

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



**DESARROLLO DE DATAWAREHOUSE PARA EL MÓDULO DE COMPRAS DE LA
EMPRESA LIBRERÍA DEL PACIFICO**

PRESENTADO POR:

JOSÉ ERNESTO HERNÁNDEZ VÁSQUEZ

MARCOS ALEXANDER PARRILLAS ORELLANA

PARA OPTAR AL TÍTULO DE:

INGENIERO DE SISTEMAS INFORMÁTICOS

UNIVERSIDAD DE EL SALVADOR

RECTOR:

MSC. ROGER ARMANDO ARIAS ALVARADO

SECRETARIO GENERAL:

MSC. FRANCISCO ANTONIO ALARCÓN SANDOVAL

FACULTAD DE INGENIERÍA Y ARQUITECTURA

DECANO:

PHD. EDGAR ARMANDO PEÑA FIGUEROA

SECRETARIO:

ING. JULIO ALBERTO PORTILLO

ESCUELA DE INGENIERÍA DE SISTEMAS INFORMÁTICOS

DIRECTOR:

ING. RUDY WILFREDO CHICAS VILLEGAS

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

Trabajo de Graduación previo a la opción al Grado de:
INGENIERO DE SISTEMAS INFORMÁTICOS

Título:

**DESARROLLO DE DATAWAREHOUSE PARA EL MÓDULO DE COMPRAS DE LA
EMPRESA LIBRERÍA DEL PACIFICO**

Presentado por:

JOSÉ ERNESTO HERNÁNDEZ VÁSQUEZ

MARCOS ALEXANDER PARRILLAS ORELLANA

Trabajo de Graduación Aprobado por:

Docente Asesor:

ING. RENE FABRICIO QUINTANILLA GÓMEZ

SAN SALVADOR, ENERO 2023

Trabajo de Graduación Aprobado por:

Docente Asesor:

Ing. Rene Fabricio Quintanilla Gómez

Diccionario de datos	1
Introducción	3
Capítulo I: Especificación de proyecto.....	5
a. Situación actual	5
I. Antecedentes	5
II. Descripción del problema	8
III. Planteamiento del problema.....	10
b. Objetivos.....	11
I. Objetivo General.....	11
II. Objetivos Específicos	11
c. Alcances	12
d. Justificación	13
Capitulo II: Análisis y diseño de la propuesta de solución	14
a. Metodología de trabajo	14
Data Warehousing y Business Intelligence	14
Introducción al Modelado Dimensional	16
Esquema de Estrella	18
Arquitectura de Kimball: Data Warehousing y Business Intelligence	22
b. Descripción de la propuesta de solución	26
c. Descripción de la tecnología a utilizar	28
PostgreSQL	28
DBeaver	29
Talen Open Studio	29
Amazon S3	30
Amazon Redshift.....	31
Amazon IAM	32
Power BI	33
d. Diagrama arquitectónico de la solución	35
e. Descripción de cada componente de la solución.....	36
Capitulo III: Estrategia de implementación de propuesta de solución	38
a. Estrategia de implementación.....	38

b. Periodicidad y mitigación de la Ejecución de Procesos ETL.....	56
Periodicidad	56
Mitigación.....	56
c. Análisis de resultados.....	57
Especificación de Necesidades que el Modelo Propuesto resolverá.	57
Conclusiones y recomendaciones	65
Conclusiones.....	65
Recomendaciones.....	66
Bibliografía.....	67
Anexos.....	68
a. BPMN Proceso de Compra.....	68
b. Cronograma de actividades	69
c. Sucursales de Librería el Pacífico	70
d. Presupuesto de Desarrollo	70
e. Presupuesto de implementación.....	73

Diccionario de datos

Datawarehouse (DW): Es un almacén de datos donde generalmente una empresa u organización mantiene una gran cantidad de información. Los datos de un data warehouse deben almacenarse de forma segura, fiable, fácil de recuperar y fácil de administrar.

Data Lake (Lago de datos): Es un repositorio de almacenamiento que contienen una gran cantidad de datos en bruto y que se mantienen allí hasta que sea necesario.

Big Data: Cuando hablamos de Big Data nos referimos a conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales, tales como bases de datos relacionales y estadísticas convencionales o paquetes de visualización, dentro del tiempo necesario para que sean útiles.

ODOO: Es el software libre de gestión empresarial capaz de cubrir todas las necesidades de un negocio, gracias a la integración de sus múltiples aplicaciones: contabilidad, inventario, marketing online, gestión de proyectos, recursos humanos, etc.

Dashboard: Son cuadros de mando que sirven para representar datos de forma visual, permite transformarlos en algo fácil de visualizar para su mejor interpretación, comprensible para cualquier persona que tenga acceso a la información y deba analizar las conclusiones y tomar decisiones.

Amazon S3: Es un servicio ofrecido por Amazon Web Services (AWS) que proporciona almacenamiento de objetos a través de una interfaz de servicio web.

Bucket: Es el servicio de almacenamiento de objetos de Amazon Web Services (AWS). Es la solución más común para almacenar datos en la nube de una forma segura, eficiente y escalable.

Redshift: Es un servicio de Amazon Web Services (AWS) de data warehouse gestionado que cuenta con varios petabytes de almacenamiento para sus clientes en la nube.

IAM: Es un servicio web que ayuda a controlar de forma segura el acceso a los recursos de AWS. Con IAM, puede administrar de forma centralizada los permisos que controlan a qué recursos de AWS pueden acceder los usuarios.

ETL: La palabra ETL corresponde a las siglas en inglés de:

- Extraer: extract.
- Transformar: transform.
- Y Cargar: load.

Con ello, queremos decir que todo proceso ETL consta precisamente de estas tres fases: extracción, transformación y carga de datos.

Business Intelligence (BI): Es un tipo de software que se alimenta de datos de negocios y presenta reportes, paneles, tablas y gráficos de forma amigable para el usuario.

Tabla de hechos: Es la tabla central de un esquema dimensional (en estrella o en copo de nieve) y contiene los valores de las medidas de negocio o dicho de otra forma los indicadores de negocio.

Tabla de Dimensiones: Las tablas de dimensiones son complementos integrales de una tabla de hechos. Las tablas de dimensiones contienen el contexto textual asociado con un evento de medición de procesos comerciales. Describen el "quién, qué, dónde, cuándo, cómo y por qué" asociado con el evento.

Esquema de estrella: Es un modelo de datos que tiene una tabla de hechos (Fact Table) que contiene los datos para el análisis, rodeada de las tablas de dimensiones.

Introducción

Con la explosión de Big Data en el mundo actual, se vuelve necesario contar con métodos sofisticados y eficientes de manejo de grandes volúmenes de datos. Las clásicas bases de datos relacionales cada vez se ven más limitadas especialmente ante datos no estructurados por lo que un sinnúmero de nuevas tecnologías han surgido para llenar estos vacíos.

Un ingeniero de datos se especializa en estas herramientas y se vale del dominio del negocio o industria en la que trabaje, para volver a los datos, uno de los activos más valiosos que una organización puede tener.

Partiendo de lo anterior, Librería del Pacífico es una empresa encargada de distribuir productos en el mercado salvadoreño, pero carece de un área de análisis de datos que ayude en la toma de decisiones gerenciales, a la hora de realizar compras de los diferentes productos que comercializa para poder abastecer a cada una de sus sucursales. La toma de decisiones se verá reflejada en función de los datos obtenidos de la aplicación ODOO, de las compras realizadas por la empresa Librería del Pacífico, para luego poder extraer, transformar y cargar la información optimizada en un dashboard (Gráficos y paneles), que apoye en la toma de decisiones estratégicas de la empresa.

Para solventar la carencia con la que cuenta Librería del Pacífico, es necesario tener presente el concepto de ingeniería de datos, la cual se define como la disciplina que se encarga de recolectar, trasladar y validar datos para su explotación. Utiliza principios de ingeniería de software para construir modelos de datos como almacenes de datos (Datawarehouse) y lagos de datos (Data Lake) garantizando la disponibilidad, consistencia, mantenibilidad, seguridad y limpieza de datos para que estos puedan ser utilizados.

Llamaremos *modelado dimensional* al conjunto de técnicas y conceptos utilizados en el diseño de almacenes de datos. Es una forma de acercar los datos a la manera en que estos serán convertidos en información útil para los usuarios del negocio. El objetivo final es que estos puedan encontrar de manera intuitiva y rápida la información que necesitan.

Finalmente, haremos uso de herramientas de Business Intelligence (BI) que se alimenta de datos de negocios y presenta reportes, paneles, tablas y gráficos de forma amigable para el usuario.

El proyecto se implementará haciendo uso de diferentes herramientas como un gestor de base de datos, una aplicación para poder construir los ETL a implementar, servicios de almacenamiento en la nube como lo es Amazon Web Services y una herramienta para visualizar y crear informes de apoyo a los usuarios finales, se puede visualizar con un diagrama arquitectónico de la solución, luego de haber presentado el diseño de la solución se llevará a cabo una estrategia de implementación de la propuesta de solución, tomando en cuenta una serie de pasos para su correspondiente implementación.

Capítulo I: Especificación de proyecto

a. Situación actual

I. Antecedentes

Librería Del Pacífico es una empresa que lleva 15 años comercializando productos en el mercado salvadoreño, principalmente material de oficina y artículos de papelería (papel en varios formatos y tamaños, bolígrafos, lápices, accesorios), pero también puede comercializar libros de texto, mochilas, y otros artículos relacionados, como cartuchos de impresora y demás.

Librería El Pacífico cuenta con diez sucursales las cuales se encuentran distribuidas a lo largo del territorio salvadoreño, estas están ubicadas en; **Zona Central:** San Salvador, Santa Tecla, Nuevo Cuscatlán, Zacatecoluca, San Vicente; **Zona Occidental:** Santa Ana, Sonsonate, Ahuachapán; **Zona Oriental:** San Miguel, Usulután. Para ver la ubicación de las sucursales ver los **Anexos (c)**.

Actualmente la empresa no cuenta con un área de análisis de datos que inspeccione, limpie y transforme grandes volúmenes de datos con el objetivo de resaltar información útil que ayude en la toma de decisiones gerenciales de la empresa.

Para el proceso de compras de productos la empresa Librería Del Pacífico utiliza el sistema ERP: ODOO, específicamente en el módulo de Gestión de Compras.

Entre los **principales beneficios** con los que ODOO cuenta, tenemos los siguientes:

- **Automatiza el flujo de compras:** Envía solicitudes de cotizaciones a los proveedores de forma automática según los niveles de existencias. Mejora el rendimiento de tus compras e inventarios con reglas de aprovisionamiento según los niveles de existencias, reglas de logística, órdenes de venta, órdenes de fabricación pronosticadas, etc. Selecciona diferentes métodos de reabastecimiento para cada producto de acuerdo con tus estrategias de fabricación y entrega.
- **Listas de precios de proveedor y disponibilidad de productos:** Toma de decisiones de compra inteligentes al aprovechar los mejores precios. Importar con facilidad las listas de precios y referencias de proveedores para tomar decisiones de compra más

inteligentes según promociones, cantidades y condiciones contractuales especiales. Llevar el seguimiento de la disponibilidad de un producto en las existencias de un proveedor y comprobar el estado de la orden desde la aplicación. Referenciar un precio de venta respecto a los precios de un proveedor.

- **Recibir las mejores ofertas de compra:** Ayuda a conseguir el mejor precio al negociar con diversos proveedores. Enviar licitaciones de compra, integrar las respuestas del proveedor en el proceso y comparar propuestas. Elegir la mejor oferta y enviar las órdenes de compra con facilidad. Utilizar los reportes para analizar la calidad de los proveedores posteriormente. También se puede utilizar un contrato marco para comprar mercancía de un proveedor a un precio negociado, de manera recurrente y durante un periodo de tiempo en concreto.
- **Obtener estadísticas de las compras:** Analiza, pronostica y planifica tus órdenes con eficiencia. Obtén estadísticas precisas sobre el rendimiento de sus proveedores gracias a reportes flexibles: retrasos en entregas, descuentos sobre los precios, cantidades compradas etc. Integra las compras con la contabilidad analítica para analizar la rentabilidad de tus contratos.
- **Gestionar múltiples empresas:** Ahorrar tiempo y esfuerzo gracias a las reglas multiempresa de ODOO. Utilizar una sola instancia de ODOO para sincronizar operaciones entre diferentes empresas o almacenes. Crear órdenes de venta, compartir clientes, proveedores y productos y gestionar facturas para todas las empresas al mismo tiempo. Ahorrar más tiempo al automatizar el proceso de facturación entre todas las empresas.

Para entender el proceso de compra es necesario comprender la **lógica del negocio** y cómo se relaciona con otras aplicaciones de ODOO.

- Con la relación de **Compra – Inventario** es posible sincronizar los niveles de existencias según las compras y establecer reglas de reabastecimiento automáticas para evitar el desabasto. Además de automatizar las solicitudes de cotizaciones y las órdenes de compra según las existencias mínimas, y llevar el seguimiento de los envíos entrantes. El módulo de Inventario se ve afectado cada vez que se hace una compra, con un movimiento de ingreso.

- Con la relación de **Compra – Facturación** es posible Convertir las órdenes de compra en facturas de proveedor para evitar la doble entrada. Además de obtener información significativa sobre las facturas de proveedores.
- Por último, con la relación **Compra – Contabilidad** es posible obtener una contabilidad más precisa al integrar órdenes de compra y facturas.

II. Descripción del problema

Librería Del Pacífico es una cadena de tiendas que cuenta con un Almacén Central encargado de realizar compras conjuntas, que abastecerán a cada sucursal ubicadas en las diferentes regiones del país de acuerdo con su necesidad.

Por tanto, es necesario contar con un sistema logístico eficiente que permita satisfacer oportunamente la demanda de los clientes, promoviendo la gestión de compras.

Para ello vemos la necesidad de ayudar a la empresa en el estudio de los datos haciendo uso de herramientas como AWS S3, IAM, AWS Redshift, Talend Open Studio, Power BI, PostgreSQL, para poder obtener información necesaria que ayude a la toma de decisiones a la empresa.

Dicho lo anterior, en el presente documento nos enfocaremos únicamente, en el proceso de compra de productos de la empresa Librería Del Pacífico. La cual utiliza el sistema ERP: ODOO, específicamente en el módulo de Gestión de Compras.

Entre sus **principales procesos** tenemos:

1. Solicitudes de Cotización.
2. Órdenes de Compra.
3. Mantenimiento de Proveedores.
4. Mantenimiento de Productos.
5. Licitaciones de Compras.
6. Reporte de Compras.
7. Facturación.
8. Reabastecimiento de Inventario.
9. Programación de Compras.

Para el análisis de compras de la empresa Librería del Pacífico haciendo uso del software ODOO definimos una serie de métricas que serán de suma importancia para la empresa en el análisis de datos y tomas de decisiones, las cuales definimos a continuación:

1. Cantidad de compras por producto (conteo).
2. Volumen de compras por producto (dinero).

3. Productos por Proveedor (conteo).
4. Volumen de Productos por Proveedor (dinero).
5. Volumen de Solicitudes por Proveedor (dinero).
6. Cantidades de licitaciones sin Respuesta (conteo).

III. Planteamiento del problema

Realizaremos el planteamiento del problema en función de una caja negra en el cual podemos observar lo siguiente:

- Las **entradas** se verán reflejadas con los datos obtenidos de la aplicación ODOO de las compras realizadas en las diferentes sucursales de la Librería del Pacífico en el país.
- Las **funciones** representarán las diferentes herramientas utilizadas en el análisis de los datos, los cuales son: la obtención de los datos haciendo uso de un gestor de base de datos y Talen Open Studio para poder obtener archivos planos con la información cruda y poder realizar la extracción, transformación y carga; AWS Services (IAM, S3, Redshift) para tener una conexión con Talend Open Studio y poder cargar los datos a la nube, finalizando con la herramienta Power BI para transformar, visualizar datos y crear informes de los mismos.
- **Los Informes Gráficos** representará los diferentes reportes que se esperan obtener de la transformación de los datos los cuales se nombran a continuación: Cantidad de compras por día, Total de compras por fecha, Total de impuesto pagados, Total de compras por producto, Total de compras por proveedor, Cantidad de licitaciones sin respuesta.

En la **Imagen 1**, se puede visualizar el planteamiento del problema en función de un diagrama de caja negra.



Imagen 1. Diagrama de Caja Negra.

b. Objetivos

I. Objetivo General

Construir una solución de análisis empresarial, que se integre al sistema transaccional de la empresa Librería del Pacífico que permita extraer, transformar y cargar datos para analizarlos y presentarlos en un Dashboard que apoye la toma de decisiones gerenciales.

II. Objetivos Específicos

- Extraer la base de datos del módulo de compras del sistema transaccional de la empresa Librería del Pacífico a través de un Sistema Gestor de Base de Datos para gestionar y administrar la información generada.
- Diseñar procesos que permitan extraer, transformar y cargar datos que provienen del sistema transaccional para su posterior análisis.
- Diseñar una estructura en la nube utilizando la herramienta S3 de AWS que soporte el almacenamiento y transformación de los datos extraídos del sistema transaccional.
- Implementar un modelo dimensional de estrella utilizando la herramienta Redshift de AWS que soporte la carga de datos proveniente de S3.
- Presentar un Dashboard (Gráficos y Paneles) de información optimizada y limpia que permita incorporar objetos visuales para apoyar la toma de decisiones gerenciales.

c. Alcances

- Diseño del modelo ER del módulo de compras del sistema transaccional de la empresa Librería del Pacífico.
- Creación de procesos que extraigan, transformen y carguen la información generada por el sistema transaccional.
- Almacenamiento en la nube de S3 en AWS que soporte la limpieza de los datos del sistema transaccional.
- Diseño de un modelo dimensional de estrella donde se cargará la información transformada.
- Presentación de informes gráficos que apoyen la toma de decisiones.

d. Justificación

En la actualidad, se vuelve necesario contar con métodos sofisticados y eficientes de manejo de grandes volúmenes de datos. Las clásicas bases de datos relacionales cada vez se ven más limitadas especialmente ante datos no estructurados por lo que un sinnúmero de nuevas tecnologías han surgido para llenar estos vacíos.

El módulo de compras del sistema transaccional de la empresa Librería del Pacífico genera un volumen considerable de datos con los cuales no es posible realizar proyecciones en las compras, ni análisis que apoyen las decisiones gerenciales de la empresa.

Por tanto, la empresa se plantea integrar una solución que permita extraer, transformar y cargar datos para analizarlos y presentarlos en un Dashboard que apoye la toma de decisiones gerenciales.

Capítulo II: Análisis y diseño de la propuesta de solución

a. Metodología de trabajo

Data Warehousing y Business Intelligence

Antes de profundizar en los detalles del modelado dimensional, es útil centrarse en los objetivos fundamentales del Data Warehousing (almacenamiento de datos) y Business Intelligence (inteligencia empresarial). Los objetivos se pueden desarrollar fácilmente caminando por los pasillos de cualquier organización y escuchando a la gestión empresarial. Estos temas recurrentes han existido durante más de tres décadas:

- “Recopilamos toneladas de datos, pero no podemos acceder a ellos”.
- “Necesitamos cortar y trocear los datos en todos los sentidos”.
- “La gente de negocios necesita acceder a los datos con facilidad”.
- “Solo muéstrame lo que es importante”.
- “Pasamos reuniones enteras discutiendo sobre quién tiene los números correctos en lugar de tomar decisiones”.
- “Queremos que las personas utilicen la información para respaldar una toma de decisiones más basada en hechos”.

Según la experiencia, estas preocupaciones siguen siendo tan universales que impulsan los requisitos básicos para el sistema DW/BI.

A continuación, se plantean los Requerimientos que debe cumplir el sistema DW/BI para solventar las preocupaciones mencionadas:

- **El sistema DW/BI debe hacer que la información sea fácilmente accesible.** El contenido del sistema DW/BI debe ser comprensible. Los datos deben ser intuitivos y obvios para el usuario comercial, no solo para el desarrollador. Las estructuras y etiquetas de los datos deben imitar los procesos de pensamiento y el vocabulario de los usuarios comerciales. Los usuarios comerciales desean separar y combinar datos analíticos en infinitas combinaciones. Las herramientas y aplicaciones de inteligencia de negocios que acceden a los datos deben ser simples y fáciles de usar. También deben

devolver los resultados de la consulta al usuario con tiempos de espera mínimos. Podemos resumir este requisito simplemente diciendo: simple y rápido.

- **El sistema DW/BI debe presentar la información de manera consistente.** Los datos en el sistema DW/BI deben ser creíbles. Los datos deben recopilarse cuidadosamente a partir de una variedad de fuentes, limpiarse, garantizarse la calidad y publicarse solo cuando sean aptos para el consumo del usuario. La coherencia también implica que se utilicen etiquetas y definiciones comunes para los contenidos del sistema DW/BI en todas las fuentes de datos. Si dos medidas de desempeño tienen el mismo nombre, deben significar lo mismo. Por el contrario, si dos medidas no significan lo mismo, deben etiquetarse de manera diferente
- **El sistema DW/BI debe adaptarse al cambio.** Las necesidades del usuario, las condiciones comerciales, los datos y la tecnología están sujetos a cambios. El sistema DW/BI debe estar diseñado para manejar este cambio inevitable con gracia para que no invalide los datos o aplicaciones existentes. Los datos y las aplicaciones existentes no deben cambiarse ni interrumpirse cuando la comunidad empresarial hace nuevas preguntas o se agregan nuevos datos al almacén. Finalmente, si se deben modificar los datos descriptivos en el sistema DW/BI, debe contabilizar adecuadamente los cambios y hacer que estos cambios sean transparentes para los usuarios.
- **El sistema DW/BI debe presentar la información de manera oportuna.** A medida que el sistema DW/BI se usa más intensamente para tomar decisiones operativas, es posible que los datos sin procesar deban convertirse en información procesable en cuestión de horas, minutos o incluso segundos. El equipo de DW/BI y los usuarios comerciales deben tener expectativas realistas de lo que significa entregar datos cuando hay poco tiempo para hacerlo, limpiarlo o validarlo.
- **El sistema DW/BI debe ser un bastión seguro que proteja los activos de información.** Las joyas de la corona informativa de una organización se almacenan en el Data Warehouse. Como mínimo, es probable que el Warehouse contenga información sobre lo que está comprando, a quién y a qué precio: detalles potencialmente dañinos en manos de las personas equivocadas. El sistema DW/BI debe controlar de manera efectiva el acceso a la información confidencial de la organización.

