

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



**ANÁLISIS DE VENTAS DE LIBRERÍA PACÍFICO POR MEDIO DE TALEND Y  
POWERBI**

**PRESENTADO POR:**

**CHRISTIAN ALEXIS PÉREZ FUENTES**

**LUIS FELIPE SERRANO AMAYA**

**JESSICA MARCELA ZA VALETA PADILLA**

**PARA OPTAR AL TÍTULO DE:**

**INGENIERO DE SISTEMAS INFORMÁTICOS**

**CIUDAD UNIVERSITARIA, ENERO DE 2023**

**UNIVERSIDAD DE EL SALVADOR**

RECTOR:

**MSC. ROGER ARMANDO ARIAS ALVARADO**

SECRETARIO GENERAL:

**ING. 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 SISTEMA INFORMÁTICOS**

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

Título:

**ANÁLISIS DE VENTAS DE LIBRERÍA PACÍFICO POR MEDIO DE TALEND Y  
POWERBI**

Presentado por:

**CHRISTIAN ALEXIS PÉREZ FUENTES**  
**LUIS FELIPE SERRANO AMAYA**  
**JESSICA MARCELA ZAVALAETA PADILLA**

Trabajo de graduación aprobado por:

Docente asesor:

**ING. FABRICIO QUINTANILLA**

**SAN SALVADOR, ENERO DE 2023**

# Contenido

Introducción .....	1
Capítulo I: Especificación de proyecto .....	2
Situación actual .....	2
I.    Antecedentes.....	2
II.   Descripción del problema.....	3
III.  Planteamiento del problema .....	3
Objetivos .....	5
Objetivo general.....	5
Objetivos específicos .....	5
Alcances .....	6
Justificación .....	7
Cronogramas de actividades .....	8
Presupuesto .....	9
Capitulo II: Análisis y diseño de la propuesta de solución.....	11
Metodología de trabajo .....	11
Descripción de la propuesta de solución.....	13
Descripción de la tecnología a utilizar.....	16
Diagrama arquitectónico de la solución.....	20
Descripción de cada componente de la solución .....	21

Capitulo III: Estrategia de implementación de propuesta de solución .....	23
Estrategia de implementación .....	23
Instalar el software Talend Open Studio.....	23
Clonar el repositorio del proyecto talend_pacifico .....	30
Instalar Power BI .....	30
Configuración en AWS.....	33
Configuración de Talend.....	52
Actualizar las credenciales de S3.....	56
Políticas de ejecución.....	61
Conexión para Redshift en Power Bi.....	63
Presupuesto de implementación.....	68
Análisis de resultados .....	72
Conclusiones y recomendaciones .....	81
Conclusiones .....	81
Recomendaciones .....	81
Bibliografía .....	83
Anexos .....	85
Anexo 1 .....	85
Anexo 2 .....	87
Anexo 3 .....	89

Anexo 4.....	91
Anexo 5.....	92
Anexo 6.....	93

## Introducción

Cada organización dispone de datos producidos en sus operaciones diarias la necesidad de aprovechar los datos y transformarlos en información útil para la organización es hoy en día esencial para cada organización, por lo tanto, es útil disponer de información para abordar sobre las decisiones en las unidades de organización.

Actualmente los sistemas de información están sometidos a una evolución, debido a que se busca mejorar el carácter de la información que estos producen, para obtener información de alto carácter decisivo para los niveles gerenciales, por ende, esta debe ser ordenada, consistente y fidedigna; todo este esquema conlleva a pensar: ¿qué tipo de tecnología podría brindarle a una institución esta infraestructura de información? Bajo este tipo de necesidades nace el Data Warehouse, el cual se fortalece como una colección de datos en la cual se encuentra integrada la información de la institución y que se usa como soporte para el proceso de toma de decisiones gerenciales. El Data Warehouse provee la esquematización de datos necesaria para la aplicación de herramientas de proceso analítico en línea. (OLAP), que explora los datos desde diversos puntos de vista definidos en dichos esquemas de datos.

