes dentro de Power BI.

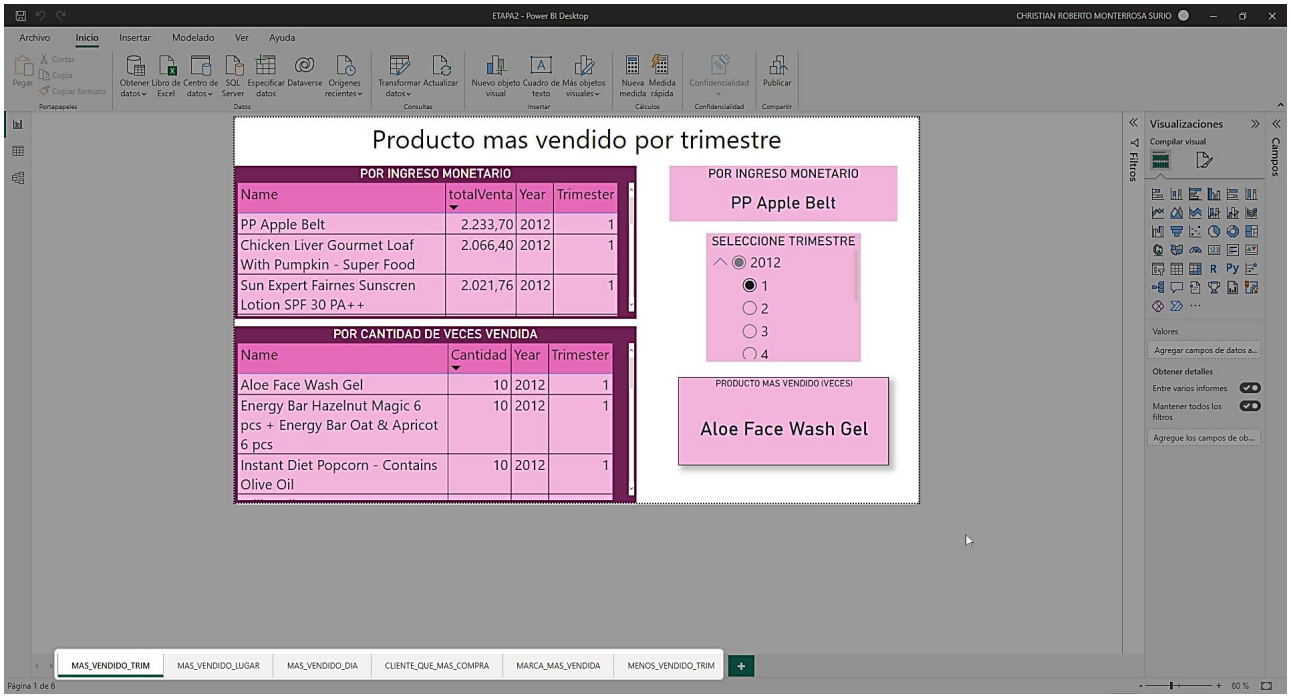


Imagen 27: Vista previa de reporte en Power BI

Conclusiones y recomendaciones

Conclusiones:

- Se realizó un análisis sobre el comercio electrónico y su respectiva base de datos transaccional, identificando las tablas del esquema relevantes para formar el Datalake que será procesado.
- Aplicando los conocimientos teóricos sobre Datawarehouse adquiridos en el curso de especialización de Business intelligence, se ha logrado realizar un diseño de modelo dimensional conocido como modelo de estrella, que representa la estructura del Datawarehouse del comercio electrónico. El diseño fue construido a partir de la estructura de la base de datos transaccional y tomando en cuenta las métricas y necesidades de información de los dueños del negocio.
- Utilizando la herramienta Talend, se ha construido un proceso ETL que establece conexión con la base de datos transaccional extrayendo los esquemas relevantes; se han realizado procesos que transforman los datos extraídos en Dimensiones del modelo dimensional; y posteriormente se han realizado procesos para la carga de datos a al repositorio en Amazon S3..
- Con el desarrollo del prototipo presentado en el presente documento, se ha logrado solventar una necesidad de información para la toma de decisiones sobre un comercio electrónico, mediante aplicación de metodologías y técnicas para la construcción de almacenes de datos, conocidos como Datawarehouse.