- **El sistema DW/BI debe servir como base autorizada y confiable para una mejor toma de decisiones.** El Data Warehouse debe tener los datos correctos para respaldar la toma de decisiones. Los resultados más importantes de un sistema DW/BI son las decisiones que se toman con base en la evidencia analítica presentada; estas decisiones brindan el impacto comercial y el valor atribuible al sistema DW/BI. La etiqueta original anterior a DW/BI sigue siendo la mejor descripción de lo que está diseñando: un sistema de soporte de decisiones.
- **La comunidad empresarial debe aceptar el sistema DW/BI para considerarlo exitoso.** No importa si creó una solución elegante utilizando los mejores productos y plataformas. Si la comunidad empresarial no adopta el entorno DW/BI y no lo usa activamente, ha fallado la implementación. A diferencia de la implementación de un sistema transaccional en el que los usuarios comerciales no tienen otra opción que usar el nuevo sistema, el uso de DW/BI a veces es opcional. Los usuarios comerciales adoptarán el sistema DW/BI si es la fuente "simple y rápida" de información procesable.

Aunque cada requisito de esta lista es importante, los dos últimos son los más críticos y, lamentablemente, a menudo los que más se pasan por alto. (Kimball, Dimensional Modeling Introduction, 2013)

Introducción al Modelado Dimensional

Ahora que se comprende los objetivos del sistema DW/BI, consideremos los conceptos básicos del modelado dimensional. El modelado dimensional es ampliamente aceptado como la técnica preferida para presentar datos analíticos porque aborda dos requisitos simultáneos:

- Entregar datos que sean comprensibles para los usuarios comerciales.
- Ofrezca un rendimiento de consulta rápido.

El modelado dimensional es una técnica de larga data para simplificar las bases de datos. Desde hace tiempo, las organizaciones de TI, consultores y los usuarios comerciales han gravitado naturalmente hacia una estructura dimensional simple para que coincida la necesidad humana fundamental de simplicidad. La simplicidad es crítica porque asegura que los usuarios puedan entender fácilmente los datos, así como también permite que el software navegue y entregar resultados de manera rápida y eficiente.

La capacidad de visualizar algo tan abstracto como un conjunto de datos de forma concreta y tangible es el secreto de comprensibilidad. Un modelo de datos que comienza simple tiene la posibilidad de permanecer simple al final del diseño. Un modelo que comienza complicado seguramente será demasiado complicado al final, resultando en Rendimiento lento de consultas y rechazo de usuarios comerciales.

Aunque los modelos dimensionales a menudo se instancian en sistemas de administración de bases de datos relacionales, son bastante diferentes de los modelos de tercera forma normal (3NF) que buscan eliminar las redundancias de datos. Las estructuras 3NF normalizadas dividen los datos en muchas entidades discretas, cada una de las cuales se convierte en una tabla relacional. Una base de datos de pedidos de ventas puede comenzar con un registro para cada línea de pedido, pero convertirse en un diagrama complejo de telaraña como un modelo 3NF, que quizás consista en cientos de tablas normalizadas.

La industria a veces se refiere a los modelos 3NF como modelos entidad-relación (ER) Los diagramas de entidad-relación (diagramas ER o ERD) son dibujos que comunican las relaciones entre tablas. Tanto los modelos 3NF como los dimensionales pueden estar representados en ERD porque ambos consisten en tablas relacionales unidas; la diferencia entre 3NF y los modelos dimensionales es el grado de normalización. Debido a que ambos tipos de modelos pueden presentarse como ERD, nos abstenemos de referirnos a los modelos 3NF como modelos ER; en cambio, los llamamos modelos normalizados para minimizar la confusión.

Las estructuras 3NF normalizadas son inmensamente útiles en el procesamiento operativo porque una transacción de actualización o inserción toca la base de datos en un solo lugar. Sin embargo, los modelos normalizados son demasiado complicados para las consultas de BI Los usuarios no pueden entender, navegar o recordar modelos normalizados que se asemejan a un mapa del sistema de autopistas de Los Ángeles. Del mismo modo, la mayoría de los sistemas de administración de bases de datos relacionales no pueden consultar de manera eficiente un modelo normalizado; la complejidad de las consultas impredecibles de los usuarios abrumba a los optimizadores de bases de datos, lo que resulta en un rendimiento de consulta desastroso. El uso del modelado normalizado en el área de presentación de DW/BI derrota la recuperación de datos intuitiva y de alto rendimiento. Afortunadamente, el modelado dimensional aborda el problema de los esquemas demasiado complejos en el área de presentación.

Un modelo dimensional contiene la misma información que un modelo normalizado, pero empaqueta los datos en un formato que brinda comprensión al usuario, rendimiento de consultas y resiliencia al cambio. (Kimball, Dimensional Modeling Introduction, 2013)

Esquema de Estrella

En las bases de datos usadas para data warehousing, un esquema en estrella es un modelo de datos que tiene una tabla de hechos (Fact Table) que contiene los datos para el análisis, rodeada de las tablas de dimensiones, como se muestra en la **Imagen 2**. Este aspecto, de tabla de hechos más grande rodeada de radios o tablas más pequeñas es lo que asemeja a una estrella, dándole nombre a este tipo de construcciones.

Las tablas de dimensiones tendrán siempre una clave primaria simple, mientras que, en la tabla de hechos, la clave principal estará compuesta por las claves principales de las tablas dimensionales.

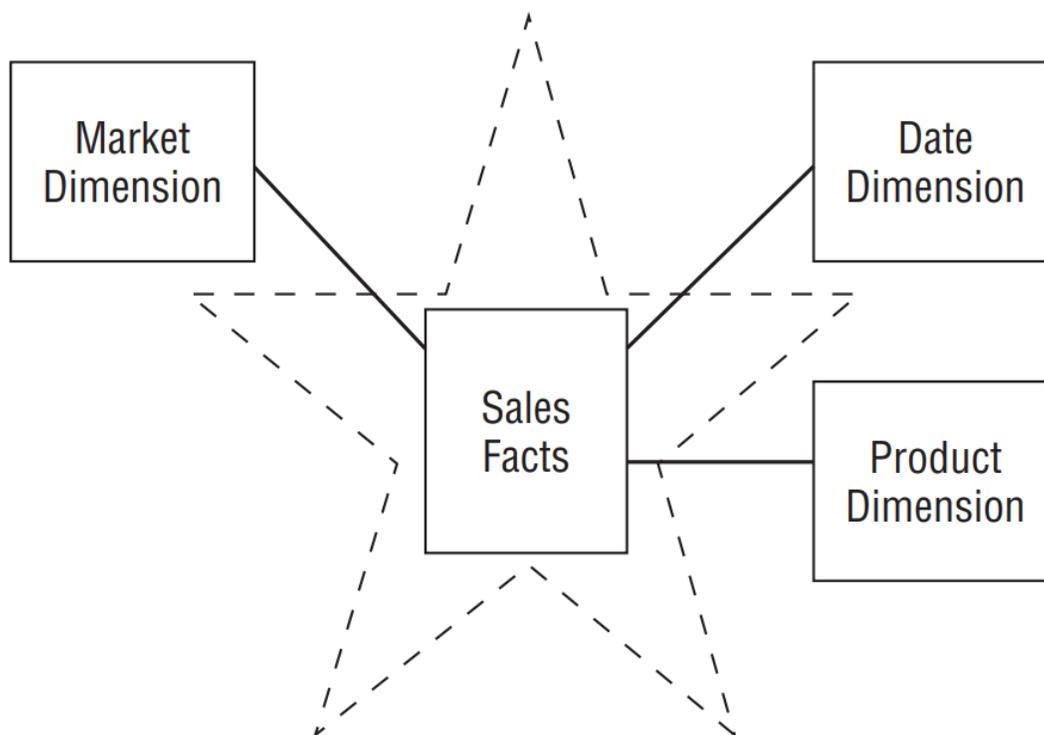


Imagen 2. Esquema de Estrella (Kimball, Star Schemas Versus OLAP Cubes, 2013)

A continuación, describiremos los dos componentes clave de un esquema de estrella.

Tabla de Hechos (Fact Table)

La tabla de hechos en un modelo dimensional almacena las medidas de desempeño resultantes de los eventos del proceso comercial de una organización. Debe esforzarse por almacenar los datos de medición de bajo nivel resultantes de un proceso comercial en un modelo unidimensional, como se muestra en la **Imagen 3**. Debido a que los datos de medición son abrumadoramente el conjunto de datos más grande, no deben replicarse en múltiples lugares para múltiples funciones organizacionales en la empresa. (Kimball, Fact Tables for Measurements, 2013)

Cada fila de una tabla de hechos corresponde a un evento de medición. Los datos de cada fila se encuentran en un nivel específico de detalle, denominado grano, como una fila por producto vendido en una transacción de venta. Uno de los principios básicos del modelado dimensional es que todas las filas de medidas en una tabla de hechos deben tener el mismo grano. Tener la disciplina para crear tablas de hechos con un solo nivel de detalle garantiza que las mediciones no se cuenten dos veces de manera inapropiada.

Retail Sales Facts
Date Key (FK)
Product Key (FK)
Store Key (FK)
Promotion Key (FK)
Customer Key (FK)
Clerk Key (FK)
Transaction #
Sales Dollars
Sales Units

Imagen 3. Tabla de Hechos (Kimball, Fact Tables for Measurements, 2013)

La idea de que un evento de medición en el mundo físico tiene una relación de uno a uno con una sola fila en la tabla de hechos correspondiente es un principio fundamental para el modelado dimensional. Todo lo demás se construye a partir de esta base.

Tablas de dimensiones

Las tablas de dimensiones son complementos integrales de una tabla de hechos. Las tablas de dimensiones contienen el contexto textual asociado con un evento de medición de procesos comerciales. Describen el "quién, qué, dónde, cuándo, cómo y por qué" asociado con el evento. (Kimball, Dimension Tables for Descriptive Context, 2013)

Las tablas de dimensiones suelen tener muchas columnas o atributos. No es raro que una tabla de dimensiones tenga de 50 a 100 atributos, como se muestra en la **Imagen 4**; aunque, naturalmente, algunas tablas de dimensiones solo tienen un puñado de atributos.

Las tablas de dimensiones tienden a tener menos filas que las tablas de hechos, pero pueden ser anchas con muchas columnas de texto grandes. Cada dimensión está definida por una sola clave primaria, que sirve como base para la integridad referencial con cualquier tabla de hechos dada a la que se une.

Product Dimension
Product Key (PK)
SKU Number (Natural Key)
Product Description
Brand Name
Category Name
Department Name
Package Type
Package Size
Abrasive Indicator
Weight
Weight Unit of Measure
Storage Type
Shelf Life Type
Shelf Width
Shelf Height
Shelf Depth
...

Imagen 4. Tablas de Dimensiones (Kimball, Dimension Tables for Descriptive Context, 2013)

Los atributos de la tabla de dimensiones juegan un papel vital en el sistema DW/BI. Debido a que son la fuente de prácticamente todas las restricciones y etiquetas de informes, los atributos de dimensión son fundamentales para que el sistema DW/BI sea usable y comprensible. Los

atributos deben consistir en palabras reales en lugar de abreviaturas crípticas. Debe esforzarse por minimizar el uso de códigos en las tablas de dimensiones reemplazándolos con atributos textuales más detallados.

Tabla de Hechos y Dimensiones unidos en un esquema de estrella

Ahora que se comprende las tablas de hechos y dimensiones, es hora de unir los componentes básicos en un modelo dimensional. Cada proceso de negocio está representado por un modelo dimensional que consta de una tabla de hechos que contiene las medidas numéricas del evento rodeadas por un conjunto de tablas de dimensiones que contienen el contexto que era cierto en el momento en que ocurrió el evento; como se muestra en la **Imagen 5**.

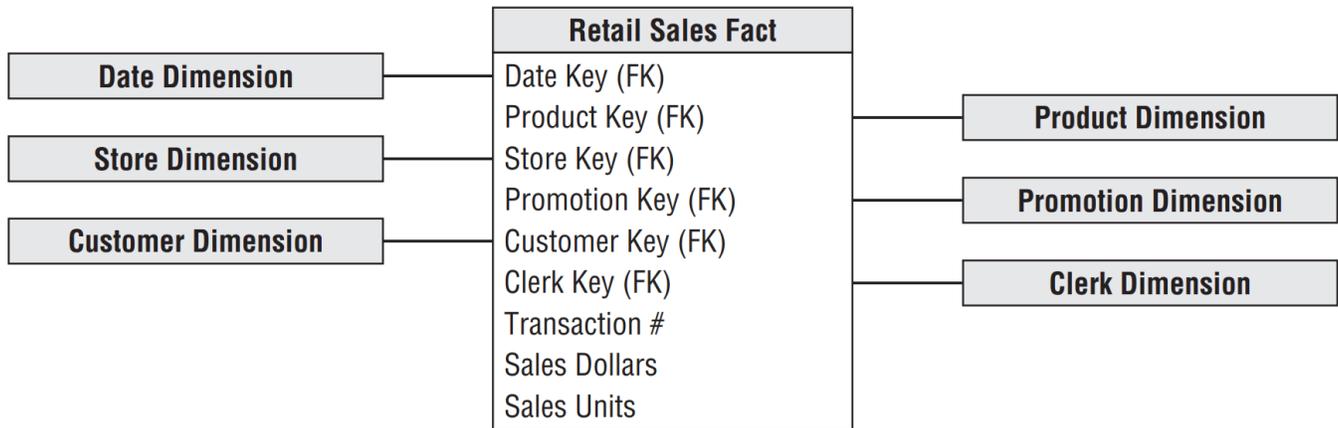


Imagen 5. Tablas de hechos y dimensiones en un modelo dimensional. (Kimball, Facts and Dimensions Joined in a Star Schema, 2013)

Lo primero que se debe notar sobre el esquema dimensional es su simplicidad y simetría. Obviamente, los usuarios comerciales se benefician de la simplicidad porque los datos son más fáciles de entender y navegar. Además, el número reducido de tablas y el uso de descriptores comerciales significativos facilitan la navegación y reducen la probabilidad de que se produzcan errores. (Kimball, Facts and Dimensions Joined in a Star Schema, 2013)

La simplicidad de un modelo dimensional también tiene beneficios de rendimiento. Los optimizadores de bases de datos procesan estos esquemas simples con menos uniones de manera más eficiente. Un motor de base de datos puede hacer suposiciones sólidas sobre primero restringir las tablas de dimensiones altamente indexadas y luego atacar la tabla de hechos de una sola vez con el producto cartesiano de las claves de la tabla de dimensiones

que satisfacen las restricciones del usuario. Sorprendentemente, con este enfoque, el optimizador puede evaluar uniones arbitrarias de n vías a una tabla de hechos en un solo paso a través del índice de la tabla de hechos. (Kimball, Star Schemas Versus OLAP Cubes, 2013)

Arquitectura de Kimball: Data Warehousing y Business Intelligence

Debe conocer la importancia estratégica de cada componente para evitar confundir su rol y función. Existen cuatro componentes separados y distintos que se deben considerar en el entorno DW/BI: sistemas de origen operativos, sistema ETL, área de presentación de datos y aplicaciones de inteligencia comercial, como se muestra en la **Imagen 6**. (Kimball, Kimball's DW/BI Architecture, 2013)

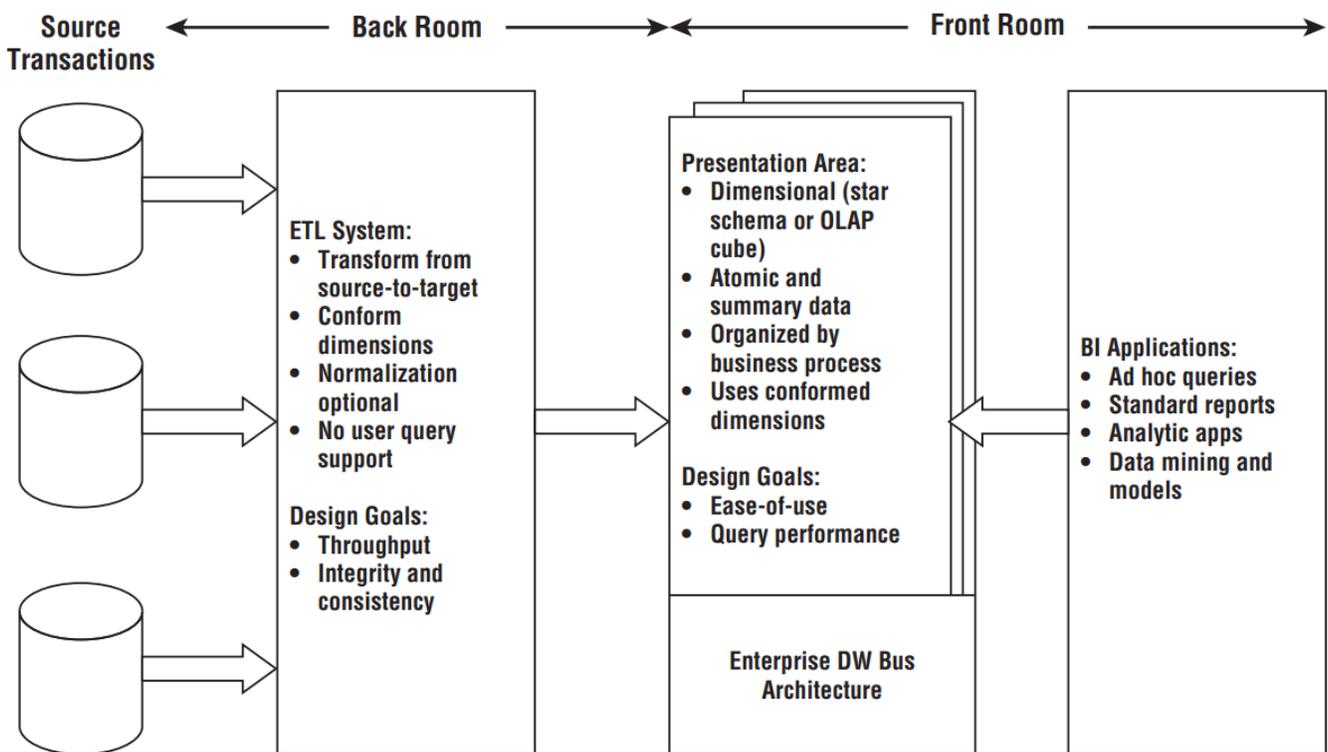


Imagen 6. Elementos centrales de la arquitectura Kimball DW/BI. (Kimball, Kimball's DW/BI Architecture, 2013)

Sistemas de Origen Operativos

Estos son los sistemas operativos de registros que capturan las transacciones del negocio. Piense en los sistemas de origen como fuera del almacén de datos porque presumiblemente tiene poco o ningún control sobre el contenido y el formato de los datos en estos sistemas

operativos. Las principales prioridades de los sistemas de origen son el rendimiento y la disponibilidad del procesamiento. Las consultas operativas contra los sistemas de origen son consultas estrechas, de un registro a la vez, que forman parte del flujo de transacciones normal y están severamente restringidas en sus demandas sobre el sistema operativo. Es seguro asumir que los sistemas de origen no se consultan de la manera amplia e inesperada en que se consultan típicamente los sistemas DW/BI.

Los sistemas fuente mantienen pocos datos históricos; un buen almacén de datos puede liberar a los sistemas fuente de gran parte de la responsabilidad de representar el pasado. En muchos casos, los sistemas de origen son aplicaciones de propósito especial sin ningún compromiso de compartir datos comunes como productos, clientes, geografía o calendario con otros sistemas operativos de la organización. (Kimball, Operational Source Systems, 2013)

Sistema de extracción, transformación y carga

El sistema de extracción, transformación y carga (ETL) del entorno DW/BI consta de un área de trabajo, estructuras de datos instanciadas y un conjunto de procesos. El sistema ETL es todo lo que se encuentra entre los sistemas fuente operativos y el área de presentación DW/BI. Este elemento es fundamental del rompecabezas general del sistema DW/BI. (Kimball, Extract, Transformation, and Load System, 2013)

- La **extracción** es el primer paso en el proceso de introducir datos en el entorno del almacén de datos. Extraer significa leer y comprender los datos de origen y copiar los datos necesarios en el sistema ETL para su posterior manipulación. En este punto, los datos pertenecen al almacén de datos.
- Una vez que los datos se extraen al sistema ETL, existen numerosas **transformaciones** potenciales, como la limpieza de los datos (corrección de errores ortográficos, resolución de conflictos de dominio, manejo de elementos faltantes o análisis en formatos estándar), combinación de datos de múltiples fuentes y eliminación. Duplicación de datos. El sistema ETL agrega valor a los datos con estas tareas de limpieza y conformidad al cambiar los datos y mejorarlos. Además, estas actividades se pueden diseñar para crear metadatos de diagnóstico, lo que finalmente conduce a la reingeniería de procesos comerciales para mejorar la calidad de los datos en los sistemas de origen con el tiempo.

- El paso final del proceso ETL es la estructuración física y la **carga** de datos en los modelos dimensionales de destino del área de presentación. Estos subsistemas son fundamentales, debido a que la misión principal del sistema ETL es entregar los datos a las tablas de dimensiones y hechos. Muchos de estos subsistemas definidos se centran en el procesamiento de tablas de dimensiones, como asignaciones de claves sustitutas, búsquedas de códigos para proporcionar descripciones apropiadas, división o combinación de columnas para presentar los valores de datos apropiados, o unión de estructuras de tablas subyacentes de tercera forma normal en dimensiones planas desnormalizadas. Por el contrario, las tablas de hechos suelen ser grandes y requieren mucho tiempo para cargarlas, pero su preparación para el área de presentación suele ser sencilla. Cuando las tablas de dimensiones y hechos en un modelo dimensional se han actualizado, indexado, suministrado con los agregados apropiados y una mayor calidad garantizada, se notifica a la comunidad empresarial que se han publicado los nuevos datos.

Área de presentación de apoyo a Business Intelligence

El área de presentación de DW/BI es donde los datos se organizan, almacenan y ponen a disposición para consultas directas de usuarios, redactores de informes y otras aplicaciones analíticas de BI. Debido a que el sistema ETL de la trastienda está fuera de los límites, el área de presentación es el entorno DW/BI en lo que respecta a la comunidad empresarial; es todo lo que la empresa ve y toca a través de sus herramientas de acceso y aplicaciones de BI.

El área de datos de presentación debe estructurarse en torno a eventos de medición de procesos comerciales. Este enfoque se alinea naturalmente con los sistemas operativos de captura de datos de origen. Los modelos dimensionales deben corresponder a eventos de captura de datos físicos; no deben estar diseñados para entregar el informe del día. Los procesos de negocios de una empresa cruzan los límites de los departamentos y funciones organizacionales. En otras palabras, debe construir una sola tabla de hechos para las métricas de ventas atómicas en lugar de completar bases de datos separadas similares, pero ligeramente diferentes, que contengan métricas de ventas para los equipos de ventas, marketing, logística y finanzas.