El presente documento se desarrollará el planeamiento e implementación de una solución para la librería pacífico con el fin de contar con un Data Warehouse implementado con tecnologías en la nube, el cual se podrá consumir por programas de inteligencia empresarial para la creación de visualizaciones que facilitaran la toma de decisiones basados en la cantidad de ventas, volumen de ventas, ganancias, etc. Se describirá cada uno de los elementos en la solución, así como las tecnologías a utilizar y se describirá la estrategia de implementación de la misma.

## Capítulo I: Especificación de proyecto

### Situación actual

#### *I. Antecedentes*

En el transcurso de los años las empresas han entendido la importancia en el manejo de la información para el crecimiento de sus compañías; es por esto que nace un concepto llamado inteligencia de negocios o sus abreviaciones en ingles BI. La inteligencia de negocios es la conversión de los datos operativos de las empresas en información que da soporte a la tomade decisiones.

Tomar mejores decisiones de negocios rápidamente es la llave para el éxito en el mercado competitivo que se vive actualmente. Comprendiendo esto, las organizaciones están buscando mejorar sus sistemas de toma de decisiones ya que pueden ser rebasados por el volumen y la complejidad de los datos disponibles provenientes de sus sistemas de producción y transaccionales. El permitir que todos estos datos estén disponibles para toda la audiencia a lo largo de la empresa es hoy por hoy uno de los retos más significativos para los profesionales involucrados en las tecnologías de información.

La empresa Librería del Pacífico opera en El Salvador y cuenta con catorce puntos de venta los cuales se encuentran distribuidos a lo largo del territorio salvadoreño, estos están ubicados en cada departamento del país.

Librería del Pacifico se dedica a vender artículos escolares para educación primaria, incluyendo libros de texto, material para manualidades, utensilios escolares, etc. Librería del Pacifico acepta dos tipos de pago que son mediante efectivo y tarjeta de crédito/débito. Se emiten ticket, factura y crédito fiscal al realizar una compra.

## ***II. Descripción del problema***

Librerías del pacífico requiere una solución integral para poder en un repositorio de datos las ventas realizadas en sus diferentes sucursales con el fin de poder analizar los datos y tomar mejores decisiones estratégicas.

El fin de la solución a plantear es para poder extraer todos los datos de la base de datos de su sistema de ventas, transformarlo para el uso dentro de un datawarehouse y cargarlo a un base de datos en la nube, para poder consumirlos mediante una aplicación de BI y esta manera poder utilizar diferentes reportes basados en las necesidades de la empresa y las métricas a plantear.

## ***III. Planteamiento del problema***

Librerías del pacífico necesita la implementación de un Data Warehouse para el procesamiento, extracción y visualización de los datos de las ventas realizadas utilizando tecnologías en la nube para organizar una gran cantidad de datos que sean de ayuda para la toma de decisiones.



*Imagen 1 Planteamiento del problema*

**Fuente de datos:** Datos que ya tuvieron un primer procesado, esto quiere decir que fueron limpiados, es útil, clasificado y se encuentran consolidados en un sistema organizado en este caso la base de datos transaccional de librerías del pacífico.

**Proceso:** Creación de un Data Warehouse que será un depósito de datos centralizados con la información obtenida de la fuente de datos, toda esta información se procesará con métodos como el ETL (Extraer, Transformar, Cargar) y luego, el resultado almacenarlo en un Data Warehouse, un almacén electrónico donde se guardará una gran cantidad de información valiosa. En este lugar, los datos disponibles se almacenan de forma segura y resultan fáciles de recuperar y de analizar.

Los datos que se almacenan en un Data Warehouse son tanto históricos como actuales, lo que además permite un panorama aún mayor. Es importante saber que por definición sólo almacena datos que fueron modelados o estructurados.