Recomendaciones:

- En el proyecto, se hizo uso de tecnologías en la nube, específicamente AWS, que ofrece grandes ventajas por su seguridad y disponibilidad. Sin embargo, para una pequeña o mediana empresa, estos servicios representan un costo considerablemente alto. Por lo que se recomienda buscar otras alternativas como el uso de equipos propios.
- Las herramientas utilizadas son muy efectivas al momento de hacer un uso óptimo de ellas. Sin embargo, también poseen limitantes por lo que requieren una versión de paga si se desea otro tipo de complemento.
- Poseer hardware de gama media-alta para soportar el procesamiento masivo de datos.
- Al ser una empresa ficticia, los datos no son en su totalidad realistas, por lo cual se recomienda usar la ciencia de datos para organizaciones reales ya que fueron generados, en su gran mayoría, por medio de un generador de datos.
- Estos análisis se deben realizar en un periodo determinado. Sin embargo, esto se deja a criterio del jefe o coordinador encargado de la organización, ya que según a lo que ésa se dedique, así será el periodo que se efectúen dichos estudios y análisis (por temporadas, por meses o por alguna fecha en especial).

Bibliografía

[1] *Kimball, R., & Ross, M. (2013). The data warehouse toolkit (3rd ed.). Wiley.*

[2] *nopCommerce Documentation*. (2022, 23 noviembre). https://docs.nopcommerce.com/en/index.html

[3] Talend. (2022, 27 septiembre). *Talend Open Studio: Open-source ETL and Free Data Integration*. Talend - A Leader in Data Integration & Data Integrity. https://www.talend.com/products/talend-open-studio/

[4] *AWS | Almacenamiento de datos seguro en la nube (S3)*. (s. f.). Amazon Web Services, Inc. https://aws.amazon.com/es/s3/

[5] *Introducción a Data Warehousing on AWS con Amazon Redshift (2:07)*. (s. f.). Amazon Web Services, Inc. https://aws.amazon.com/es/redshift/

[6] *¿Qué es Power BI? Definición y características | Microsoft Power BI*. (s. f.). https://powerbi.microsoft.com/es-es/what-is-power-bi/

[7] *Paso 1: Crear tu primer Becket de S3 - Amazon Simple Storage Ser vice*. (s. f.). https://docs.aws.amazon.com/es_es/AmazonS3/latest/userguide/creating-bucket.html

[8] Official, A. (2022, 30 septiembre). *How can I grant a user access to a specific folder in my Amazon S3 bucket?* Amazon Web Services, Inc. https://repost.aws/knowledge-center/s3-folder-user-access

[9] JMB Auditores y Consultores. (2022, 26 enero). *Calcular prestaciones*. https://jmbauditores.com/calcular-prestaciones/

Glosario

A

Analistas de datos: persona encargada de recopilar grandes cantidades de datos, procesarlos y generar informes de forma estadistica con el fin de crear estrategias en beneficio de la empresa, ... 9

ASP.NET: plataforma de desarrollo compuesta de herramientas, lenguajes de programacion y librerias que permiten desarrollar diferentes tipos de aplicaciones, ... 19

AWS: siglas para Amazon Web Services, es una plataforma que ofrece servicios de computacion en la nube, ... 19

B

Big Data: concepto que se refiere a grandes cantidades de datos structurados y no estructurados provenientes de distintos tipos de fuentes los cuales requieren aplicaciones informaticas dedicadas para su tratamiento, ...9

Bucket: proviene de Amazon S3, un bucket es un contenedor capaz de almacenar cualquier tipo de objeto, ...28

C

Clúster de Amazon Redshift: un cluster es un grupo de nodos que tienen un nodo lider encargado de crear planes de acción que son ejecutados en paralelo por los nodos de computo cuando llega una query, ...34