Los datos en el área de presentación consultable del sistema DW/BI deben ser dimensionales, atómicos (complementados con agregados que mejoran el rendimiento), centrados en procesos comerciales y adherirse a la arquitectura de bus de almacenamiento de datos empresarial. Los datos no deben estar estructurados de acuerdo con la interpretación de los datos por parte de departamentos individuales. (Kimball, Presentation Area to Support Business Intelligence, 2013)

Aplicaciones de inteligencia empresarial

El último componente principal de la arquitectura Kimball DW/BI es la aplicación de inteligencia empresarial (BI). El término aplicación de BI se refiere vagamente al rango de capacidades proporcionadas a los usuarios comerciales para aprovechar el área de presentación para la toma de decisiones analíticas.

Por definición, todas las aplicaciones de BI consultan los datos en el área de presentación DW/BI. Consultar, obviamente, es el objetivo principal de usar datos para mejorar la toma de decisiones.

Una aplicación de BI puede ser tan simple como una herramienta de consulta o tan compleja como una sofisticada aplicación de modelado o extracción de datos. Las herramientas de consulta, a pesar de lo poderosas que son, pueden ser entendidas y utilizadas de manera efectiva por solo un pequeño porcentaje de la población potencial de usuarios comerciales de DW/BI. Es probable que la mayoría de los usuarios de negocios accedan a los datos a través de aplicaciones y plantillas preconstruidas basadas en parámetros que no requieren que los usuarios construyan consultas directamente. Algunas de las aplicaciones más sofisticadas, como las herramientas de modelado o pronóstico, pueden volver a cargar los resultados en los sistemas de origen operativos, el sistema ETL o el área de presentación. (Kimball, Business Intelligence Applications, 2013)

b. Descripción de la propuesta de solución

Como punto de partida mostraremos el diagrama ER del módulo de compras del sistema de origen transaccional de la empresa Librería del Pacífico; Diagrama Entidad Relación del módulo de compras se muestra en la **Imagen 7**.

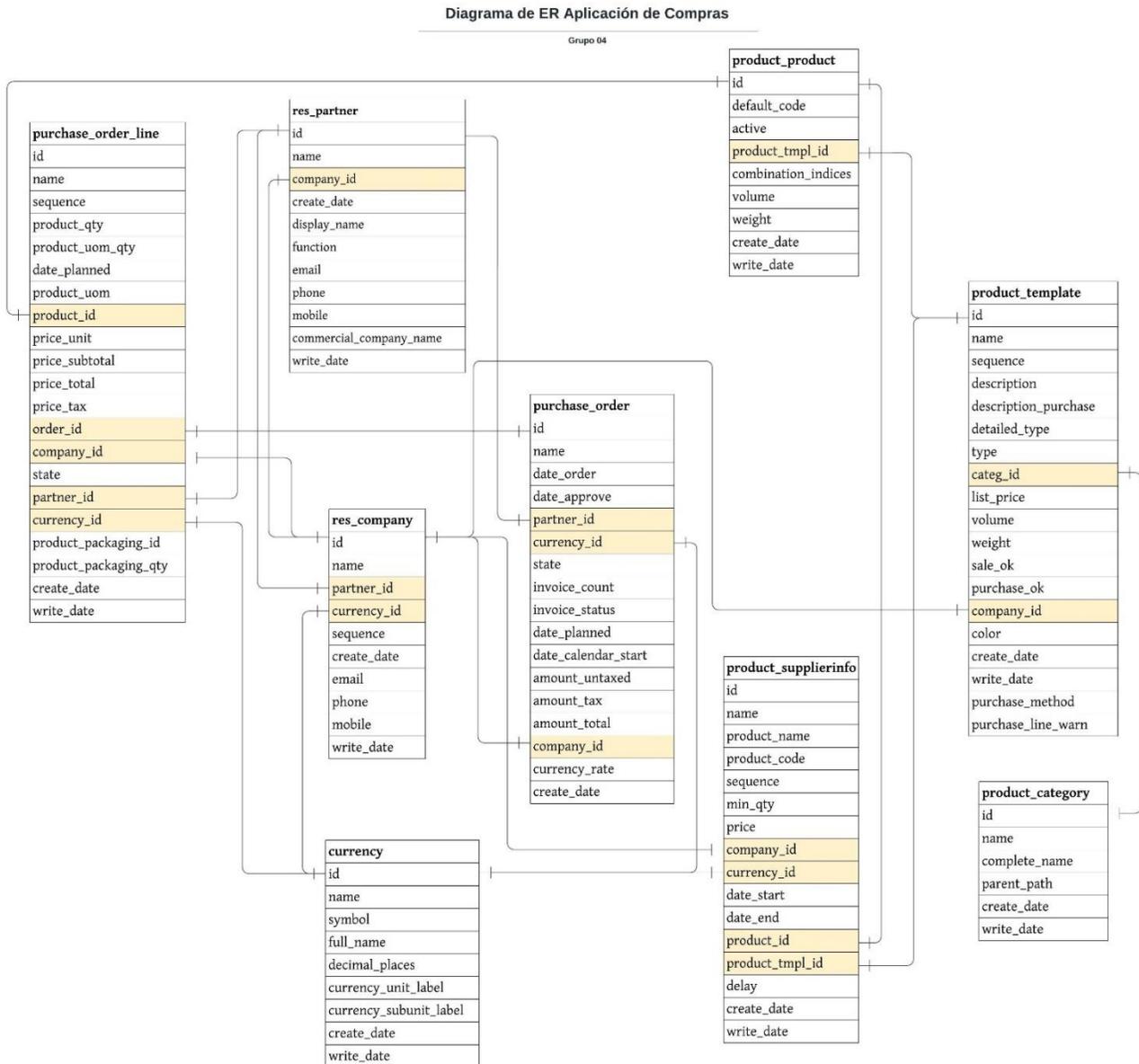


Imagen 7. ER Módulo de Compras

Con base en el diagrama ER, hemos propuesto el siguiente modelo dimensional, compuesto por 1 tabla de hechos y 4 dimensiones, siguiendo el Esquema de Estrella; como se muestra en la **Imagen 8**.

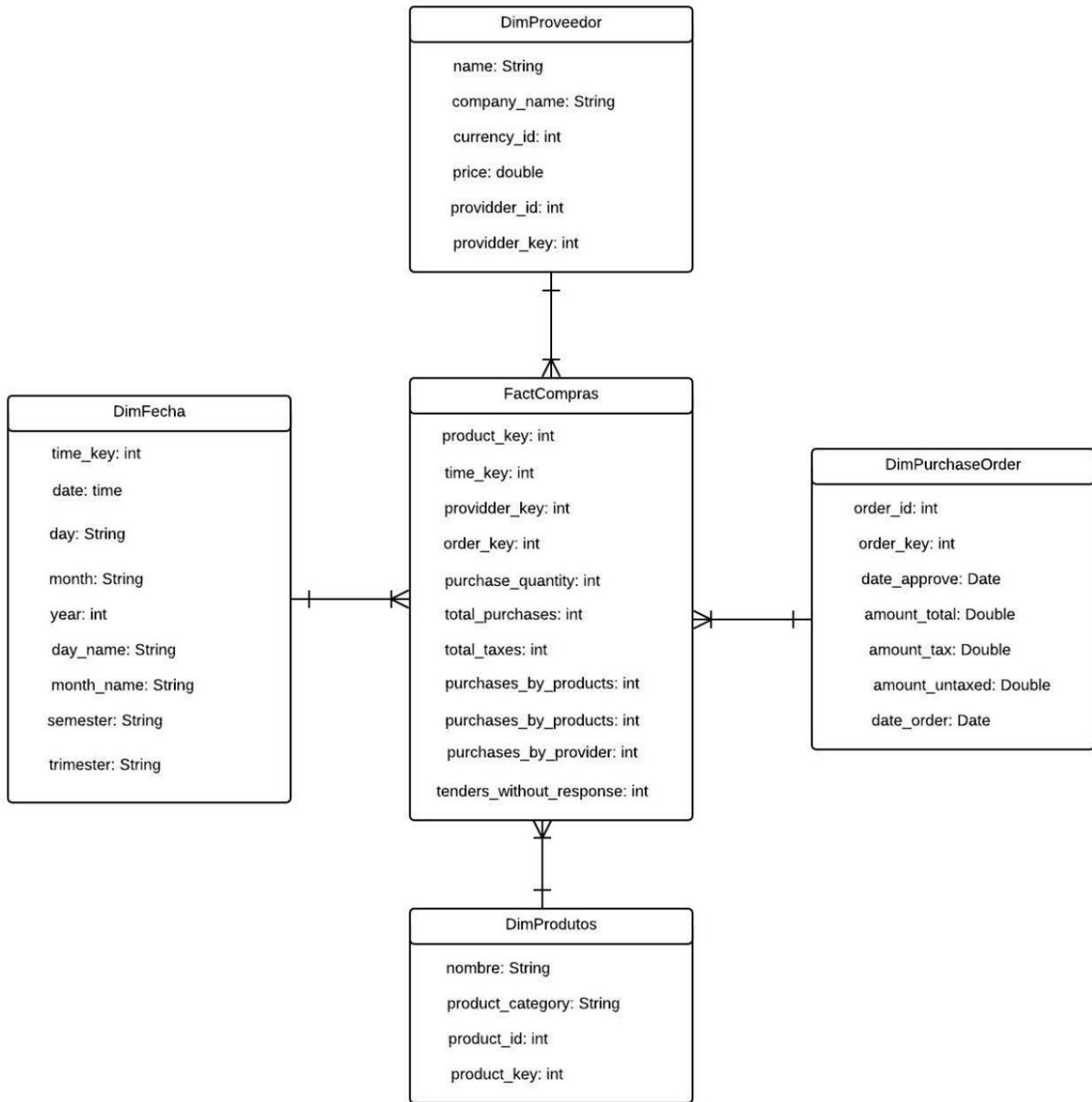


Imagen 8. Modelo Dimensional Propuesto

c. Descripción de la tecnología a utilizar

PostgreSQL

¿Qué es PostgreSQL?

PostgreSQL es un sistema gestor de bases de datos relacionales, está orientado a objetos, es multiplataforma y open source.

Está desarrollado desde 1996 por la comunidad partir del SGBD POSTGRES, que surgió a partir de un proyecto de investigación militar estadounidense con participación civil.

Como ya hemos dicho es orientado a objetos, es decir, todos los elementos de nuestra base de datos van a poder tratarse como objetos, algo parecido a un lenguaje de programación.

Es multisistema, por tanto, PostgreSQL puede ser instalado en Microsoft Windows, GNU/Linux, MacOS, BSD y muchos otros sistemas operativos.

Es extensible, podemos añadir funcionalidades que no vengan provistas de serie. Si quieres saber más sobre ello tenemos un artículo detallado de cómo instalar extensiones en PostgreSQL.

PostgreSQL es escalable y puede manejar bases de datos enormes, de más de 100 Terabytes y funciona bajo licencia libre, podemos usarlo para cualquier propósito sin ningún problema.

¿Para qué se usa PostgreSQL?

PostgreSQL se utiliza en diversos ámbitos, como, por ejemplo:

- Almacenamiento de datos (DWH).
- En servicios como Amazon Web Services Redshift.
- Para procesado de datos, almacenado tanto en la propia instancia como en otros servicios que puedan conectarse.
- En sistemas de información geográfica, como el servicio de mapas web o también en servicios móviles OpenStreetMap.
- En bases de datos para servicios web.
- En CMS como Drupal o WordPress.

- En la conocida base de datos de cine IMDb.

DBeaver

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

¿Qué funcionalidades permite el software de DBeaver?

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

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

¿Qué bases de datos incorpora?

Esta herramienta permite la conexión con la mayoría de las bases de datos SQL de alto nivel (PostgreSQL, MySQL, MariaDB, Oracle, SQLite) y también para bases de datos NoSQL (MongoDB, Cassandra). Además, nos incorpora acceso a otros tipos de información como archivos CSV, DBF, JSON o XML.

Talen Open Studio

¿Qué es Talend Open Studio?

Talend Open Studio (TOS) es una suite que aporta un conjunto muy complejo, variado y completo de herramientas para llevar a cabo la integración de datos que se ofrece en una

versión de código libre (open source). Precisamente por ello, esta es una de las herramientas de integración ETL (extract, transform, load) más utilizadas dentro del mundo Big Data; es más, es la cuarta en la lista después de Informatica Powercenter, IBM InfoSphere Datastage y Oracle Data Integrator (ODI).

Por otra parte, esta suite cuenta con un Community Edition (CE) totalmente funcional. Además, podrás utilizar una gran cantidad de componentes (más o menos 900) para llevar a cabo una gestión de datos personalizada.

Talend ofrece una colección de componentes genéricos destinados a los procesos de transformación de datos entre los que se incluyen las funcionalidades de normalizar, denormalizar, extraer o insertar campos en varios formatos, separar filas, agregar y ordenar filas, convertir tipos, etc.

Asimismo, ofrece conjuntos de componentes de procesamiento de datos que son específicos para diferentes tecnologías. Estos componentes tienen la finalidad de ejecutar dichas transformaciones de la forma más eficiente posible.

Amazon S3

¿Qué es Amazon S3?

Antes de que profundices en las características de Amazon S3, es relevante que conozcas que el sistema de Amazon Simple Storage Service se define como una herramienta de AWS que se encarga del almacenamiento de objetos, ofreciendo disponibilidad de datos, escalabilidad, alto rendimiento y niveles de seguridad.

Amazon S3 puede implementarse en los procesos de protección y almacenamiento de datos, sin importar su cantidad. Además, este servicio incluye casos de uso como copias de seguridad, análisis de Big Data, aplicaciones móviles, sitios web y demás.

Características de Amazon S3

El sistema de Amazon S3 incluye una serie de características que permiten su funcionamiento, como, por ejemplo, que tiene la capacidad de almacenar y mover datos entre las storage classes de S3.

Otras de las características de Amazon S3 es que contribuye en los procesos de configuración y aplicación de controles relacionados con el acceso a datos, protegiéndolos de usuarios sin autorización para ingresar al sistema.

De la misma manera, Amazon S3 puede utilizarse para realizar una revisión del uso de almacenamiento, así como las tendencias referentes a la actividad en la organización del cliente.

Amazon Redshift

¿Qué es Amazon Redshift?

La herramienta de Amazon Redshift se define como un servicio que se encarga de la gestión de procesos de operación, escalado y ajuste del almacenamiento de datos.

La opción de Amazon Redshift también se caracteriza por ser un servicio de data warehouse gestionado que cuenta con varios petabytes en la nube.

Características de Amazon Redshift

El servicio de Amazon Redshift incluye una serie de propiedades que caracterizan y permiten su funcionamiento, como, por ejemplo, que facilita la ejecución y el escalado del análisis de todos los datos en poco tiempo, sin que sea necesario la gestión de la infraestructura de almacenamiento.

El diseño de Amazon Redshift se caracteriza, además, por encargarse del control y la administración de las labores del almacenamiento de datos, como es el caso del aprovisionamiento de capacidad, la realización de copias de seguridad para el clúster y su supervisión o la aplicación de parches de actualización, entre otras.

Este servicio también permite realizar la conversión de datos en información en tan solo segundos, eliminando la necesidad de gestión de infraestructura, al tiempo que ofrece seguridad y fiabilidad en sus actividades.

Cabe destacar que el funcionamiento de Amazon Redshift se lleva a cabo gracias al uso del lenguaje de dominio específico SQL, que permite el análisis de datos semiestructurados y

estructurados para el almacenamiento de datos, bases de datos operativas y demás, lo que proporciona un alto rendimiento por un bajo costo.

Amazon IAM

¿Qué es AWS IAM?

AWS Identity and Access Management (IAM) es un servicio web que lo ayuda a controlar de forma segura el acceso a los recursos de AWS. Se utiliza IAM para controlar quién está autenticado (ha iniciado sesión) y autorizado (tiene permisos) para utilizar recursos.

IAM ofrece las siguientes características:

Acceso compartido a la cuenta de AWS

Puede conceder permiso a otras personas para administrar y utilizar los recursos de su cuenta de AWS sin tener que compartir su contraseña o clave de acceso.

Permisos detallados

Puede conceder diferentes permisos a diferentes personas para diferentes recursos. Por ejemplo, puede permitir que algunos usuarios completen el acceso a Amazon Elastic Compute Cloud (Amazon EC2), Amazon Simple Storage Service (Amazon S3), Amazon DynamoDB, Amazon Redshift y otros servicios de AWS. En el caso de otros usuarios, puede permitir el acceso de solo lectura a solo algunos buckets de S3 o conceder permiso para administrar solo algunas instancias EC2 o para tener acceso a la información de facturación, pero nada más.

Acceso seguro a los recursos de AWS para aplicaciones que se ejecutan en Amazon EC2

Puede utilizar características de IAM para proporcionar de forma segura credenciales para las aplicaciones que se ejecutan en instancias EC2. Estas credenciales proporcionan permisos a la aplicación para obtener acceso a otros recursos de AWS. Entre los ejemplos se incluyen buckets de S3 y tablas de DynamoDB.

Multi-Factor authentication (MFA)

Puede agregar una autenticación de dos factores a la cuenta y a los usuarios individuales para mayor seguridad. Con MFA usted o sus usuarios deben proporcionar no solo una contraseña o

clave de acceso para trabajar con la cuenta, sino también un código de un dispositivo configurado específicamente. Si ya utiliza una clave de seguridad FIDO con otros servicios y esta tiene una configuración compatible con AWS. Para obtener más información, consulte Configuraciones admitidas para usar las claves de seguridad FIDO.

Identidad federada

Puede permitir que los usuarios que ya tienen contraseñas en otros lugares, por ejemplo, en la red corporativa o en un proveedor de identidad de Internet, obtengan acceso temporal a la cuenta de AWS.

Power BI

Power BI es un servicio gratuito de análisis de negocio basado en la nube y visualización de datos, de negocio. Esta herramienta de Business Intelligence (BI), incorporada en la suite de productividad Microsoft Office 365, permite controlar la salud de un negocio mediante un dashboard en vivo, crear informes interactivos con Power BI Desktop y acceder a los datos en cualquier lugar con las aplicaciones nativas de móvil.

Actualmente cuenta con más de 5 millones de usuarios y es utilizado por más de 200.000 empresas. Es ampliamente utilizado en agencias de analítica web y empresas especializadas en Business Intelligence.

¿Por qué Power BI?

Power BI es una solución de análisis empresarial basado en la nube, que permite unir diferentes fuentes de datos, analizarlos y presentar un análisis de estos a través de informes y paneles. Con Power BI se tiene de manera fácil acceso a datos dentro y fuera de la organización casi en cualquier dispositivo. Estos análisis pueden ser compartidos por diferentes usuarios de la misma organización; por lo que directivos, financieros, comerciales, etc., pueden disponer de la información del negocio en tiempo real.

Se conforma fundamentalmente de estos componentes:

- Power BI Desktop: aplicación gratuita de escritorio para transformar, visualizar datos y crear informes de los mismos.

- Power BI Service: servicio online (SaaS) con funcionalidad similar a la aplicación desktop y permite publicar informes y configurar la actualización de datos automáticamente para que el personal de la organización tenga los datos actualizados.
- Power BI Mobile: aplicación móvil disponible para Windows, iOS y Android para visualizar informes y que se actualiza automáticamente con los cambios de los datos.

El flujo de trabajo de Power BI sigue el siguiente orden:

1. En Power BI Desktop se obtienen los datos a través de los diferentes conectores existentes.
2. Se cruzan los datos y se generan visualizaciones de datos e informes.
3. Se publican los informes y visualizaciones a través del servicio Power BI.
4. Diferentes usuarios pueden consultar dichos informes a través de apps en sus móviles de Android e iOS.

d. Diagrama arquitectónico de la solución

Representación del diagrama arquitectónico de la solución, se compone del gestor de base de datos, procesos ETL, almacenamiento en la nube AWS S3, Redshift y la herramienta para la representación gráfica de la solución Power BI, como se muestra en la **Imagen 9**.

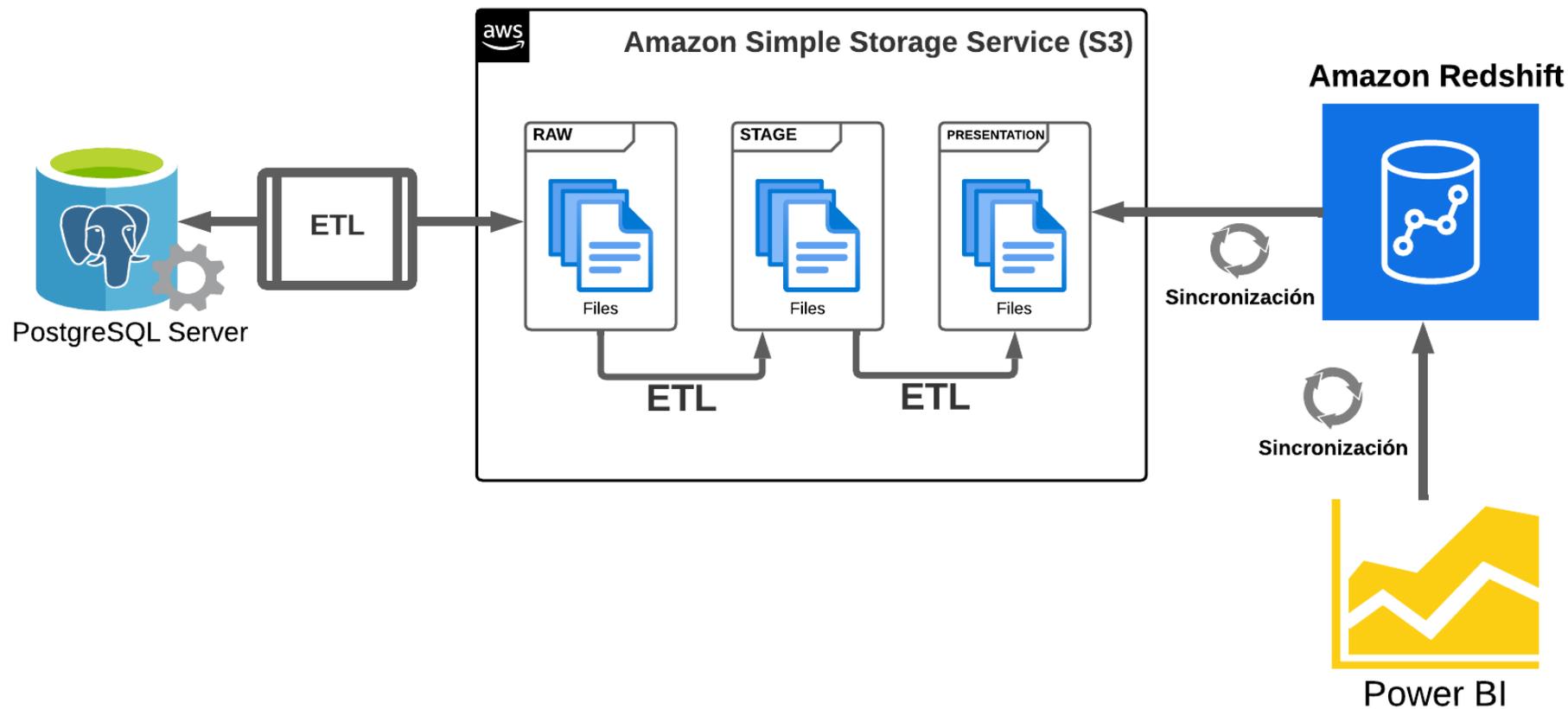


Imagen 9. Diagrama Arquitectónico

e. Descripción de cada componente de la solución.