**Salida de datos:** Datos organizados y transformados que facilitan la consulta de los mismos por parte de los usuarios. Con ello, es posible generar reportes históricos, análisis estadísticos y análisis cualitativos, predicciones, comparaciones, proyecciones y demás. Esto con la finalidad de poder tomar decisiones estratégicas cada vez más acertadas.

## **Objetivos**

### ***Objetivo general***

Desarrollar un Data Warehouse para el proceso de ventas para la empresa Librería del Pacífico que utiliza el software ERP Odoo.

### ***Objetivos específicos***

- Aplicar todos los conocimientos técnicos adquiridos en la especialización de ingeniería de datos.
- Utilizar los pasos necesarios para diseñar un modelo dimensional adecuado para el proceso de negocio.
- Utilizar la herramienta DataCleaner para la realización de un análisis de datos en las tablas involucradas en la aplicación Punto de ventas que está siendo utilizada en el software Odoo y así describir los datos más relevantes y de interés.
- Presentar los datos analizados mediante la herramienta Power BI para que el usuario final pueda visualizar los datos y comprenderlos mejor

## Alcances

- La creación de procesos ETL (son las siglas en inglés de Extraer, Transformar y Cargar) que permitan la extracción de datos contenidos en los sistemas transaccionales logrando cumplir con las 3 operaciones básicas del ETL las cuales son: extracción, transformación y carga.
- Realizar un análisis detallado del proceso de negocio que se quiere modelar.
- Documentar el modelo dimensional y proporcionar las especificaciones técnicas para su implementación.
- Utilizando los requisitos de información identificados, diseñar el modelo dimensional adecuado para el proceso de negocio. Incluyendo las tablas de hechos y dimensiones, y la definición de las relaciones entre ellas.
- Desarrollar una aplicación en PowerBi para la visualización Y comprensión de los datos.
- Hacer un análisis de datos de la cantidad de ventas por producto, construirlo y visualizarlo en PowerBI.
- Hacer un análisis de datos del volumen de ventas por producto, construirlo y visualizarlo en PowerBI.
- Hacer un análisis de datos del costo por producto vendido construirlo, y visualizarlo en PowerBI.
- Hacer un análisis de datos del margen de ventas, construirlo y visualizarlo en PowerBI.
- Hacer un análisis de datos del monto de descuentos, construirlo y visualizarlo en PowerBI.
- Hacer un análisis de datos de la cantidad de impuesto pagado por producto, construirlo y visualizarlo en PowerBI.

## **Justificación**

El creciente nivel de complejidad proveniente del desarrollo de los mercados en los que se desenvuelven las organizaciones y de los avances de las tecnologías de información, demanda a las empresas una mayor y mejor gestión de la información, la cual es considerada como la materia prima para la toma de decisiones, y que sirve de apoyo para enfrentar los retos presentes y anticiparse a los futuros, transformando esta información en conocimiento para la generación de soluciones útiles y oportunas para determinadas problemáticas.

El Business Intelligence ofrece una cantidad amplia de herramientas que facilitan a las compañías organizar y analizar los datos, permitiendo tomar sabias decisiones dentro de las que anteriormente se realizaban de manera manual y que requerían invertir más tiempo por parte de los gerentes del área. Esto lo hace atractivo y novedoso porque es una era tecnológica que debe ser aprovechada y explotada al máximo donde estar a la vanguardia e ir un paso delante de la competencia puede ser detonante para el éxito de una empresa.

La implementación de una Data Warehouse como parte de una solución de Inteligencia de Negocios que suplirá las necesidades de información de las distintas áreas de la compañía como: Disponibilidad, Análisis y Eficiencia a nivel estratégico que permita el apoyo a la toma de decisiones en beneficio de la empresa.