CSV: siglas provenientes del inglés para “Comma Separated Values”, los cuales son archivos de texto plano que contienen datos separados por comas , ... 29

D

Dashboard: en el contexto de Power BI son cuadros de mando o visualizaciones de un conjunto de datos lo que apoya a la toma de decisiones en la empresa, ...19

Data Warehouse: es un almacen de datos electronico de gran capacidad capaz de almacenar datos de forma segura, fiable, facil de recuperar y administrar , ... 9

Datos no estructurados: datos que no se almacenan en una base de datos estructurada como pueden ser: audios, archivos de texto plano, datos geoespaciales,etc, ... 9

Dimensiones: en el modelado dimensional, las dimensiones denotan el contexto del hecho que se esta queriendo analizar, se enlazan a la tabla de hechos a traves de un campo clave, ...24

DW: siglas utilizadas para Data Warehouse, ... 28

E

Esquema estrella: en el modelado dimensional, un esquema estrella esta conformado por una tabla de hechos al centro y varias tablas de dimensiones enlazadas a la tabla de hechos, ... 25

ETL: siglas proveniente del inglés “Extract Transform Load” utilizadas para los procesos de estraccion, transformacion y carga de datos, ...13

G

Granularidad: en el modelado dimensional, la granularidad define el nivel minimo de detalle de los datos con que cuenta el sistema fuente, ...24

I

IAM: siglas provenientes del ingles para “Identity and Access Management”, servicio de AWS para la administracion de identidades y acceso a los recursos de AWS de forma segura, ...36

Ingeniería de Datos: se encarga de construir y mantener la estructura de datos y las arquitecturas necesarias que alimentan aplicaciones que utilizan grandes cantidades de datos,..... 14

Inteligencia de negocios: tambien abreviado BI (del inglés Business Intelligence) conjunto de estrategias y herramientas utilizadas para la mejora en la toma de decisiones basadas en el analisis de los datos, ... 34

J

JDK: siglas provenientes del inglés “Java Development Kit”, un conjunto de herramientas de software para desarrollar programas en lenguaje Java, ...37

M

Métricas: como parte del análisis dimensional, las metricas definen qué es lo que el usuario quiere medir en funcion de un atributo numerico presente en la tabla de hechos, ... 24

Modelado Dimensional: conjunto de tecnicas y conceptos utilizado para el diseño de almacenes de datos

N

P

R

S

T

Anexos

Anexo 1. Script para creación de base de datos en RedShift: **creadb.txt**

```
create table dimcustomer(
   CustomerId integer,
   Email varchar(255),
   Name varchar(255),
primary key(CustomerID) );

create table dimdate(
   DateId integer,
   Day integer,
   Month integer,
   Year integer,
   DayName varchar(255),
   MonthName varchar(255),
   Semester integer,
   Trimester integer,
primary key(DateId) );

create table dimproduct(
   ProductId integer,
   Name varchar(255),
   Manufacturer varchar(255),
   Price decimal (82),
   Category varchar(255),
primary key(ProductId) );

create table dimaddress(
   AddressId integer,
   Country varchar(255),
   City varchar(255),
   StateProvince varchar(255),
   County varchar(255),
   ZipPostalCode varchar(255),
primary key(AddressId) );

create table factorder(
   OrderKey integer,
   AddressKey integer,
   ProductKey integer,
   CustomerKey integer,
   Quantity integer,
   Price decimal (8,2),
   DateKey integer,
primary key(OrderKey),
foreign key(AddressKey) references
dimaddress(AddressId),
foreign key(ProductKey) references
dimproduct(ProductId),
foreign key(CustomerKey) references
dimcustomer(customerid),
foreign key(DateKey) references
dimdate(dateid) );
```