Los componentes de la solución se describen en la **Tabla 1**.

Componente	Descripción
PostgreSQL Server	Utilizaremos el gestor de base de datos PostgreSQL para poder obtener las transacciones del día a día de la empresa Librería del Pacifico.
Procesos ETL	<p>Los procesos ETL es todo lo que se encuentra entre los sistemas fuente operativos y el área de presentación DW/BI. Este elemento es fundamental del rompecabezas general del sistema DW/BI.</p> <ul style="list-style-type: none"> • Extraer significa leer y comprender los datos de origen y copiar los datos necesarios en el sistema ETL para su posterior manipulación. En este punto, los datos pertenecen al almacén de datos. • Las transformaciones, como la limpieza de los datos (corrección de errores ortográficos, resolución de conflictos de dominio, manejo de elementos faltantes o análisis en formatos estándar), combinación de datos de múltiples fuentes y eliminación. Duplicación de datos. El sistema ETL agrega valor a los datos con estas tareas de limpieza y conformidad al cambiar los datos y mejorarlos. • La carga de datos en los modelos dimensionales de destino del área de presentación. Estos subsistemas son fundamentales, debido a que la misión principal del sistema ETL es entregar los datos a las tablas de dimensiones y hechos.
S3	Utilizaremos S3 para el almacenamiento de los datos procedentes de la base de datos transaccional. S3 será dividido en 3 escenarios, los cuales almacenarán las transformaciones que los datos vayan sufriendo a través de los diferentes procesos ETL que la solución lo requiera. Todo esto a través de una interfaz de servicio web.
RAW-S3	El primer escenario de S3 será RAW, el cuál almacenará la data cruda, directamente del gestor de base de datos, dichos datos aún no han sufrido ninguna transformación, limpieza o eliminación.
STAGE-S3	El segundo escenario de S3 será STAGE, el cuál almacenará los datos que han sufrido algún tipo de transformación, limpieza, eliminación, etc.
PRESENTATION-S3	El último escenario de S3 será PRESENTATION, el cuál almacenará los datos finales, listos para ser almacenados en el modelo dimensional de la solución.

Amazon RedShift	En Redshift crearemos nuestro modelo dimensional (Tabla de Hecho y Dimensiones), el cuál almacenará los datos provenientes de S3 (Habiendo completado los procesos ELT). Redshift cumplirá la función de base de datos en la nube, la cual cuenta con la capacidad para almacenar y analizar grandes cantidades de registros o fuentes de datos.
Power BI	Power BI será el pasó final del proyecto. Cumple la función de Business Intelligence. Estará conectado con Redshift. Será el encargado de presentar todos los datos de manera gráfica. El objetivo principal es usar los datos para mejorar la toma de decisiones.

Tabla 1. Componentes de la Solución

Capítulo III: Estrategia de implementación de propuesta de solución

a. Estrategia de implementación

Hemos definido que el producto a entregar será un producto por integración, es decir, será una implementación que se integrará a otra solución existente, nos referimos al sistema transaccional que se encuentra operando en la empresa Librería del Pacífico. Esta solución se integrará específicamente al módulo de compras, la cual respaldará la toma de decisiones. Los resultados más importantes de un sistema DW/BI son las decisiones que se toman con base en la evidencia analítica presentada; estas decisiones brindan el impacto comercial y el valor atribuible al sistema DW/BI. La etiqueta original anterior a DW/BI sigue siendo la mejor descripción de lo que está diseñando: un sistema de soporte de decisiones.

Una vez que hemos presentado el diseño de la solución, consideramos que los pasos a seguir para su correspondiente implementación son:

PostgreSQL

El primer paso para la implementación del modelo de la solución es haciendo uso del gestor de base de datos PostgreSQL para poder obtener las transacciones del día a día de la empresa Librería del Pacífico, creamos una base de datos con el nombre “**libreriapacifico**”, esta base nos ayuda a guardar todos los datos relacionados a las transacciones de compras de la empresa obtenidas del sistema ERP: ODOO, se puede visualizar el gestor de la base de datos en la **Imagen 10**.

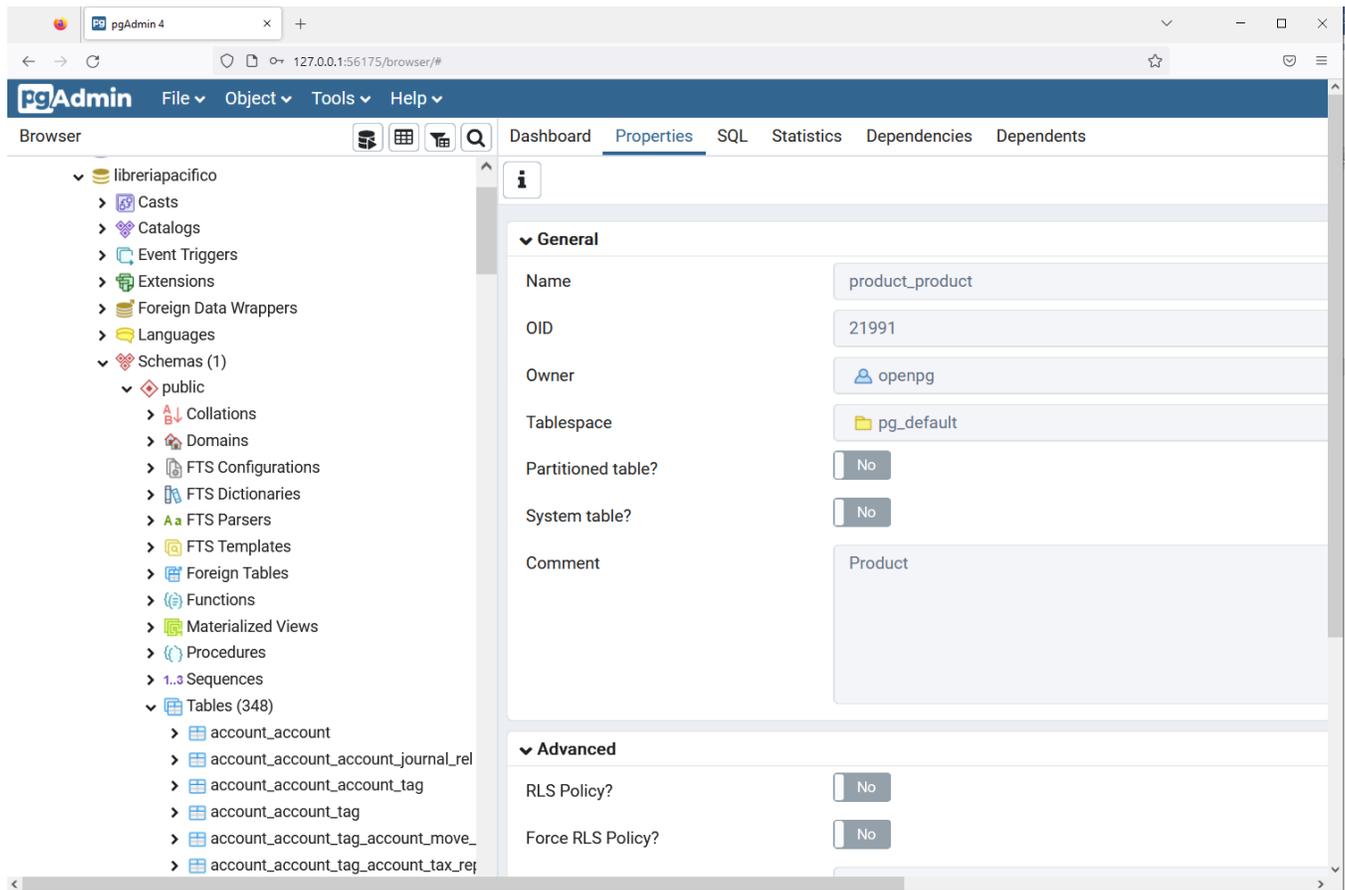


Imagen 10. Gestor de base de datos PostgreSQL con la base de datos transaccional.

DBeaver

Para la elaboración del diagrama entidad-relación nos apoyamos en la herramienta DBeaver utilizada en la gestión de bases de datos ya que permite una conexión con la mayoría de bases de datos SQL de alto nivel en este caso para nuestro uso PostgreSQL donde tenemos creada la base de datos “**libreriapacifico**” realizamos la configuración de conexión a PostgreSQL; la configuración de la conexión al gestor de base de datos PostgreSQL se puede visualizar en la **Imagen 11**. Luuego, gracias a una de sus funcionalidades a través de su interfaz amigable pudimos visualizar como se relacionaban todas las tablas de la base de datos y así poder crear el diagrama del modelo entidad-relación, como se muestra en la **Imagen 12**.

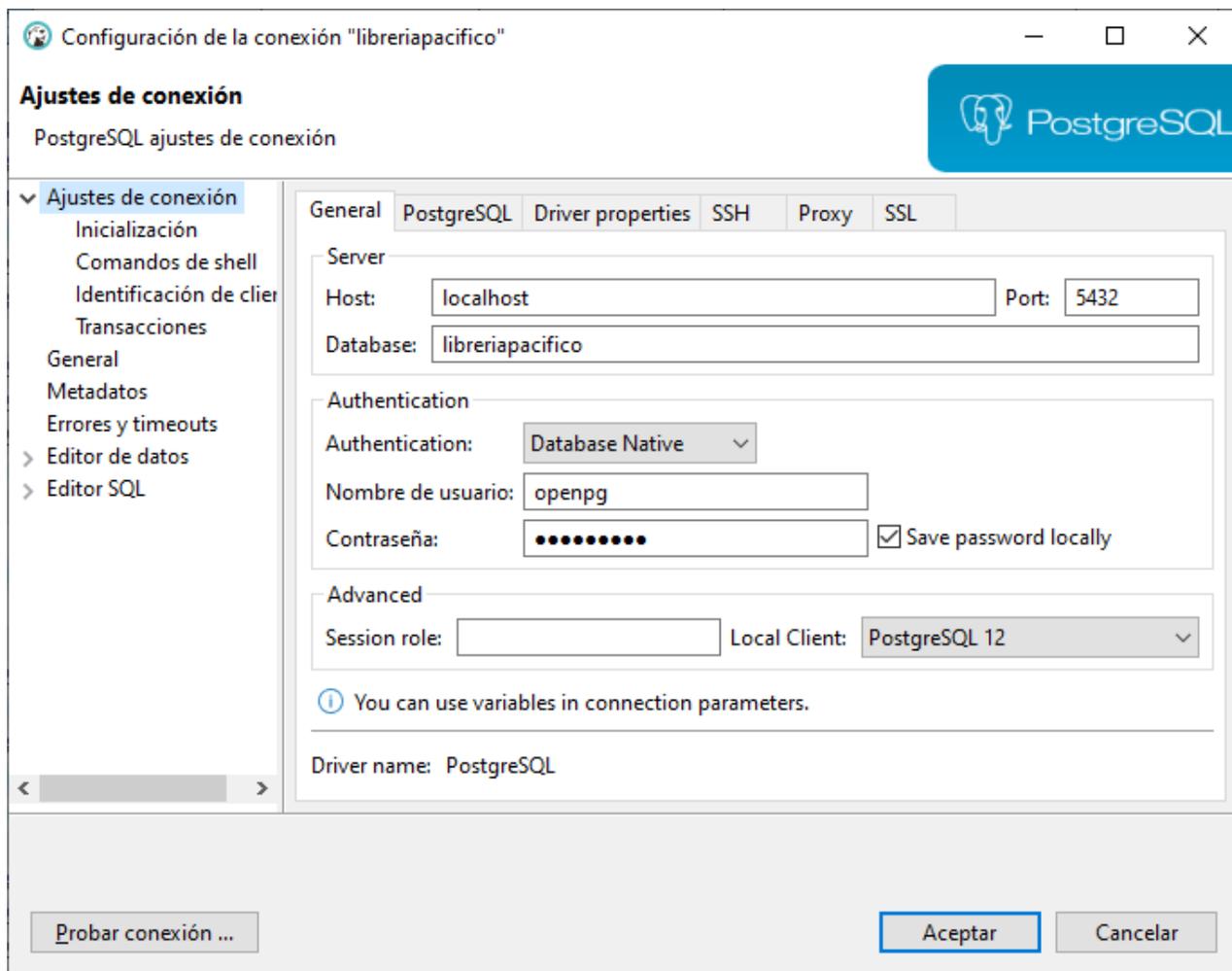


Imagen 11. Configuración de conexión a PostgreSQL.

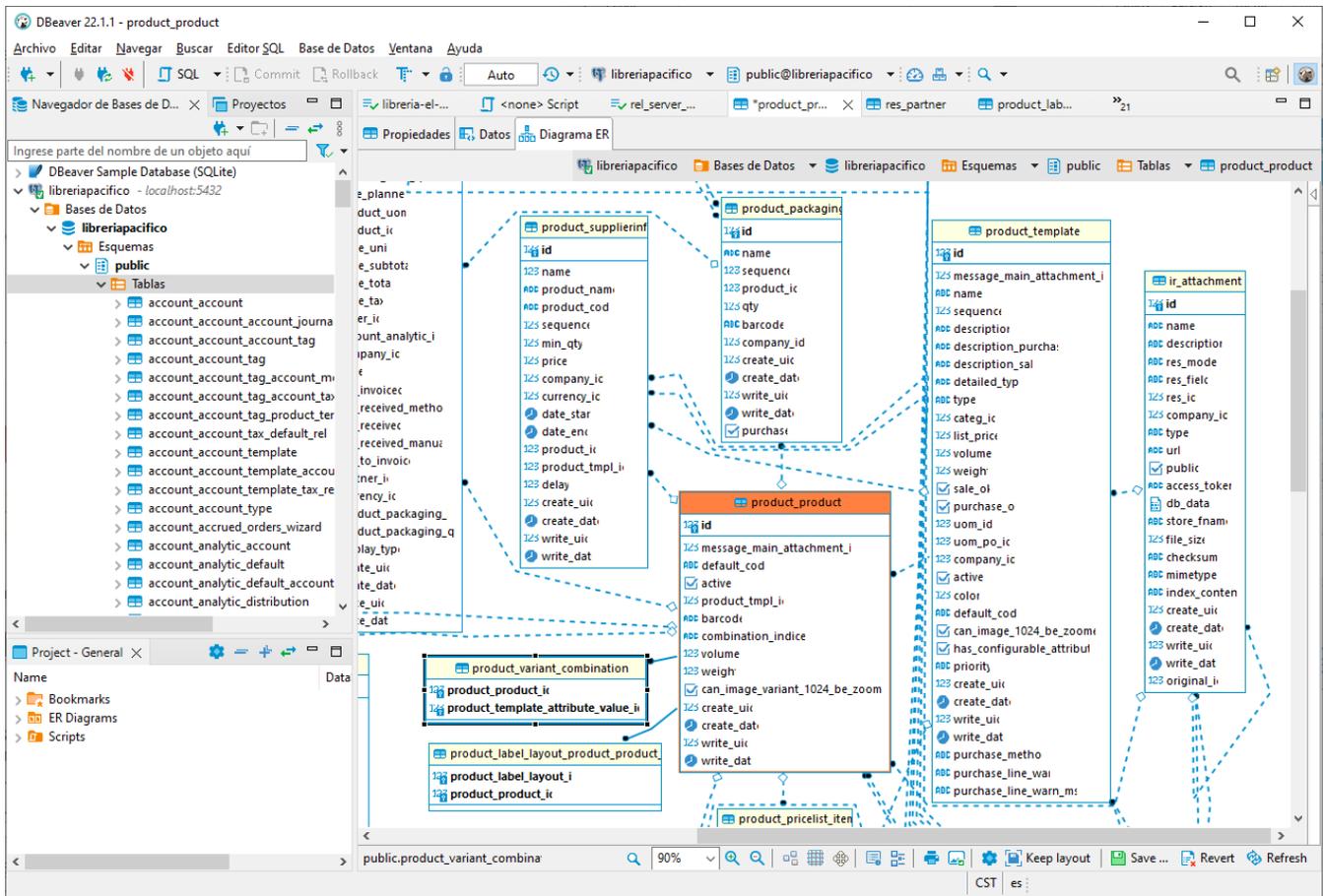


Imagen 12. Visualización de las relaciones entre tablas de la Base de datos, diagrama Entidad-Relación.

Talend Open Studio

Siguiendo con la implementación tenemos el proceso ETL (extraer, transformar, cargar) el cual es muy relevante dentro del análisis de datos. Existe una gran variedad de herramientas y programas que lo implementan para llevar a cabo el procesamiento de los macrodatos, por ejemplo, Talend Open Studio. Con la ayuda de esta herramienta creamos Jobs para poder realizar una conexión a la base de datos de PostgreSQL, generar la extracción de los archivos csv de las tablas que vamos a utilizar en el diseño de nuestra solución, transformar los datos y finalizar con la carga de los datos en Amazon S3.

Los pasos por seguir para llevar a cabo la configuración de la herramienta son los siguientes:

1. Luego de haber instalado la herramienta en el equipo procedemos a ejecutar la aplicación y nos aparece la pantalla de inicio, Podemos seleccionar uno de los proyectos que tengamos creados en el nuestro espacio de trabajo. Si es la primera vez que entramos creamos un nuevo proyecto con “Create a new Project”, también desde esta pantalla inicial, disponemos de la opción” Import an existing Project”, que permite importar un proyecto a partir de un fichero local. Nosotros vamos a seleccionar la opción “Select an existing project”, seleccionar el nombre del proyecto y dar clic en el botón “Finish”. la **Imagen 13**, muestra la selección de las opciones para poder abrir el proyecto correctamente.

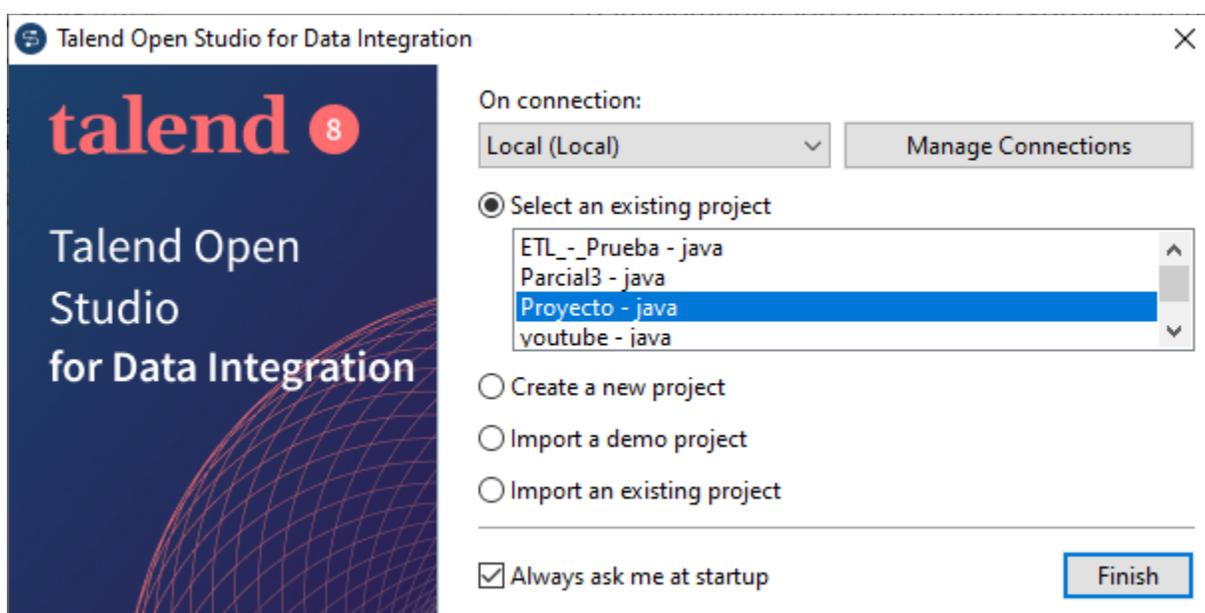


Imagen 13. Pantalla de inicio de la aplicación.

2. Crear la estructura de las carpetas para llevar un orden de trabajo, se crean las carpetas RAW, STAGE y PRESENTATION en las cuales vamos a almacenar los diferentes Jobs creados para la extracción, carga y transformación de los datos (ETL). La estructura de las carpetas creadas en el proyecto se puede visualizar en la **Imagen 14**.