## Cronogramas de actividades

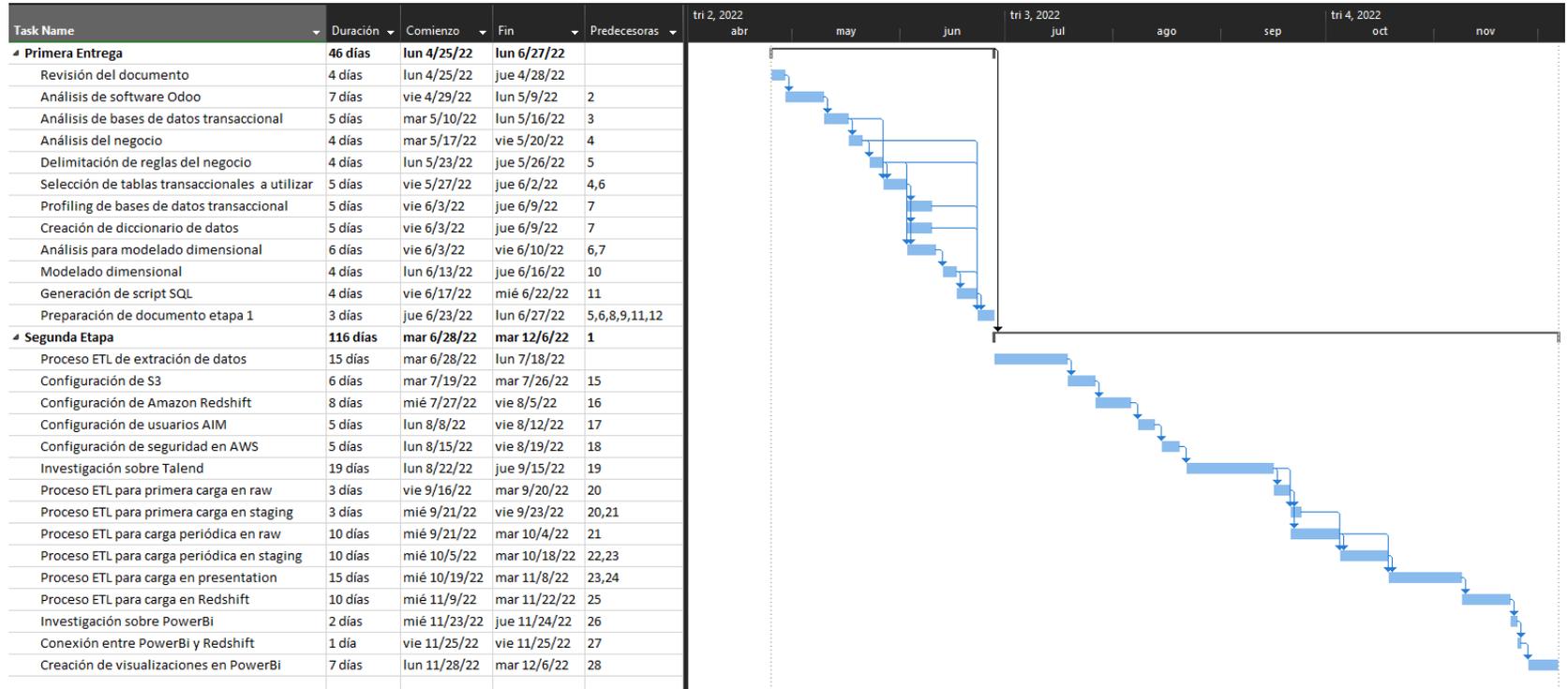


Imagen 2 Cronograma de actividades realizadas durante el desarrollo de la solución

## Presupuesto

### Costo de Amazon Redshift.

Durante la fase de desarrollo se utilizó Amazon Redshift para la elaboración y prueba de la base de datos para el Data Warehouse, por la cantidad de datos que se manejaban no se pudo utilizar la versión de prueba gratis y tuvimos que incurrir en gastos en la plataforma.