Anexo 2.

```
COPY dimaddress
FROM 's3://surio-bucket-001/03
presentation/DimAddress.csv'
credentials
'aws_access_key_id=AKIA2ME4RBDKS52P
FQ6J;aws_secret_access_key=GXbmaM4LR
YjAtLJzaJDeU6oiJxxMApy5K+00lI32'
ignoreheader 1
CSV;

COPY dimproduct
FROM 's3://surio-bucket-001/03
presentation/DimProduct.csv'
credentials
'aws_access_key_id=AKIA2ME4RBDKS52P
FQ6J;aws_secret_access_key=GXbmaM4LR
YjAtLJzaJDeU6oiJxxMApy5K+00lI32'
ignoreheader 1
CSV;

COPY dimdate
FROM 's3://surio-bucket-001/03
presentation/DimDate.csv'
credentials
'aws_access_key_id=AKIA2ME4RBDKS52P
FQ6J;aws_secret_access_key=GXbmaM4LR
YjAtLJzaJDeU6oiJxxMApy5K+00lI32'
ignoreheader 1
CSV;

COPY dimcustomer
FROM 's3://surio-bucket-001/03
presentation/DimCustomer.csv'
credentials
'aws_access_key_id=AKIA2ME4RBDKS52P
FQ6J;aws_secret_access_key=GXbmaM4LR
YjAtLJzaJDeU6oiJxxMApy5K+00lI32'
ignoreheader 1
CSV;

COPY factorder
FROM 's3://surio-bucket-001/03
presentation/FactOrder.csv'
credentials
'aws_access_key_id=AKIA2ME4RBDKS52P
FQ6J;aws_secret_access_key=GXbmaM4LR
YjAtLJzaJDeU6oiJxxMApy5K+00lI32'
ignoreheader 1
CSV;
```

Anexo 3.

Depreciación de una computadora de gamma media

Valor residual = Valor inicial / vida útil en años

Valor residual = $600 / 10 = $60

Depreciación = (valor inicial - valor residual) / vida útil en años

Depreciación = ($600 - $60) / 10 = $54

Anexo 4.

Cronograma de actividades

ACTIVIDADES	MARZO				ABRIL				MAYO				JUNIO				JULIO				AGOSTO				SEPTIEMBRE				OCTUBRE				NOVIEMBRE				DICIEMBRE			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Definición y ajuste de propuesta.																																								
Modelado dimensional (definición de granularidad, dimensiones y métricas).																																								
Diseño del Datawarehouse.																																								
Diseño del ETL.																																								
Configuraciones en AWS S3.																																								
Diseño del Data Profiling.																																								
Configuración de usuario y roles en AWS IAM.																																								
Configuración en AWS Redshift.																																								
Instalación y configuración de herramientas de software. (TOP y PBI)																																								

Revisión y configuraciones extras de las herramientas de software.																																								
Preparación y transformación de datos según métricas en Power BI																																								
Presentación de datos transformados de manera tabulada (Power BI)																																								

Anexo 4.

Nº	Actividad o Recurso	Descripción	cantidad	Precio unitario (dólares)	Costo total (dólares)
1	Pasajes, viáticos, etc.	Movilidad para buscar la organización y actividades similares (alimentación a la hora de trabajar en el proyecto). Costo estimado diario.	10 días	5	50
2	Energía eléctrica	Para las computadoras y dispositivos extra con las que se desarrolló el proyecto. Costo estimado mensual por miembro del equipo.	9 meses	10	90
3	Hardware	Memoria RAM 8GB extra por demanda del software utilizado. Costro aproximado.	2	50	100
4	AWS Redshift	Utilizada para bases de datos en la nube (se usó más tiempo de la capa gratuita y aumenta su cobro por hora) 0,25 USD por hora según documentación de AWS. Costo aproximado.	40 horas	0.25	10
5	AWS S3	Utilizada para almacenamiento en la nube (se usó por 1 mes fuera de la capa gratuita) 0,023 USD por GB en los Primeros 50 TB/mes (según documentación de AWS)	1 Gb	0.023	0.023
6	Depreciación de computadoras y periféricos.	Monto calculado por dispositivo (monto aproximado por miembro del equipo y promediado. Ver anexo 3)	5 pc	54	270
7	Servicio de internet residencial	Internet utilizado durante 9 meses con una capacidad promedio de 10 Mbps (valor aproximado por mes).	18 meses (9 por 3 miembros)	28	84