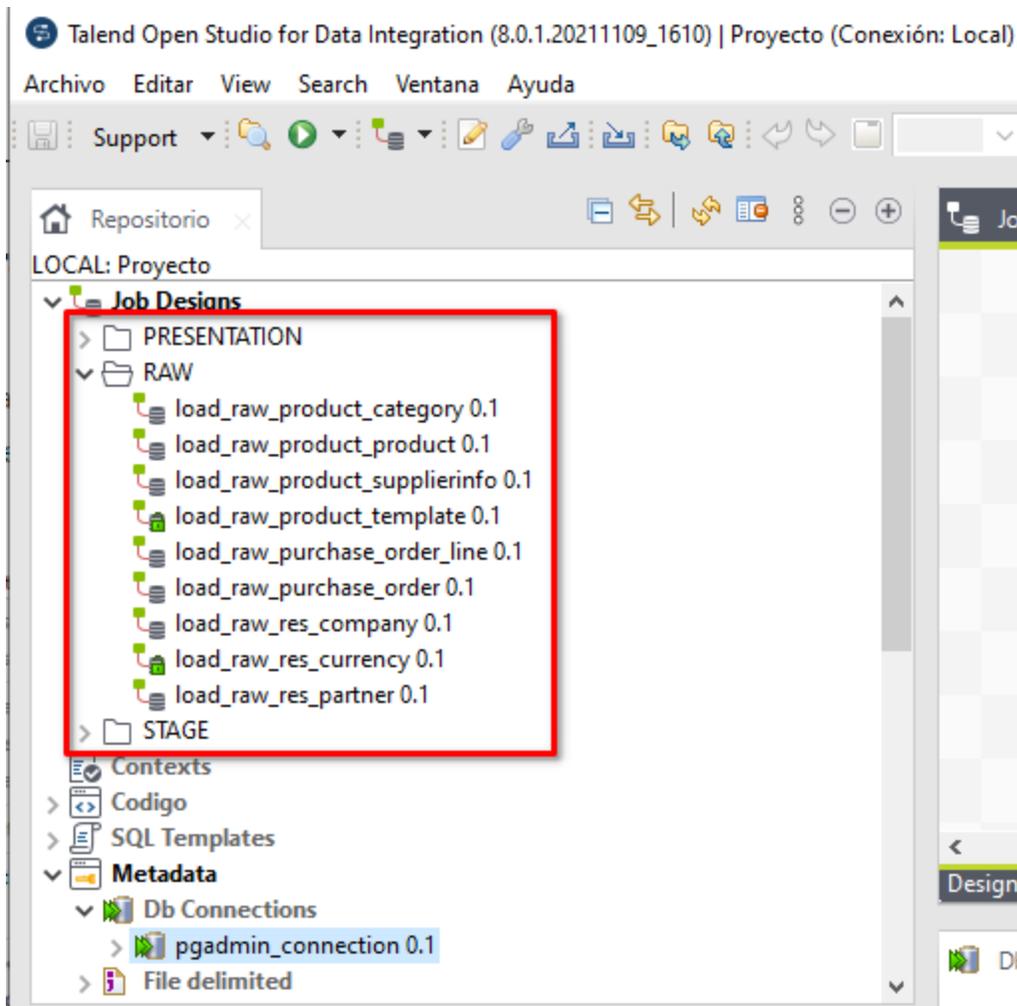


Imagen 14. Estructura de carpetas del proyecto (RAW, STAGE y PRESENTATION).

3. Creación de los diferentes Jobs a utilizar en el ETL, vamos a dirigirnos a la carpeta creada y dar clic derecho para poder seleccionar la opción "Create Job", luego de abrir la ventana para el diseño del Job, tenemos una paleta de opciones donde buscaremos aquellos componentes necesarios para la ejecución del ETL.
 - tPostgresqlInput: Configuración de la conexión con el gestor de base de datos PostgreSQL para poder leer las tablas de la base y extraer sus campos por medio de consultas SQL. La configuración de este componente la podemos visualizar en la **Imagen 15**.

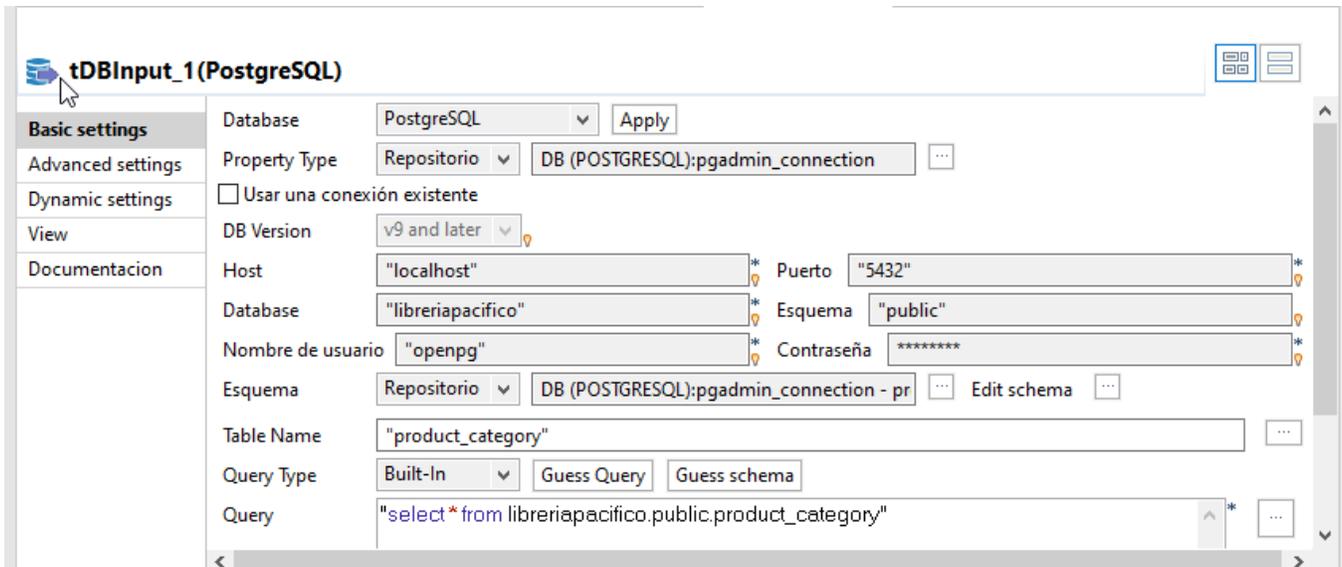


Imagen 15. Configuración de la conexión con PostgreSQL componente tPostgresqlInput.

- tMap: Este componente nos ayuda para poder manejar múltiples entradas y salidas, realizar uniones, transformaciones y más, en su configuración tendremos los datos de entrada con los datos de salida seleccionando los campos que vamos a necesitar mostrar en nuestra salida. La configuración de este componente la podemos visualizar en la **Imagen 16**.

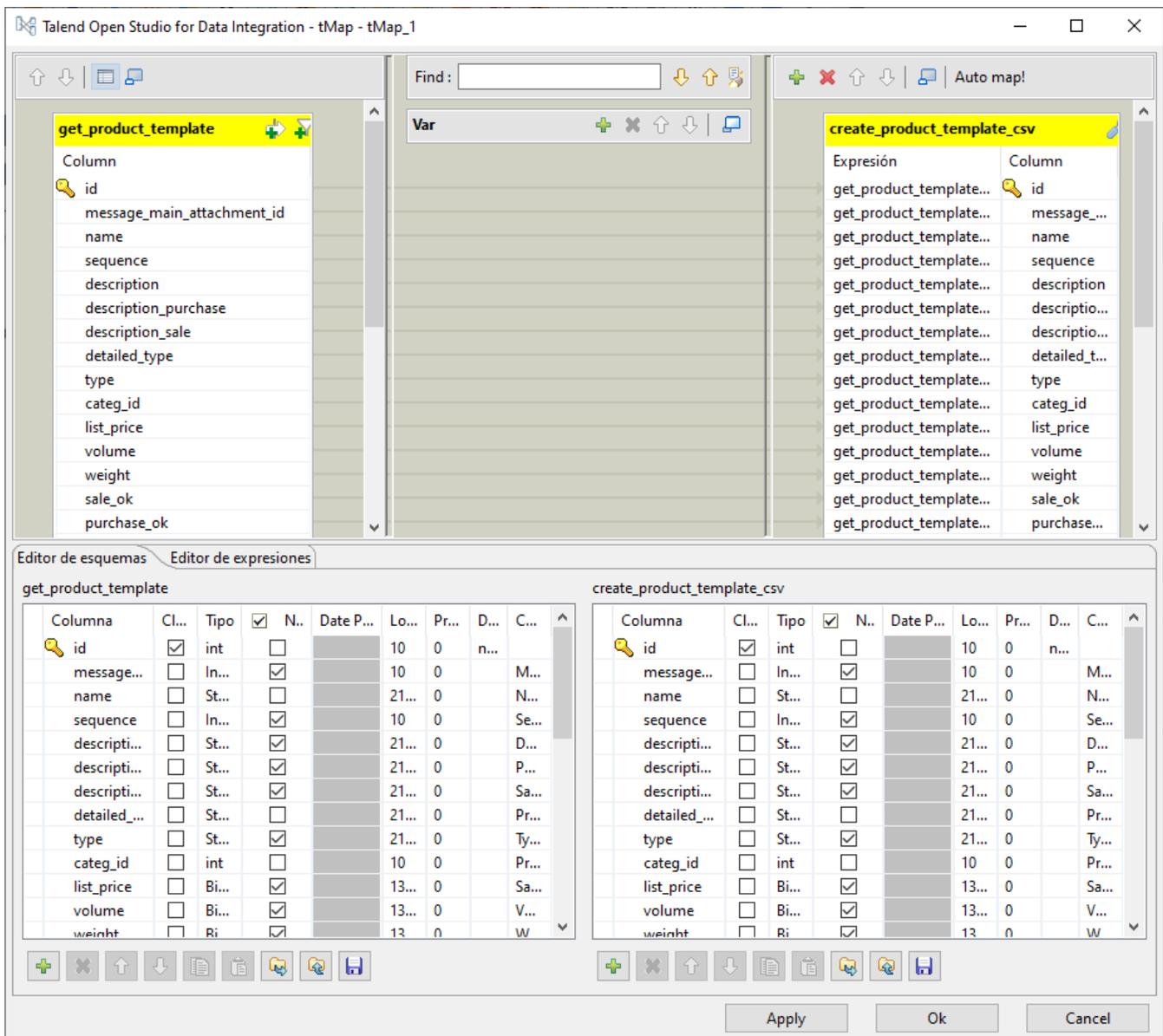


Imagen 16. Map Editor, componente tMap, tabla product_template.

- tFileOutputDelimited:** Este componente lee un archivo delimitado fila por fila para dividirlo en campos y luego envía los campos como se define en el esquema al siguiente componente. Esto quiere decir que genera un archivo csv y se almacena localmente dependiendo la dirección que le agreguen al asignar el nombre del archivo, y posteriormente se envía este archivo plano a nuestro almacenamiento en AWS S3. La configuración de este componente la podemos visualizar en la **Imagen 17**.

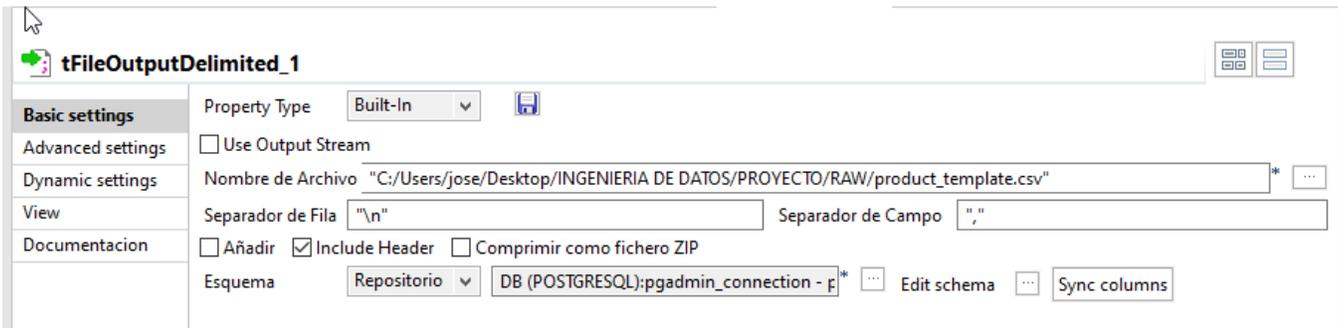


Imagen 17. Configuración componente tFileOutputDelimited, tabla producto_template

- **tS3Connection:** Este componente permite establecer una conexión con Amazon S3 para almacenar y recuperar datos. El componente Standard tS3Connection pertenece a la familia Cloud. Se utilizará para poder realizar una conexión entre Talend Open Studio con AWS S3. La configuración de este componente la podemos visualizar en la **Imagen 18**.

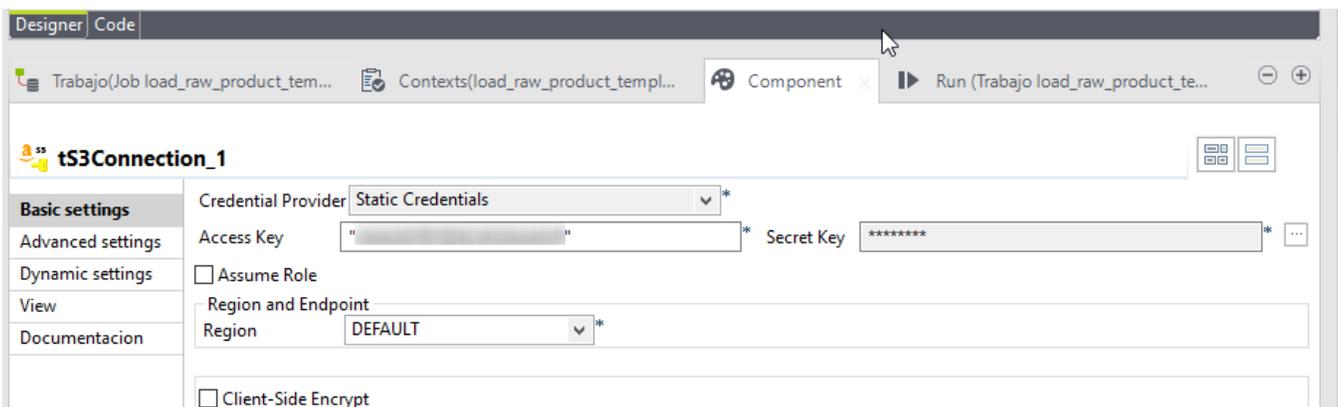


Imagen 18. Configuración componente tS3Connection.

- **tS3Put:** Carga datos en Amazon S3 desde un archivo local o desde la memoria caché a través del modo de transmisión. Para poder hacer un uso correcto de este componente necesitamos tener una conexión existente a AWS S3, se selecciona esa casilla para utilizar la conexión creada en el componente tS3Connection, se define el nombre del Bucket creado en S3 y la dirección en donde se almacenará el archivo csv que tenemos almacenado localmente en nuestra computadora. La configuración de este componente la podemos visualizar en la **Imagen 19**.

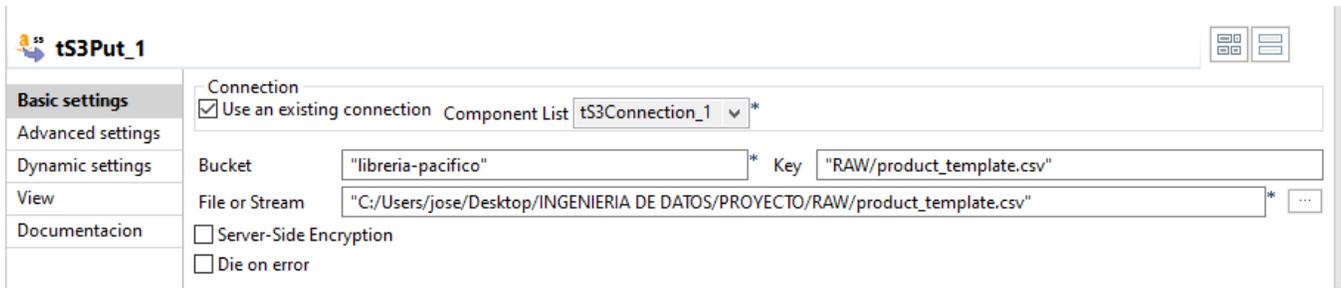


Imagen 19. Configuración del componente tS3Put.

La **Imagen 20**, refleja un ETL, el cual realiza el proceso de conexión a la base de datos de PostgreSQL, extracción de los datos y guardado en un archivo de plano (CSV), la conexión con AWS S3 y finalizando con la carga del archivo de la tabla `producto_category` a la carpeta RAW del Bucket **“libreria-pacifico”** creado en Amazon S3.

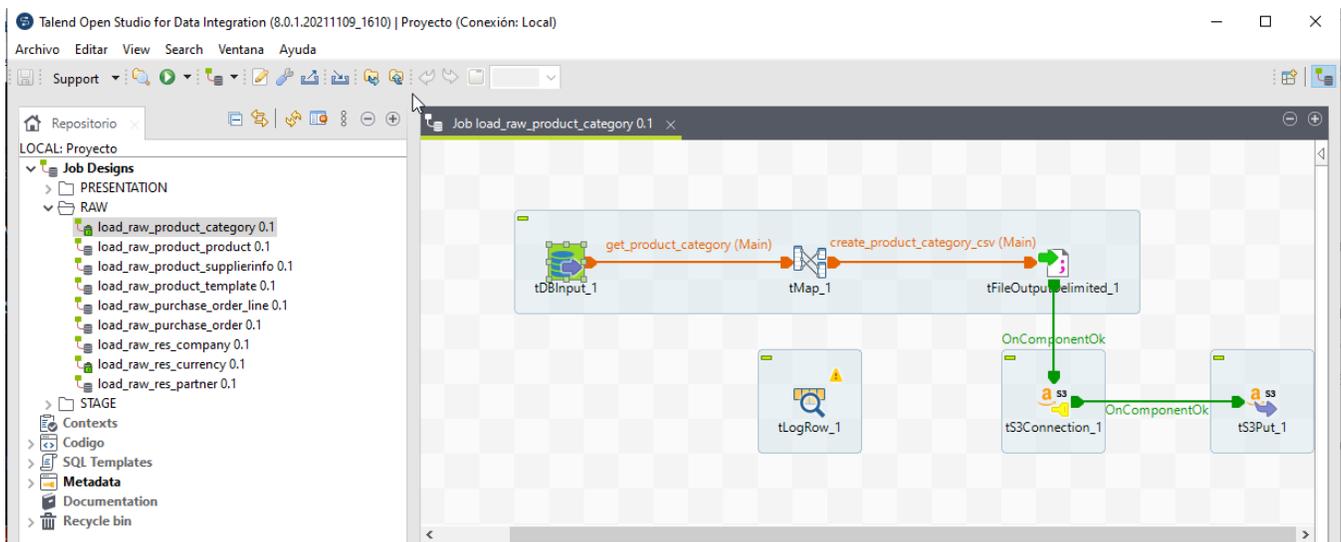


Imagen 20. ETL para obtener un archivo csv de la tabla `product_category` y cargarlo en el bucket de Amazon S3.

Amazon S3

Se creó un bucket llamado “**libreria-pacifico**” en Amazon S3 para la implementación de la solución. La creación del Bucket utilizado en el proyecto lo podemos visualizar en la **Imagen 21**.

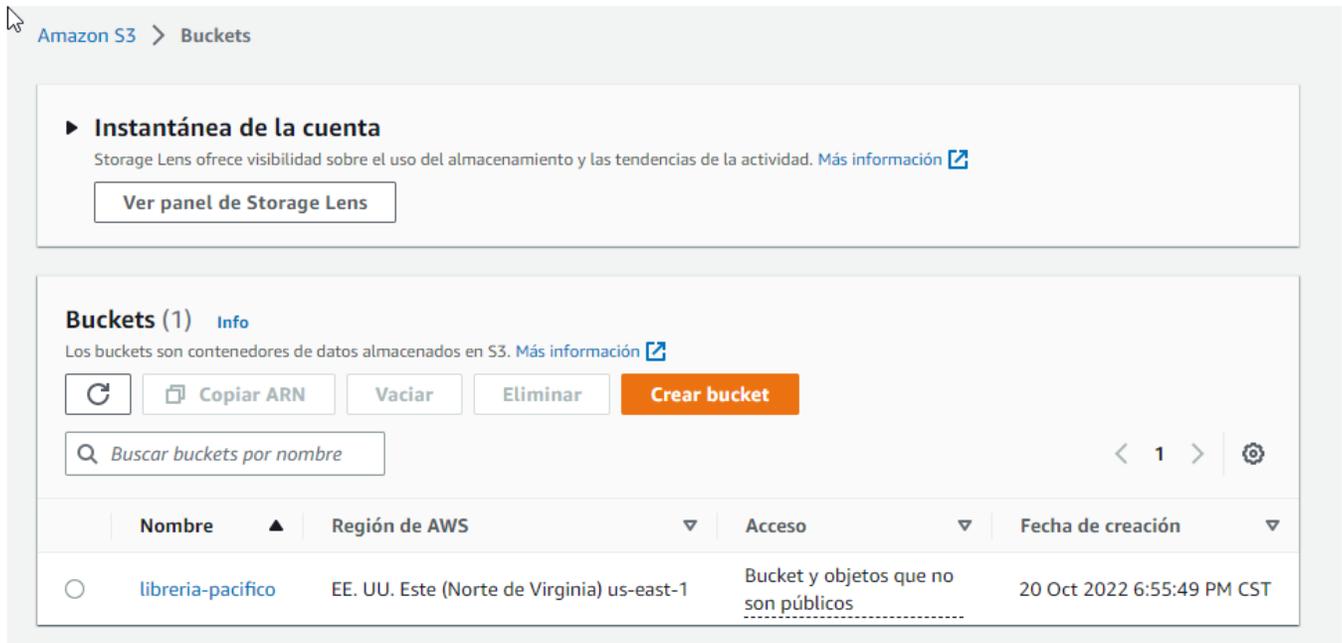


Imagen 21. Bucket librería-pacifico, creado en Amazon S3

Creación de la estructura de almacenamiento en S3 con las siguientes carpetas dentro del bucket:

1. RAW.
2. STAGE.
3. PRESENTATION.

La estructura de las carpetas creadas en Amazon S3 para el almacenamiento en la nube del proyecto, la podemos visualizar en la **Imagen 22**.