**Costo mensual de redshift = Cantidad de nodos X \$4.8 por hora X cantidad de horas activas**

**Costo de redshift = 1 X \$4.80/hora X 36 hora**

**Costo mensual redshift = \$172.80**

### Costo de Power Bi.

Para la creación de visualizaciones se necesita una licencia de Power Bi Pro, la cual cuesta \$9.99 mensuales por un usuario, esta se utilizó para las pruebas de conexión, así como para los primeros prototipos de las visualizaciones, esto con la idea de ir probando nuestros procesos ETL y asegurarnos que los datos llegaran de una manera correcta a Power Bi, esta licencia solo se pagó un mes que fue el tiempo de la construcción de las visualizaciones

### Costo por desarrollador.

En el proyecto se contará con 3 ingenieros que estarán a cargo de todas las actividades y tareas del mismo. Se ha acordado que se pagará \$10 por hora, por un tiempo máximo de 165 días.

**Costo por desarrollo de proyecto = Cantidad de ingenieros X \$10 por hora X 165 días**

**Costo por desarrollo de proyecto = 3 ingenieros X \$10.00 X 8 horas X 165 días**

**Costo por desarrollo de proyecto = \$39,600.00**

El costo total por el desarrollo de este proyecto es de:

**Costo por proyecto = Costo de Amazon Redshift + Costo de Power Bi + Costo por desarrollo**

**Costo por proyecto = \$172.80 + \$9.99 + \$39,600.00**

**Costo por proyecto = \$39,782.79**

## **Capítulo II: Análisis y diseño de la propuesta de solución**

### **Metodología de trabajo**

La solución está diseñada con base a la arquitectura de Kimball para Data Warehouses, propuesta por Ralph Kimball, esta arquitectura es conocida como "Arquitectura Kimball", la cual propone lineamientos lógicos y buenas prácticas para construir un Data Warehouse.

Este modelo divide los datos en la Fact Table y las Dimensiones. La Fact table se trata de datos numéricos medibles, y las Dimensiones complementan la información de la Fact Table, en su mayoría con datos no numéricos.

Este modelo propone un "Star Schema", o esquema de estrella, para el modelo del negocio. En este Star Schema, la Fact Table está relacionada con múltiples dimensiones. El esquema en estrella es el elemento fundamental del modelo de almacén de datos dimensional. La combinación de una tabla de hechos con varias tablas dimensionales a menudo se denomina esquema en estrella. El modelado dimensional de Kimball permite a los usuarios construir varios esquemas en estrella para satisfacer diversas necesidades de generación de informes. La ventaja del esquema en estrella es que las consultas de tablas dimensionales pequeñas se ejecutan instantáneamente.