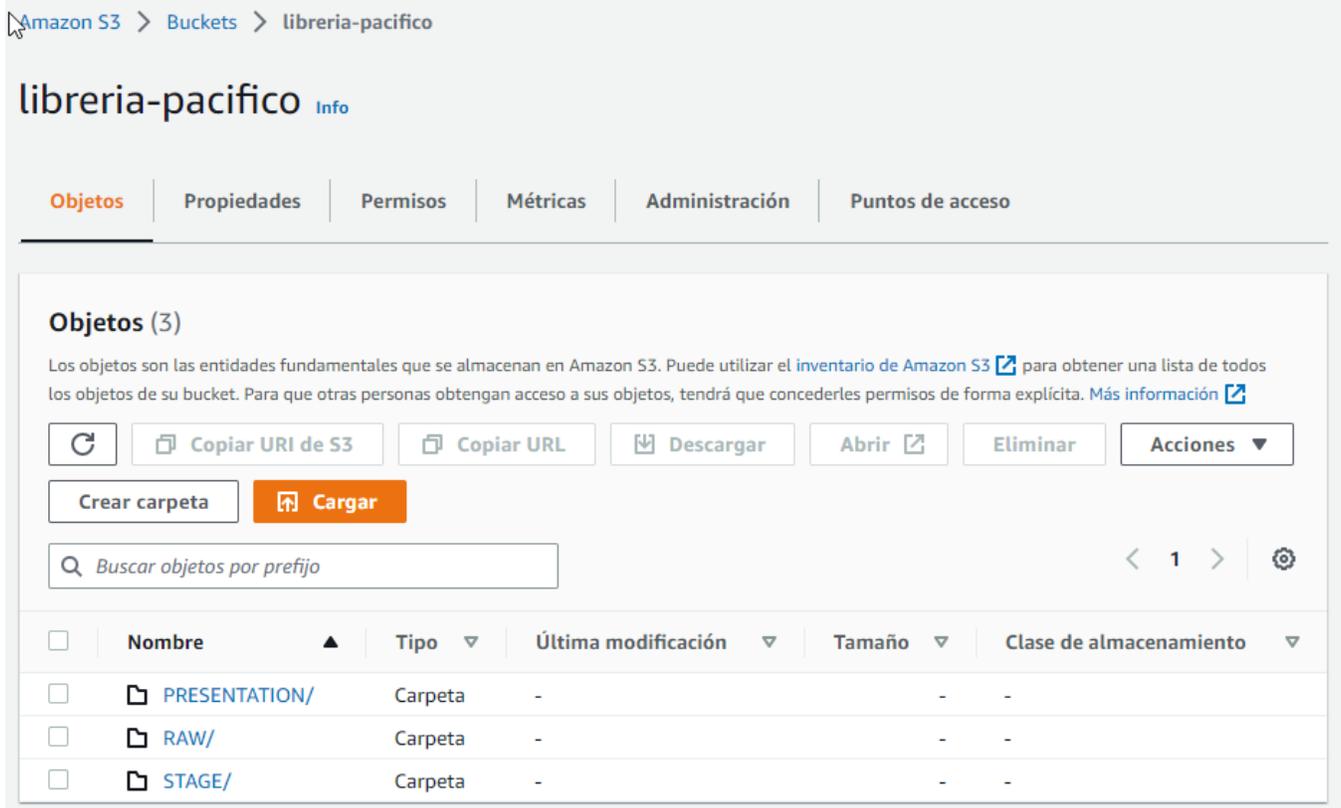


Imagen 22. Estructura de almacenamiento en el Bucket de S3, carpetas RAW, STAGE y PRESENTATION.

Para poder acceder al Bucket de Amazon S3 es necesario realizar una configuración al componente tS3Connection en Talend Open Studio, agregando el Access Key y la Secret Key para lograr el acceso a nuestro bucket “librería-pacifico”, el cual se muestra en la siguiente imagen. La configuración de este componente la podemos visualizar en la **Imagen 23**.

Amazon Redshift

La herramienta de Amazon Redshift se define como un servicio que se encarga de la gestión de procesos de operación, escalado y ajuste del almacenamiento de datos. La opción de Amazon Redshift se utilizará para obtener un almacenamiento de la base de datos en la nube, para ello se llevó a cabo la creación de un clúster en AWS Redshift con el nombre

“**libreriapacifico**”, lo podemos visualizar en la **Imagen 23**, en el cual vamos a poder almacenar los datos de la capa de presentación obtenidos del proceso ETL de Talend Open Studio y su carga de datos en Amazon S3 para su posterior visualización gráfica en Power BI y ayudar a la empresa en sus futuras decisiones estratégicas.

Amazon Redshift > Panel de clústeres aprovisionados

Panel de clústeres aprovisionados Información

[Pruebe Amazon Redshift Serverless](#) [Adquirir nodos reservados](#) [Crear clúster](#)

Información general sobre los recursos

Datos de recursos para la región de US East (N. Virginia).

Total de nodos	Nodos bajo demanda	Nodos reservados	Nodos reservados disponibles	Instantáneas automatizadas	Instantáneas manuales
1	1	0	0 (0 de 0 utilizado)	0	0

Información general acerca del clúster (1)

[Cualquier estado](#)

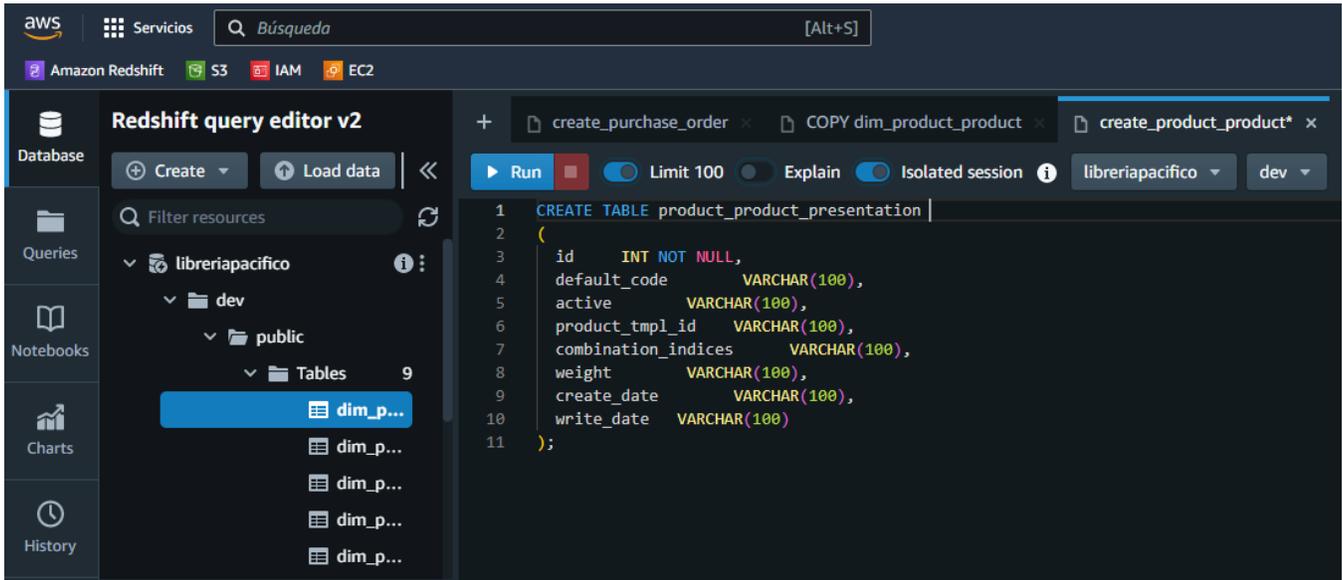
Clúster	Estado
libreriapacifico	⊖ Paused

[Ver todos los clústeres](#)

Imagen 23. Clúster libreriapacifico, creado en Amazon Redshift.

Para lograr la conexión entre Redshift y la capa de PRESENTATION de S3 se ejecutó el siguiente procedimiento:

Ejecutamos un script con el cual creamos la estructura de la tabla de hecho y sus dimensiones. Como ejemplo: Dimensión Producto. El script se puede visualizar en la **Imagen 24**.



The screenshot shows the AWS Redshift Query Editor v2 interface. The top navigation bar includes the AWS logo, 'Servicios', a search bar with 'Búsqueda' and '[Alt+S]', and service icons for Amazon Redshift, S3, IAM, and EC2. The main interface is divided into a left sidebar and a main editor area. The sidebar contains sections for 'Database', 'Queries', 'Notebooks', 'Charts', and 'History'. Under 'Database', there is a tree view showing the hierarchy: 'libreriapacifico' > 'dev' > 'public' > 'Tables' (9 items). One table, 'dim_p...', is highlighted. The main editor area shows a SQL script for creating a table named 'product_product_presentation'. The script is as follows:

```
1 CREATE TABLE product_product_presentation |
2 (
3     id INT NOT NULL,
4     default_code VARCHAR(100),
5     active VARCHAR(100),
6     product_tmpl_id VARCHAR(100),
7     combination_indices VARCHAR(100),
8     weight VARCHAR(100),
9     create_date VARCHAR(100),
10    write_date VARCHAR(100)
11 );
```

Imagen 24. Script de creación de estructura de Dimensión Producto

Luego, ejecutamos un script el cual traerá los datos de la Dimensión Productos que se encuentra en la capa PRESENTATION de S3. Ver script ejecutado para la extracción de los datos de S3 a Redshift en **Imagen 25**.

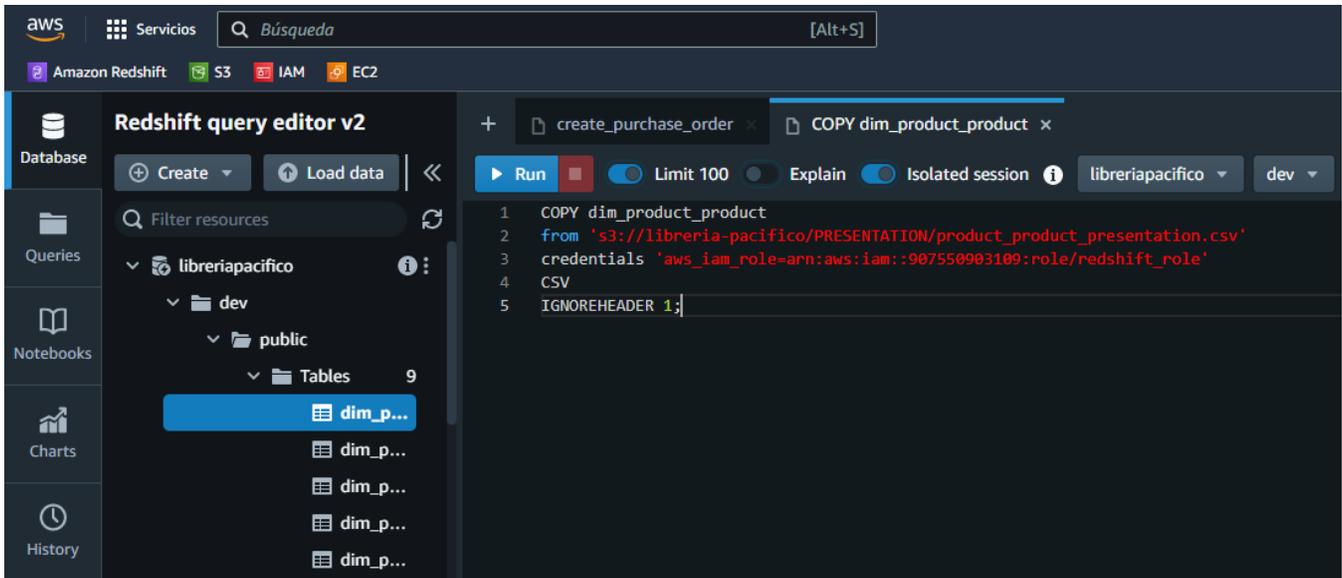


Imagen 25. Extracción de datos de S3 a Redshift.

Power BI

Finalmente se utilizará la herramienta Power BI ya que es una solución de análisis empresarial basado en la nube, y permite unir diferentes fuentes de datos, analizarlos y presentar un análisis de estos a través de informes y paneles. Realizando una conexión con la base de datos de AWS Redshift podemos acceder a los datos almacenados en la base de datos con el modelo del DataWarehouse definido en la implementación de la solución, obtenido del ETL generado con Talend Open Studio y almacenado en la capa de presentación en AWS S3. Estos análisis pueden ser compartidos por diferentes usuarios de la misma organización; por lo que directivos, financieros, comerciales, etc., pueden disponer de la información del negocio en tiempo real y poder tomar decisiones estratégicas con la información obtenida. En la **Imagen 26**, podemos visualizar el primer paso para seleccionar la base de datos de Amazon Redshift en power BI para poder realizar la conexión. La **Imagen 27**, muestra los campos obligatorios para poder conectar Power BI con Amazon Redshift. **Imagen 28**, muestra las tablas de la base de datos en Redshift luego de la conexión.

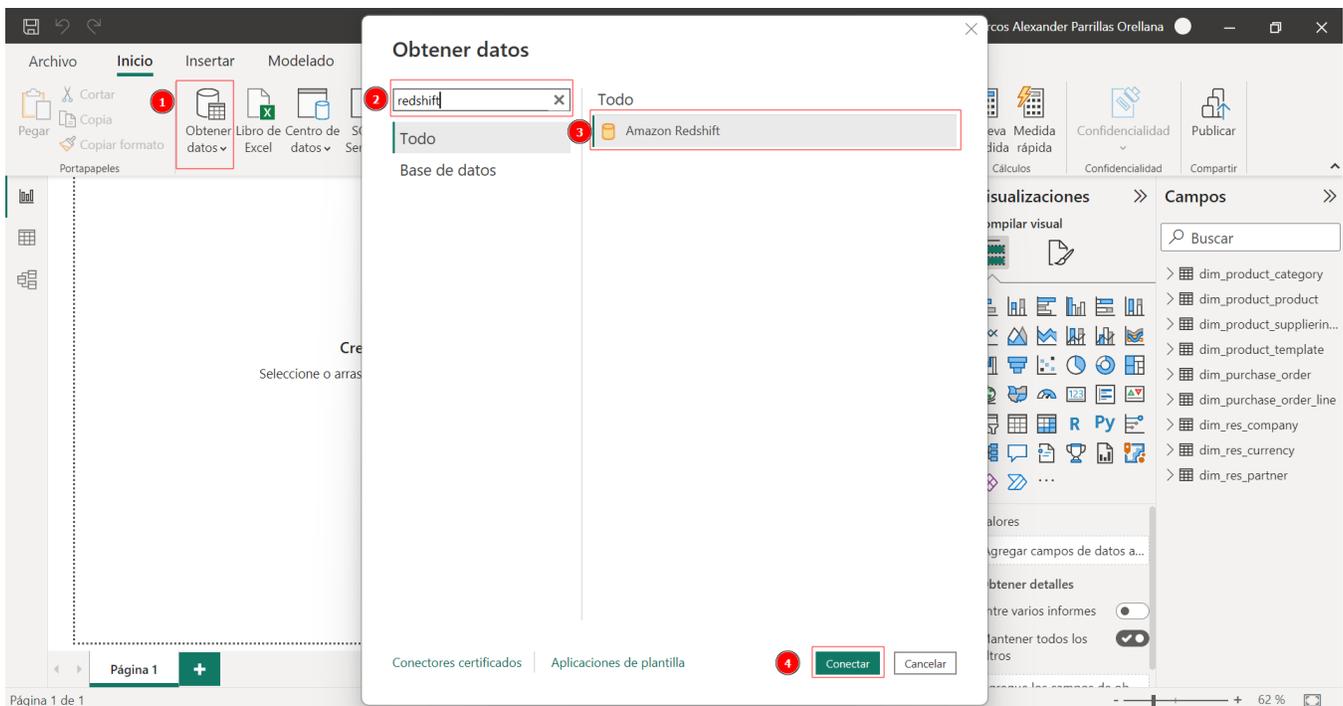


Imagen 26. Conexión Power BI con Redshift



Imagen 27. Conexión con el Clúster de Redshift

Navegador



Opciones de presentación ▾



- Amazon Redshift [3]
 - catalog_history
 - pg_automv
- public [9]
 - dim_product_catalog
 - dim_product_brand
 - dim_product_supplier
 - 1** dim_product_template
 - dim_product_type
 - dim_product_style_line
 - dim_prod_attributes
 - dim_prod_category
 - dim_prod_gender

dim_product_template



Vista previa descargada el jueves, 8 de diciembre de 2022

id	name	sequence	type	categ_id	list_pr
2	El Señor de los anillos	1	service	3	40.
12	luciernagas	1	consu	1	16.
8	Desk Combination	1	consu	8	45.
9	Customizable Desk	1	consu	8	75.
1	Restaurant Expenses	1	service	3	14.
3	Virtual Interior Design	1	service	5	31.
4	Virtual Home Staging	1	service	5	38.
5	Office Chair	1	consu	8	70.
6	Office Lamp	1	consu	8	40.
7	Office Design Software	1	consu	7	28.
10	Corner Desk Right Sit	1	consu	8	14.
11	Large Cabinet	1	consu	8	32.
13	Large Desk	1	consu	8	17.
14	Pedal Bin	1	consu	8	47.
15	Cabinet with Doors	1	consu	8	14.
17	Office Chair Black	1	consu	8	12.
18	Corner Desk Left Sit	1	consu	8	85.
19	Drawer Black	1	consu	8	25.
20	Flipover	1	consu	8	19.
21	Desk Stand with Screen	1	consu	8	21.
22	Individual Workplace	1	consu	8	88.

Seleccionar tablas relacionadas

2

Cargar

Transformar datos

Cancelar

Imagen 28. Carga de Tabla de Hechos y Dimensiones

La implementación de un Data Warehouse proporciona a las empresas una visión más clara y detallada del negocio, lo que permite a los altos mandos tomar las mejores decisiones en el menor tiempo posible. La **Imagen 29**, muestra la representación gráfica del análisis de resultados en la herramienta Power BI.

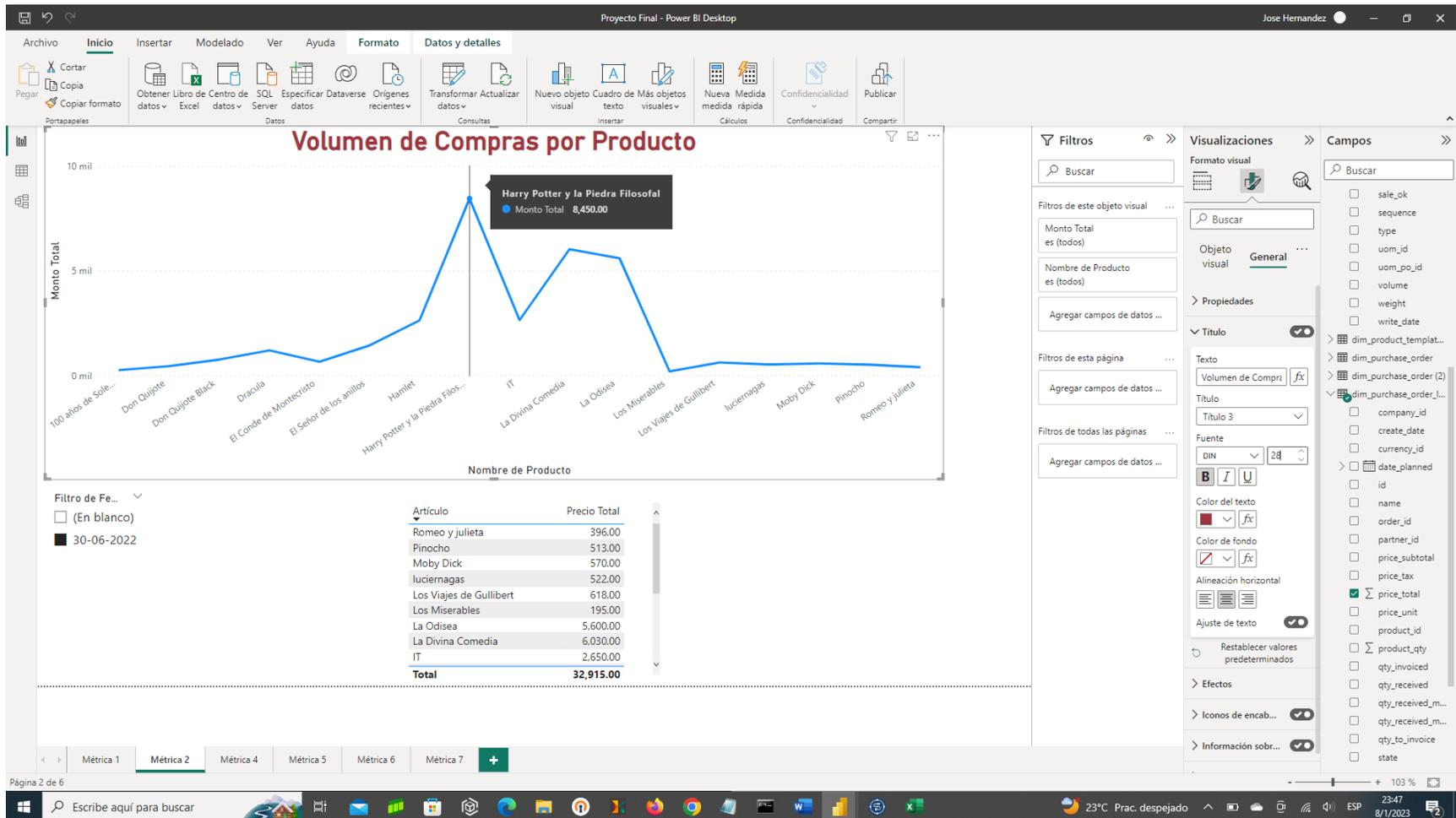


Imagen 29. Dashboard de Power BI

b. Periodicidad y mitigación de la Ejecución de Procesos ETL

Periodicidad

Dentro del proyecto hemos definido el siguiente tipo de carga de datos para la ejecución de los Procesos ETL:

Carga de datos Completa: Se cargará diariamente (a la zona RAW) toda la información de la base de datos transaccional, borrando toda la información previa que existía en el mismo y volviéndola a cargar, de acuerdo con los datos existentes en ese momento en las fuentes de datos.

Mitigación

Problema	Mitigación
Fallas en la ejecución de los ETL en procesos programados.	Monitoreo diario de la correcta ejecución de los procesos ETL, en caso de presentar errores: comunicación con equipo de soporte de sistemas. Los resultados obtenidos deben ser descartados hasta la correcta ejecución de los procesos ETL.
Error en la conectividad con la base de datos transaccional.	La ejecución de los procesos ETL no se llevará a cabo hasta que equipo de soporte de sistemas restablezca la conexión con la base de datos transaccional.
Baja calidad en los datos originarios de la base de datos transaccional.	Retroalimentación al personal operativo sobre las causas de la baja calidad en los datos que alimentan los procesos ETL. Mantener bases de datos que contengan la información replicada de la base de datos transaccional.

Tabla 2. Mitigación de problemas en la ejecución de los Procesos ETL

c. Análisis de resultados

Especificación de Necesidades que el Modelo Propuesto resolverá.

Reporte de cantidad de compras por producto (conteo): Genera un informe que permite visualizar las compras por producto en un rango de fecha seleccionado o una fecha específica. Ver [Imagen 30.](#)



Imagen 30. Cantidad de compras por producto.

Reporte de volumen de compras por producto (dinero): Genera un informe que permite visualizar el monto de las compras por producto en un rango de fecha seleccionado o una fecha específica. Ver **Imagen 31**.

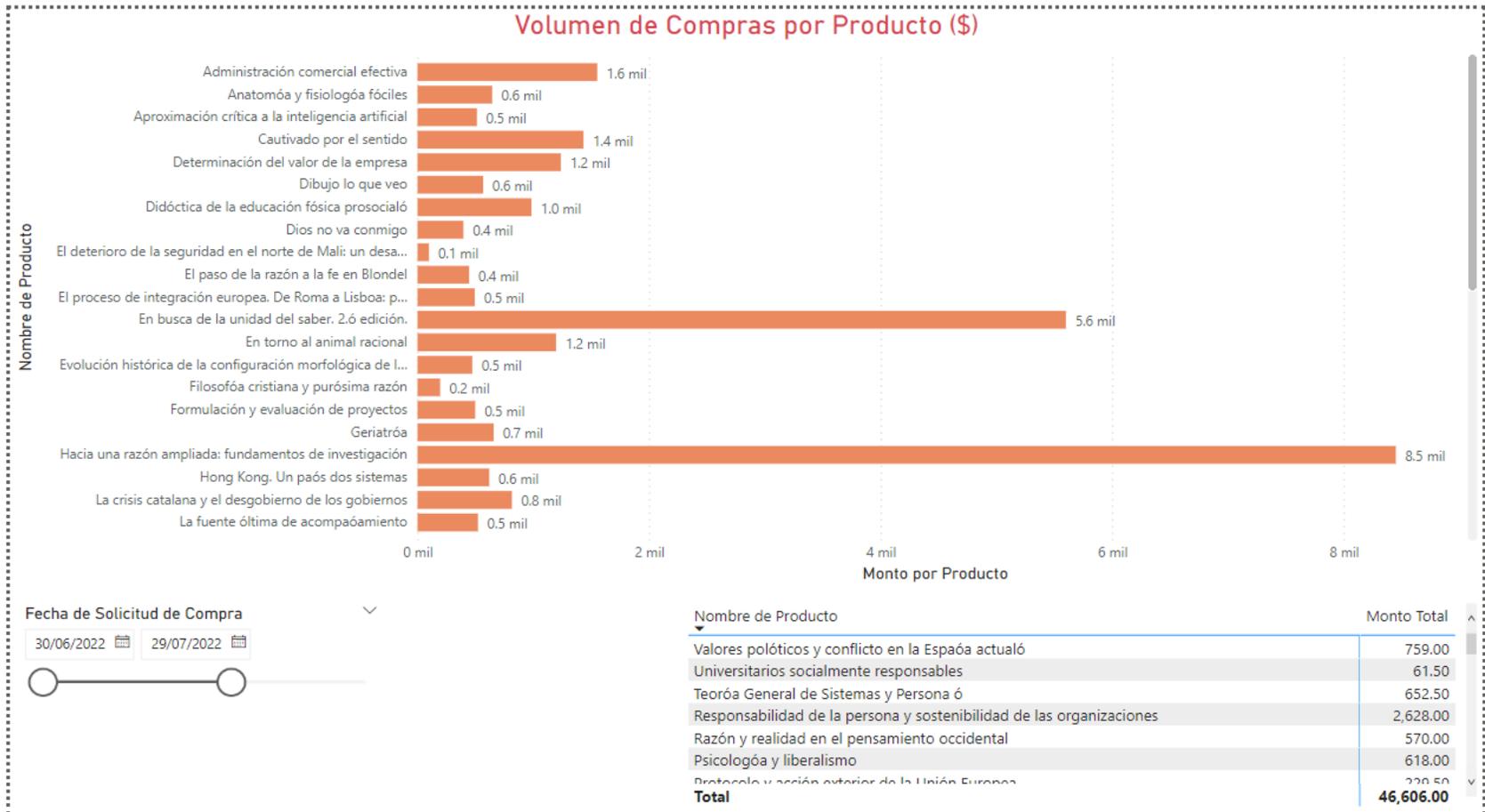


Imagen 31. Volumen de compras por producto (\$).

Reporte de productos por Compañía (conteo): Genera un informe que permite visualizar el total de productos adquiridos por cada compañía en un rango de fecha seleccionado o una fecha específica. Ver **Imagen 32.**

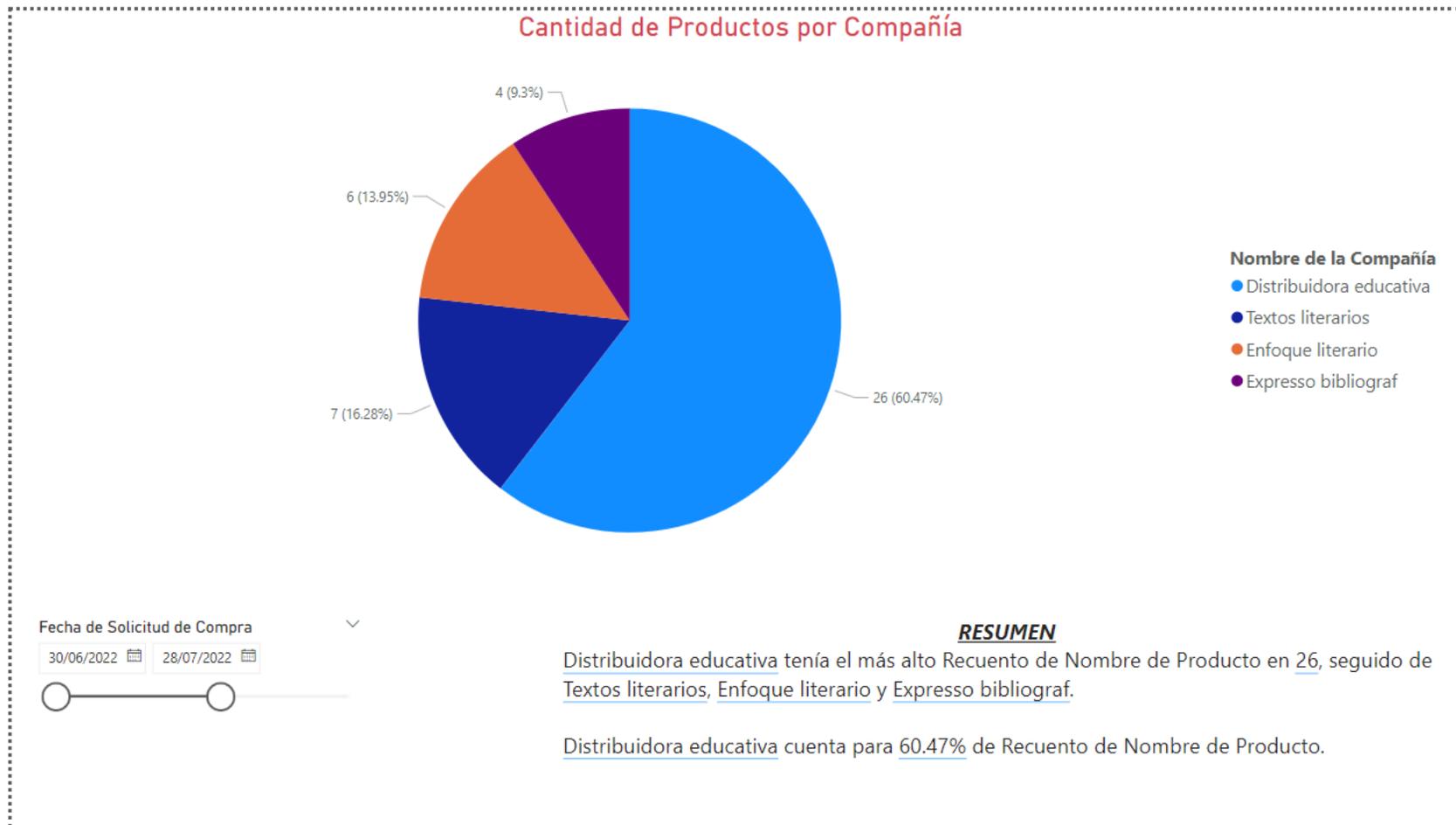


Imagen 32. Cantidad de productos por compañía.

Reporte de productos por proveedor (conteo): Genera un informe que permite visualizar el total de productos adquiridos por cada proveedor en un rango de fecha seleccionado o una fecha específica. Ver **Imagen 33.**

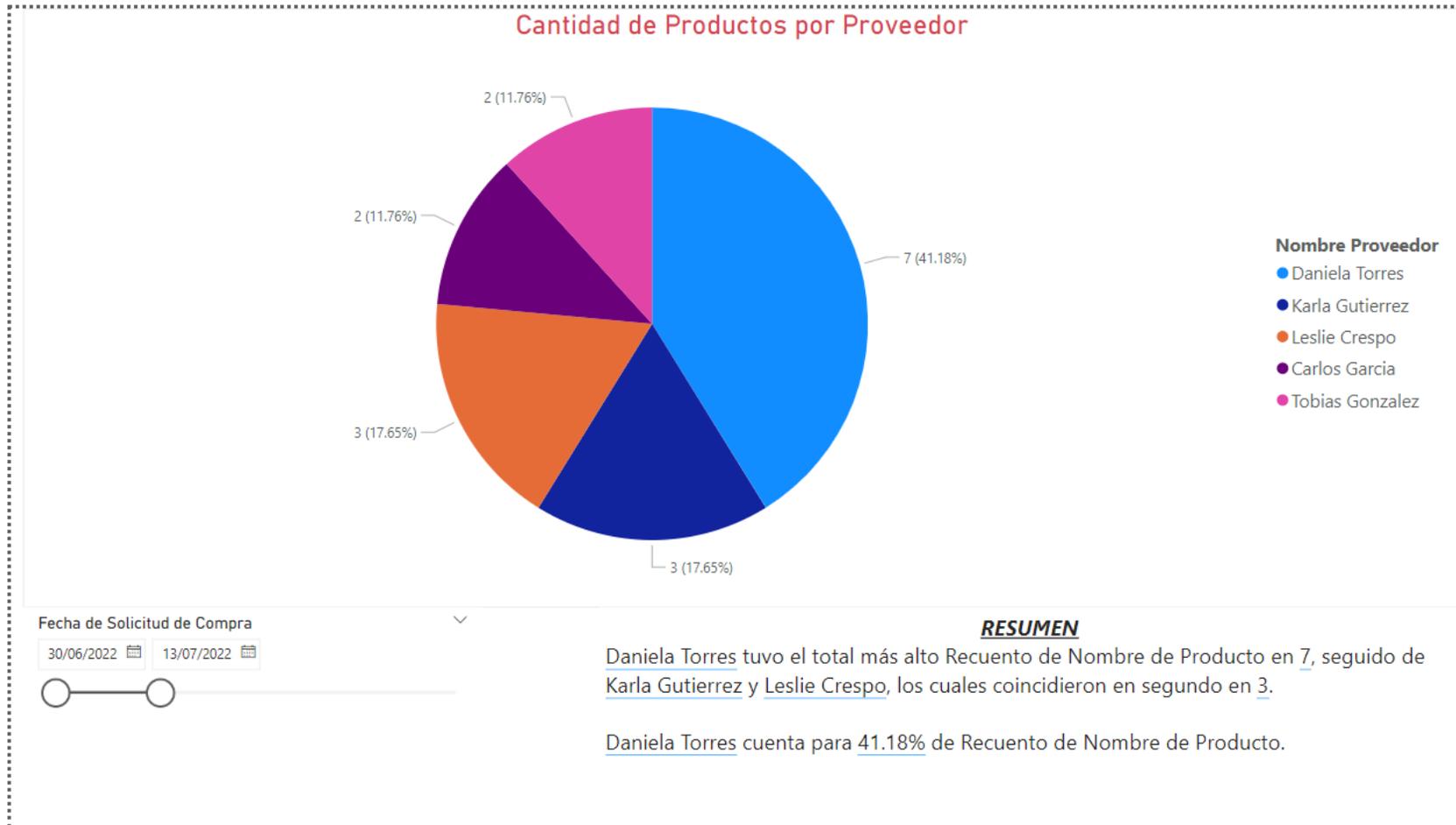


Imagen 33. Cantidad de productos por proveedor.

Reporte de volumen de Productos por Compañía (dinero): Genera un informe que permite visualizar el monto total de los productos adquiridos por cada compañía en un rango de fecha seleccionado o una fecha específica. Ver **Imagen 34.**

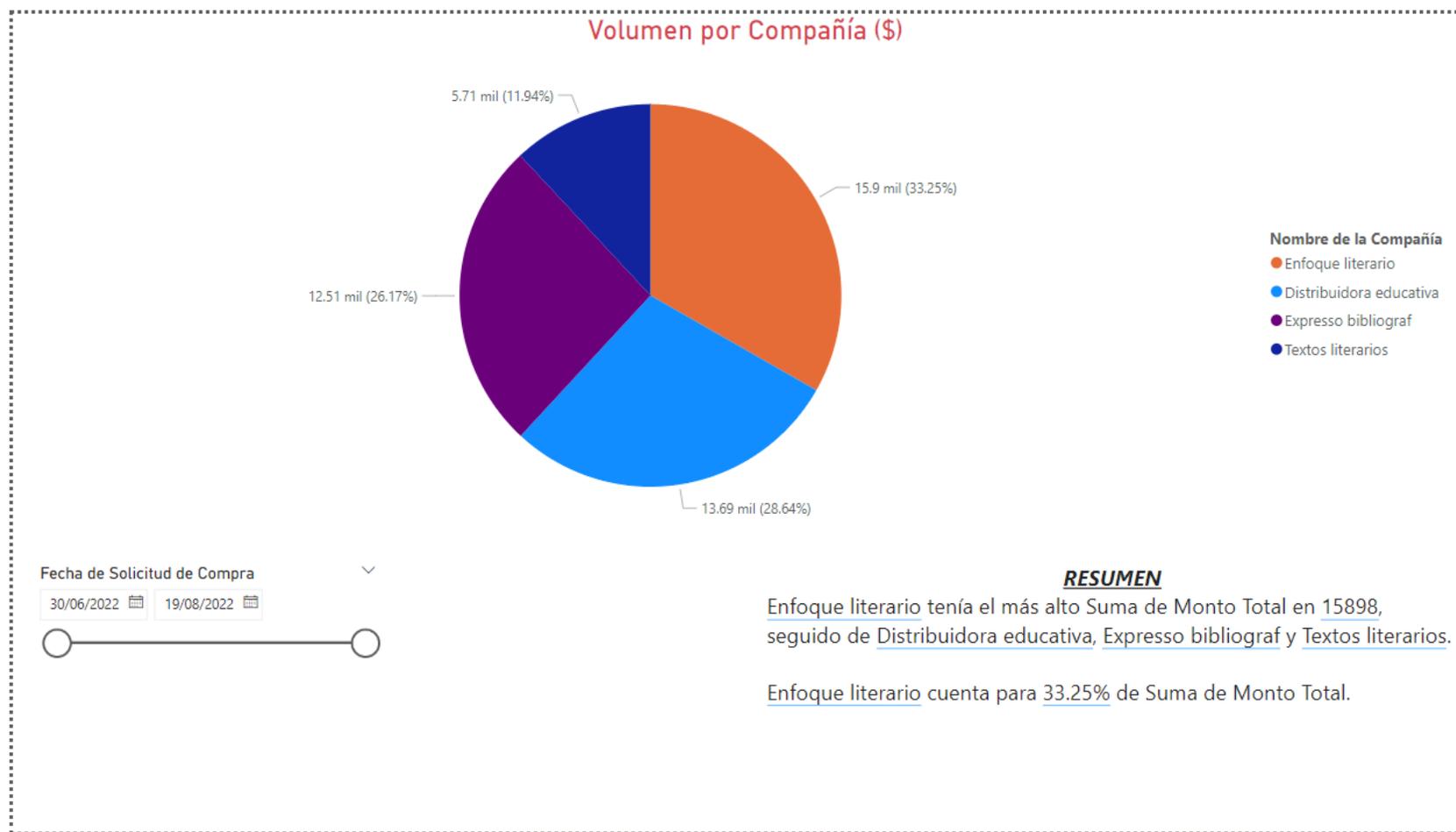


Imagen 34. Volumen por compañía (\$).

Reporte de volumen de Productos por Proveedor (dinero): Genera un informe que permite visualizar el monto total de los productos adquiridos por cada proveedor en un rango de fecha seleccionado o una fecha específica. Ver **Imagen 35.**

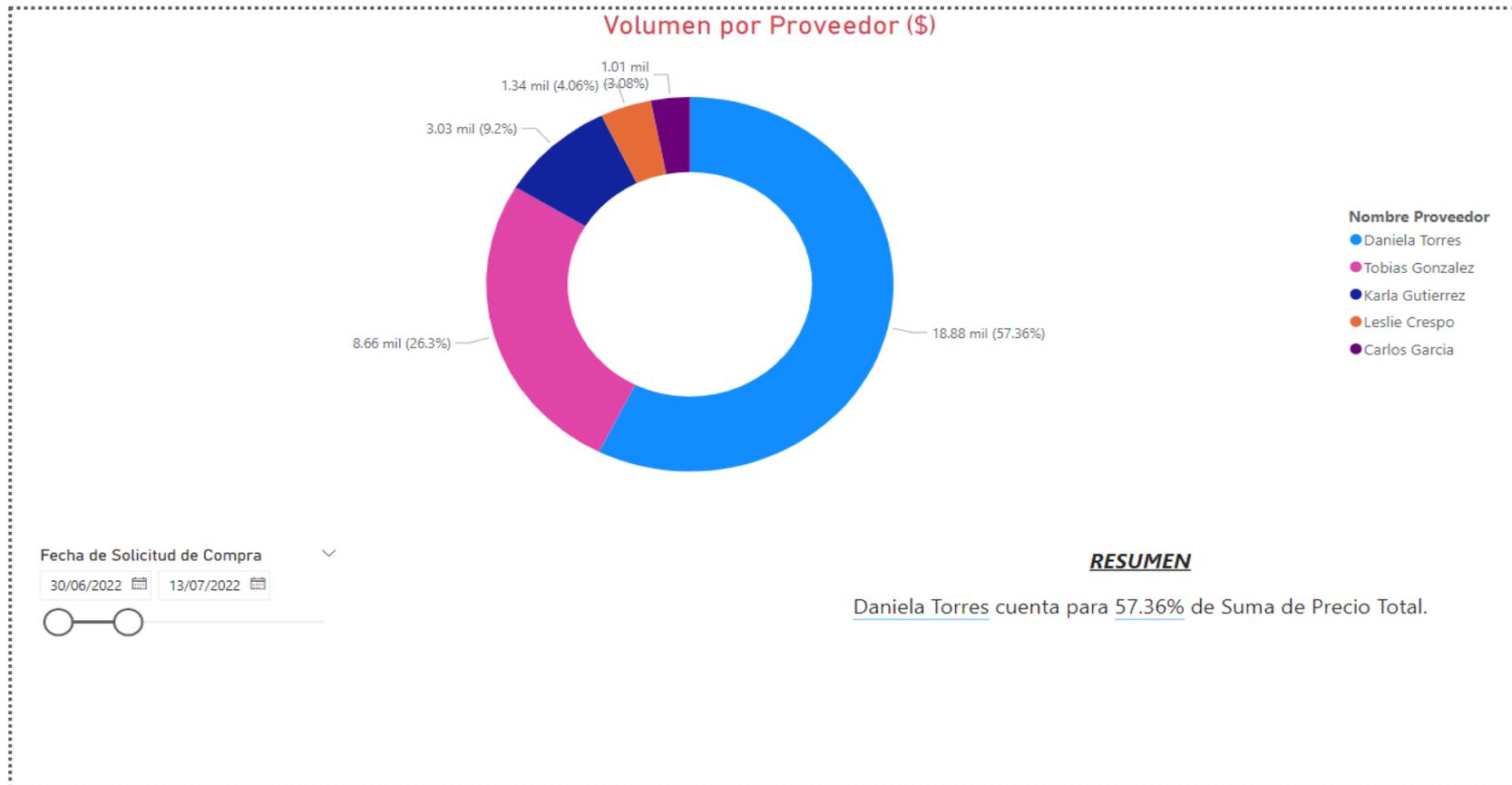


Imagen 35. Volumen por proveedor (\$).

Reporte de Solicitudes de Compra por Compañía (conteo): Genera un informe que permite visualizar la cantidad de solicitudes de compras por cada compañía en un rango de fecha seleccionado o una fecha específica. Ver **Imagen 36.**

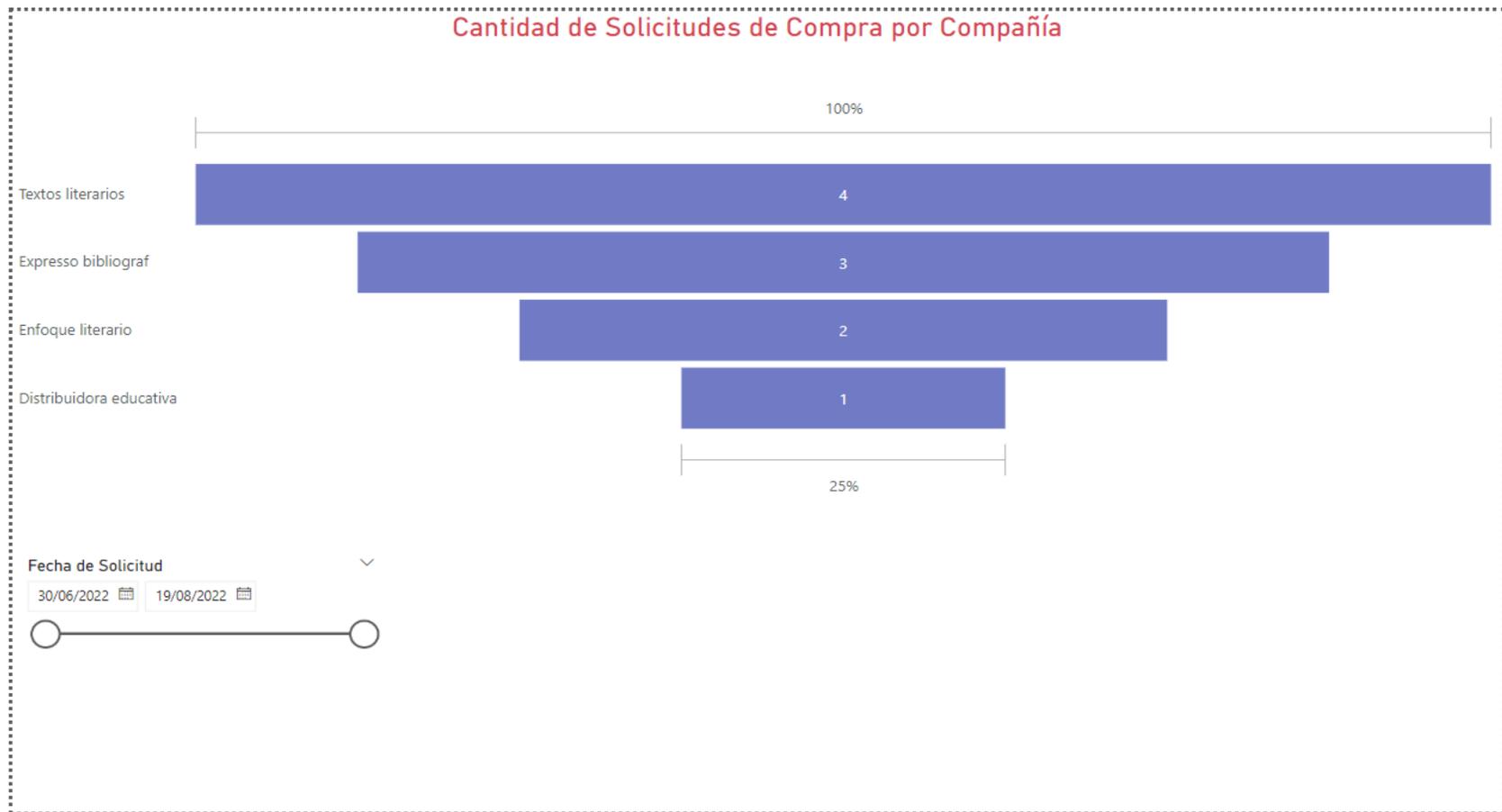


Imagen 36. Cantidad de solicitudes de compra por compañía.

Reporte de Solicitudes por estado (conteo): Genera un informe que permite visualizar la cantidad de solicitudes de compra dependiendo del estado en que se encuentra en un rango de fecha seleccionado o una fecha específica. Ver **Imagen 37.**

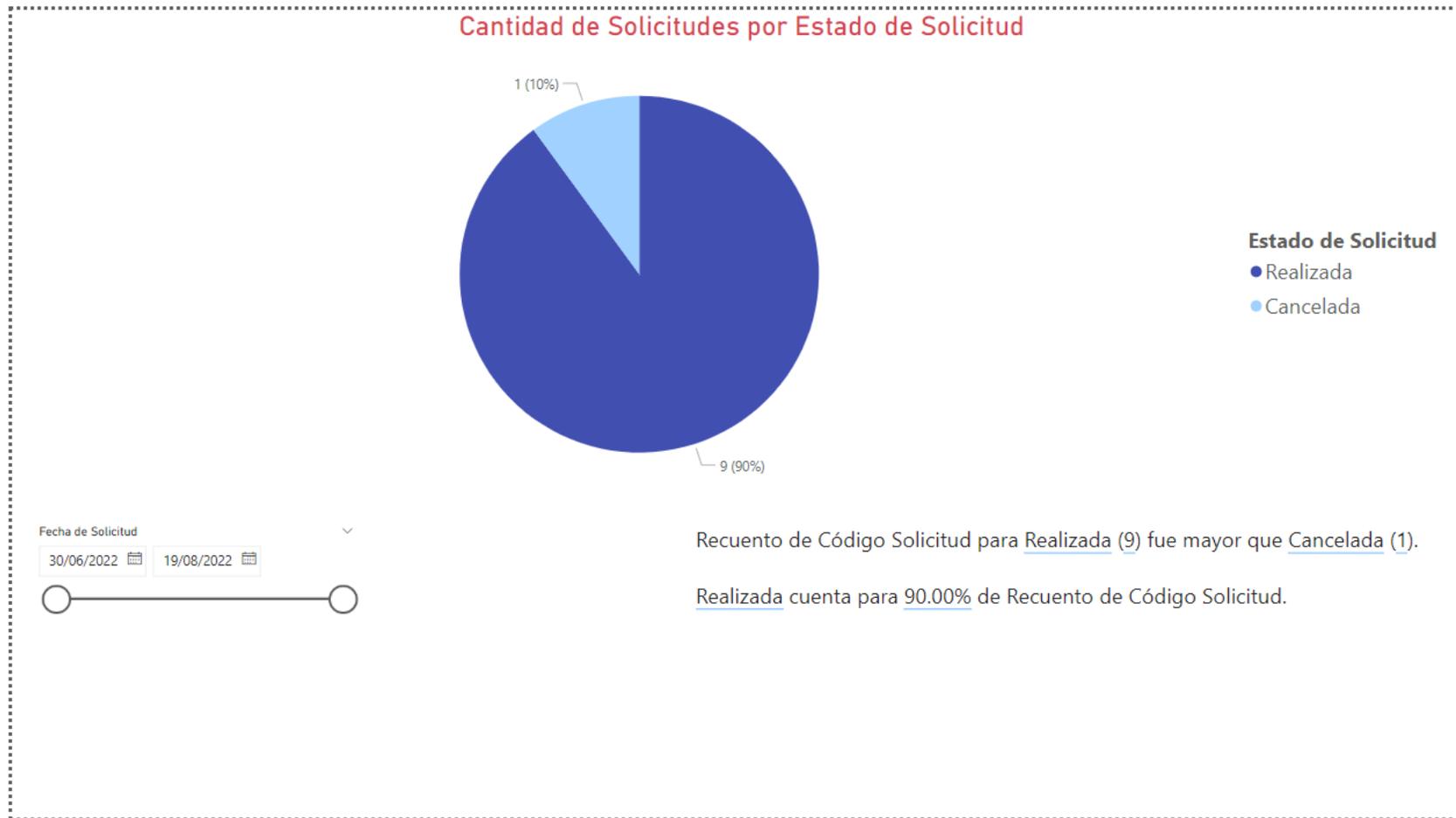


Imagen 37. Cantidad de solicitudes por estado de solicitud.

Conclusiones y recomendaciones

Conclusiones

- El análisis y comprensión inicial del sistema informático transaccional de la empresa es clave para determinar el camino que definirá el éxito de los reportes finales, debido a, de darse el caso de un mal análisis, los datos que lleguen a la herramienta de BI podrán evidenciar la carencia de información o la incongruencia de los datos; por tanto, diseñar un buen diagrama de ER de sistema transaccional es parte importante de todo el proceso.
- Concluimos que los procesos ETL son la parte más fundamental del diseño de la solución, ya que, de cierta manera, son el puente entre los grandes volúmenes de datos que genera el sistema informático transaccional, con los cuales resulta imposible analizarlos y tomar decisiones con base en ellos; y finalizan en la herramienta de Business Intelligence, la cual cumple el objetivo de apoyar en la toma de decisiones de la parte gerencial de la empresa. Todo esto es posible mediante el proceso ETL que sufren los datos desde su concepción hasta su culminación en la herramienta de BI.
- Debido al gran volumen de datos que generan los sistemas transaccionales de las empresas, es primordial contar con una potente herramienta de almacenamiento (S3), con gran disponibilidad, seguridad y soporte todos los procesos de ETL de la solución. De la misma manera se debe contar con una herramienta que soporte consultas de análisis de datos (Redshift), con un alto rendimiento a través de las mismas herramientas basadas en SQL y aplicaciones de inteligencia empresarial que la solución implemente.
- Debido a que la culminación del proyecto resulta ser la más importante para los usuarios finales, es de suma importancia que la herramienta de BI sea una herramienta fácil de interpretar por los usuarios y que permita a las organizaciones crear cuadros de mando basados en datos, colaborativos y adaptables para realizar un seguimiento de las métricas empresariales clave.

Recomendaciones

Para que un Data Warehouse sea exitoso, necesita ser parte funcional de las operaciones de una empresa y evolucionar a medida que el negocio también crezca. Por lo tanto, una buena documentación de cómo se ha venido desarrollando el almacén de datos y los logros que se han alcanzado gracias a él garantizarán su continuidad.

Bibliografía

- Kimball. (2013). Business Intelligence Applications. In R. & Kimball, *The data warehouse toolkit (3rd ed.)* (pp. 22-26). Indianapolis: Wiley.
- Kimball. (2013). Dimension Tables for Descriptive Context. In R. & Kimball, *The data warehouse toolkit (3rd ed.)* (pp. 13-15). Indianapolis: Wiley.
- Kimball. (2013). Dimensional Modeling Introduction. In R. & Kimball, *The data warehouse toolkit (3rd ed.)* (pp. 7-8). Indianapolis: Wiley.
- Kimball. (2013). Extract, Transformation, and Load System. In R. & Kimball, *The data warehouse toolkit (3rd ed.)* (pp. 19-21). Indianapolis: Wiley.
- Kimball. (2013). Fact Tables for Measurements. In R. & Kimball, *The data warehouse toolkit (3rd ed.)* (pp. 10-12). Indianapolis: Wiley.
- Kimball. (2013). Facts and Dimensions Joined in a Star Schema. In R. & Kimball, *The data warehouse toolkit (3rd ed.)* (pp. 16-18). Indianapolis: Wiley.
- Kimball. (2013). Kimball's DW/BI Architecture. In R. & Kimball, *The data warehouse toolkit (3rd ed.)* (pp. 18-22). Indianapolis: Wiley.
- Kimball. (2013). Operational Source Systems. In R. & Kimball, *The data warehouse toolkit (3rd ed.)* (pp. 18-19). Indianapolis: Wiley.
- Kimball. (2013). Presentation Area to Support Business Intelligence. In R. & Kimball, *The data warehouse toolkit (3rd ed.)* (pp. 21-22). Indianapolis: Wiley.
- Kimball. (2013). Star Schemas Versus OLAP Cubes. In R. & Kimball, *The data warehouse toolkit (3rd ed.)* (pp. 8-18). Indianapolis: Wiley.
- Urrutia, D. (2021b, agosto 5). *Qué es Power Bi - Definición, significado y ejemplos*. Arimetrics. <https://www.arimetrics.com/glosario-digital/power-bi>
- KeepCoding, R. (2022, 23 mayo). *¿Qué es Talend Open Studio?* KeepCoding Tech School. <https://keepcoding.io/blog/talend-open-studio/>
- FreshCommerce, E. (2022, 26 julio). *Herramientas ETL: Talend - Parte III*. FreshCommerce. <https://freshcommerce.es/blog/herramientas-etl-talend/>
- KeepCoding, R. (2022b, julio 12). *¿Qué es Amazon Redshift?* KeepCoding Tech School. <https://keepcoding.io/blog/que-es-amazon-redshift/>
- KeepCoding, R. (2022c, julio 13). *Características de Amazon S3*. KeepCoding Tech School. <https://keepcoding.io/blog/caracteristicas-de-amazon-s3/>
- ¿Qué es IAM? - AWS Identity and Access Management.* (s. f.). https://docs.aws.amazon.com/es_es/IAM/latest/UserGuide/introduction.html

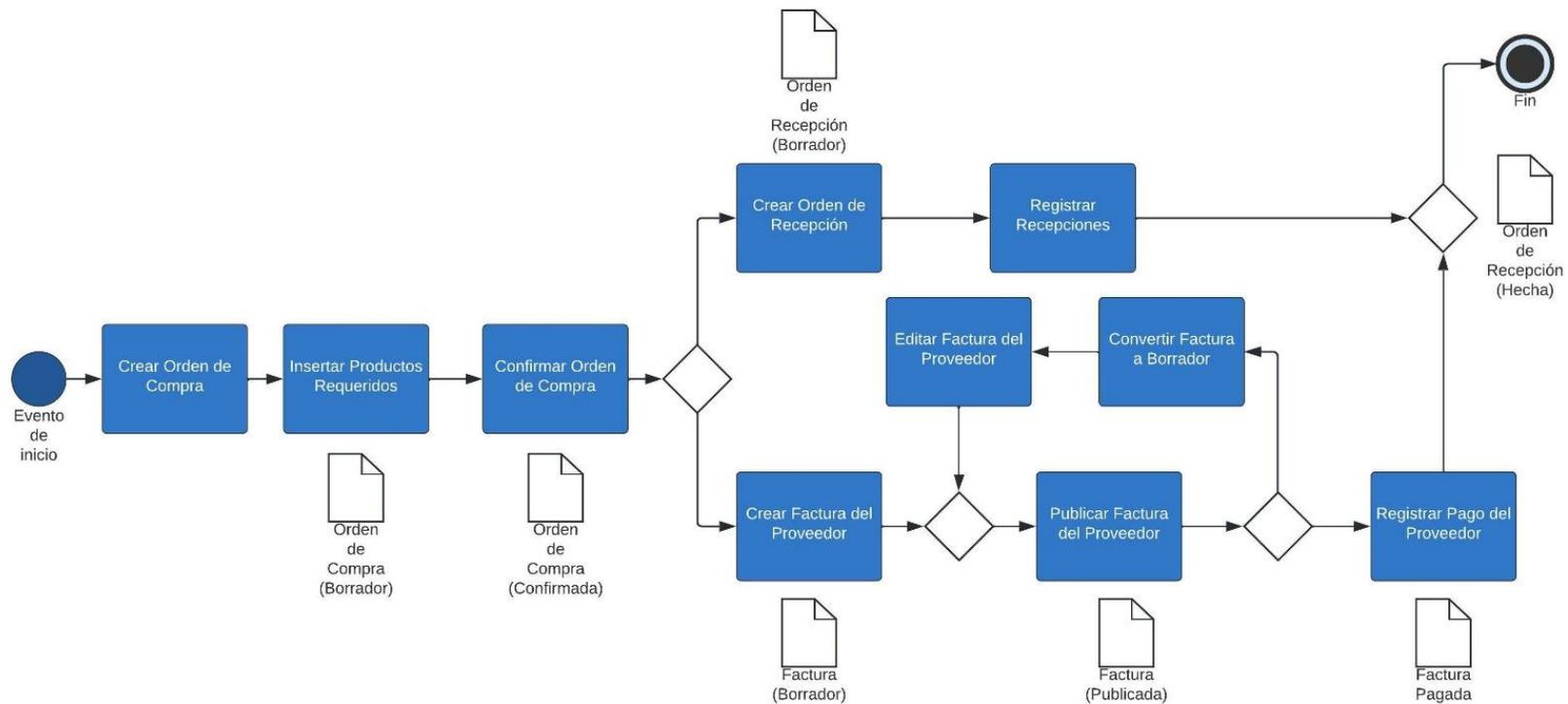
Anexos

a. BPMN Proceso de Compra

Para una mejor comprensión, se presenta el siguiente BPMN que describe el proceso de compras del sistema transaccional de la empresa Librería del Pacífico.

Flujo Estándar del Proceso de Compra

Grupo 04



b. Cronograma de actividades

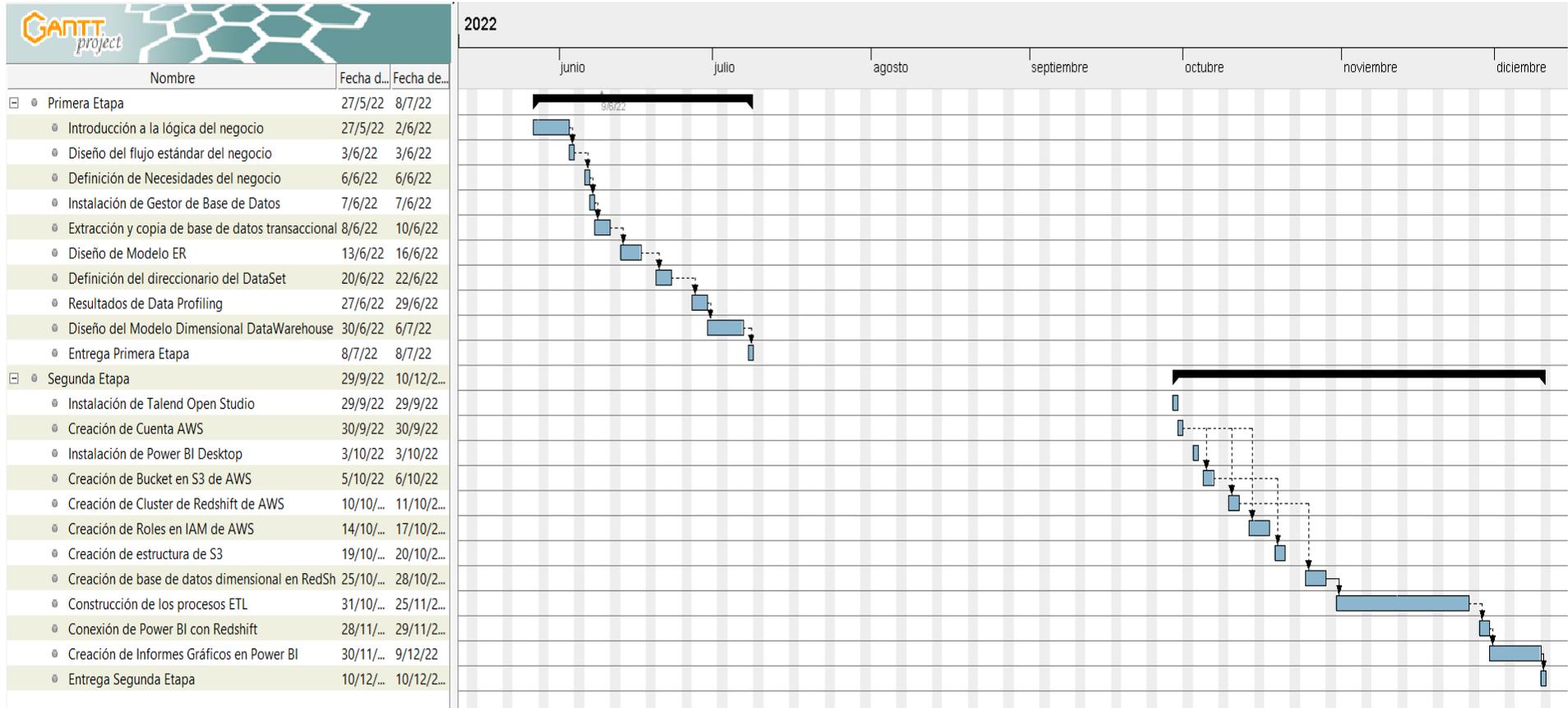


Imagen 38. Cronograma de Actividades

c. Sucursales de Librería el Pacífico

Librería El Pacífico cuenta con diez sucursales los cuales se encuentran distribuidas a lo largo del territorio salvadoreño, estas están ubicadas en:

1. Zona Central:

San Salvador: 25 Av. Sur y 6ª. C. Poniente, Col. Flor Blanca, San Salvador.

Santa Tecla: Carr. Panamericana & 6 Avenida Sur, Santa Tecla.

Nuevo Cuscatlán: Calle Antigua a Huizúcar, Nuevo Cuscatlán,

Zacatecoluca: Km. 63 Carretera Litoral, Cantón Penitente Abajo, Jurisdicción de Zacatecoluca, Zacatecoluca.

San Vicente: 5A Calle Poniente #25, San Alejo.

2. Zona Occidental:

Santa Ana: 25 C. Poniente y 18 Av. Sur, Santa Ana.

Sonsonate: Km. 64, Blvd. Las Palmeras, Sonzacate.

Ahuachapán: KM 99 Carretera a Santa Ana, Colonia San José.

3. Zona Oriental:

San Miguel: 3ª. Av. Norte y 8ª. C. Poniente, Frente ex-cine Mundial., San Miguel.

Usulután: CA-2 Km 128, Santa María

d. Presupuesto de Desarrollo

AWS

La empresa debe contar con un almacenamiento en la nube, para esta solución se tomará en cuenta Amazon Web Services, del cual se optará por usar los servicios de IAM, Redshift y S3 que nos ofrece esta plataforma. Los precios estimados mensuales y anuales se obtuvieron por medio de AWS Pricing Calculator un servicio con el que cuenta la plataforma, el presupuesto se detalla a continuación:

Título estimado: **Presupuesto AWS**

Resumen de la estimación		
Costo inicial	Costo mensual	Costo total de 12 months
0,00 USD	218,34 USD	2620,08 USD
		Incluye el costo inicial

Estimación detallada

Nombre	Grupo	Región	Costo inicial	Costo mensual
Amazon Redshift	No se ha aplicado ningún grupo	US East (Ohio)	0,00 USD	206,45 USD
Descripción: Resumen de la configuración: Nodos (1), Tipo de instancia (dc2.large), Utilización (solo bajo demanda) (100 %Utilized/Month), Modelo de precios (OnDemand), Almacenamiento de copias de seguridad adicional (1 TB), Transferencia de datos de entrada a (20 GB)				
Amazon Simple Storage Service (S3)	No se ha aplicado ningún grupo	US East (Ohio)	0,00 USD	11,89 USD
Descripción: Resumen de la configuración: Almacenamiento de S3 Estándar (20 GB por mes) DT Entrada: Internet (100 GB al mes), DT Salida: Internet (100 GB al mes)				

Imagen 39. Presupuesto estimado obtenido de AWS Pricing Calculator.

RRHH

Cada integrante del proyecto recibirá un salario por los servicios profesionales realizados a la empresa Librería del Pacífico, el cual se detalla en la siguiente tabla:

Empleado	Mes Inicio	Mes Fin	Salario Mensual	Duración del proyecto (Meses)	Salario Total
José Ernesto Hernández Vásquez	Junio	Diciembre	\$800	6	\$4,800
Marcos Alexander Parrillas Orellana	Junio	Diciembre	\$800	6	\$4,800
				Presupuesto TOTAL	\$9,600

Tabla 3. Presupuesto del Proyecto con relación a Recursos Humanos.

Presupuesto Mensual

Costo de Servicio	AWS	RRHH	Presupuesto Mensual
Costo	\$218.34	\$1,600	\$1,818.34

Tabla 4. Presupuesto Mensual en relación con los servicios de AWS y Recursos Humanos.

e. Presupuesto de implementación

La empresa Librería del Pacífico cuenta un servicio en línea para poder realizar los procesos de ventas y con acceso a internet por lo cual no es necesario realizar una inversión en el servicio de internet. Es necesario contratar servicios profesionales para la implementación de la solución, y posteriormente capacitaciones a los recursos de la empresa en el uso correcto de las herramientas necesarias para la visualización de los informes estratégicos que ayuden en la toma de decisiones. Dicho costo se realizará tomando en cuenta la implementación y capacitación que se llevará a cabo en un lapso de 15 días.

Se realizará el cálculo tomando en cuenta el salario otorgado por la empresa que es de \$800 mensuales, se realizará por ambos integrantes del proyecto. Realizando los cálculos correctos determinamos que el salario diario de cada persona es de \$26.66 por día y por hora trabajada \$3.33.

Recurso Humano

Recursos Humanos	Precio por Hora	Cantidad de horas para la implementación	Costo Total
2	\$3.33	120	\$799.20

Tabla 5. Presupuesto de implementación de los Recursos Humanos.

Hardware

Es necesario que la empresa cuente con un equipo que cumpla con los recursos mínimos y que sea capaz de soportar la ejecución simultanea de las diversas herramientas utilizadas en el desarrollo de la solución. Tomando en cuenta esto se optó por la compra de un equipo con las siguientes características:

- INTEL CORE i5-4570.
- RAM 8GB.
- SSD 240GB.

Precio valorado de \$329.00