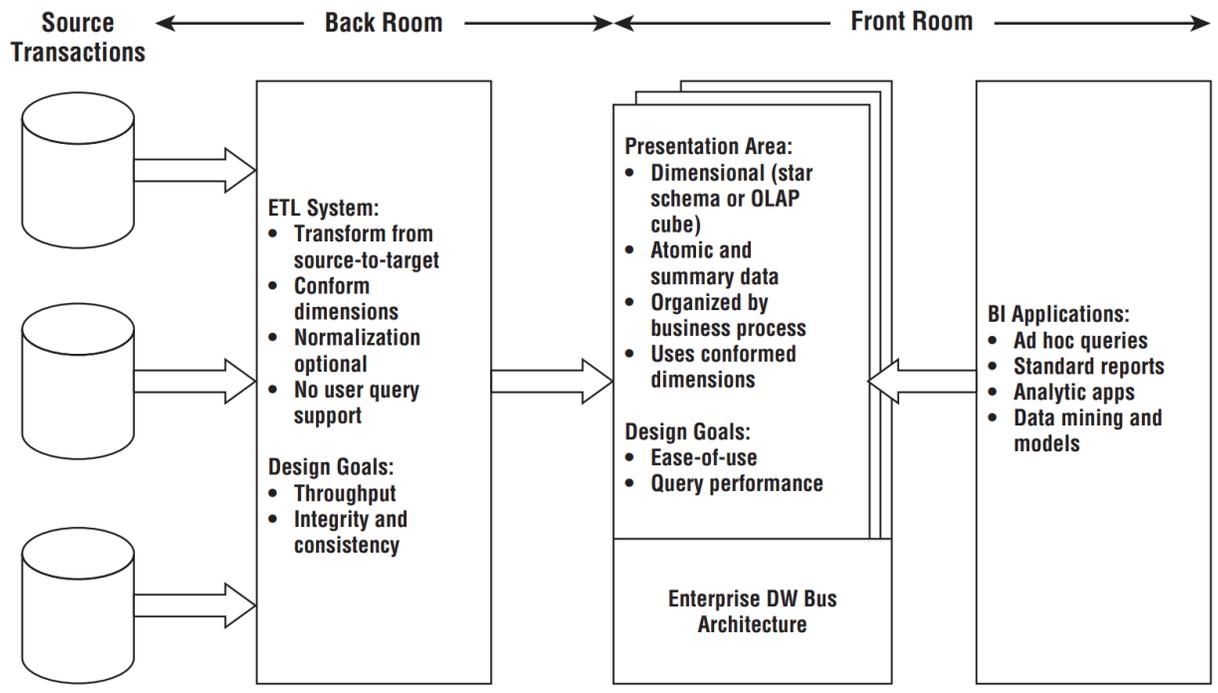


Imagen 3 Diagrama de arquitectura de Kimball Ross, M., & Kimball, R. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd ed.). Wiley.

## Descripción de la propuesta de solución

La solución propuesta busca resolver las preguntas del proceso de ventas para la empresa Librería del Pacifico. Por lo cual se utilizó la arquitectura de Kimball para diseño del DW y mediante el cual se busca responder las preguntas del negocio. Es necesario extraer los datos de la base de datos transaccional, luego de ello transformarlos y cargarlos en una capa de presentación la cual permitirá visualizar los datos de una mejor manera.

A continuación, se detallan cada una de las fases que son necesarias para la solución:

1. **Fuentes de datos:** como fuente de datos tenemos el sistema origen, el cual es el software RPG ODOO que es utilizado para realizar sus ventas.
2. **ETL (Extract - Transform - Load):** El sistema ETL consiste en un área de trabajo, estructuras de datos y un conjunto de procesos que se encargan de extraer, transformar y cargar los datos en Redshift.

Como primer paso tenemos la extracción que comprende la conexión a las fuentes de datos y transporte de los datos necesarios al sistema ETL para su manipulación.

El segundo paso es la transformación de los datos obtenidos en la fase de extracción, empezando por la limpieza de los datos, convertir a un formato específico, eliminar duplicación de datos entre las diferentes fuentes. Como último tenemos la carga de los datos los cuales son enviados a S3, donde podrán ser migrados fácilmente a Redshift y ser utilizados como capa de presentación por medio de la aplicación de BI.

3. **Capa de presentación:** almacena los datos de todas las dimensiones y la Fact Table. Aquí, los datos se encuentran listos para ser consumidos mediante aplicaciones BI.

4. **Aplicación de BI:** se conecta a la capa de presentación y permite visualizar los datos.

Para la solución propuesta se utiliza Power BI, donde se han creado informes para poder responder a cada una de las métricas.

### **Modelo Dimensional propuesto.**

1. **Selección de proceso de negocio.**

El proceso de negocio es el proceso de ventas para la empresa Librería del Pacífico que utiliza el software ERP Odoó.

2. **Definición de granularidad.**

Se requiere tener un nivel de detalle a nivel tiempo, cuanto se vendió una fecha específica siendo la unidad mínima el día.

3. **Identificación de dimensiones**

- Payment Method Dimension
- Product Dimension
- Store Dimension
- Date Dimension
- Attribute Dimension

4. **Identificación de métricas.**

- Cantidad de ventas por producto.
- Volumen de ventas por producto.
- Costo por producto vendido.
- Margen de ventas.
- Cantidad de descuento por venta.
- Cantidad de impuesto pagado por producto.
- Cantidad de descuento por venta.

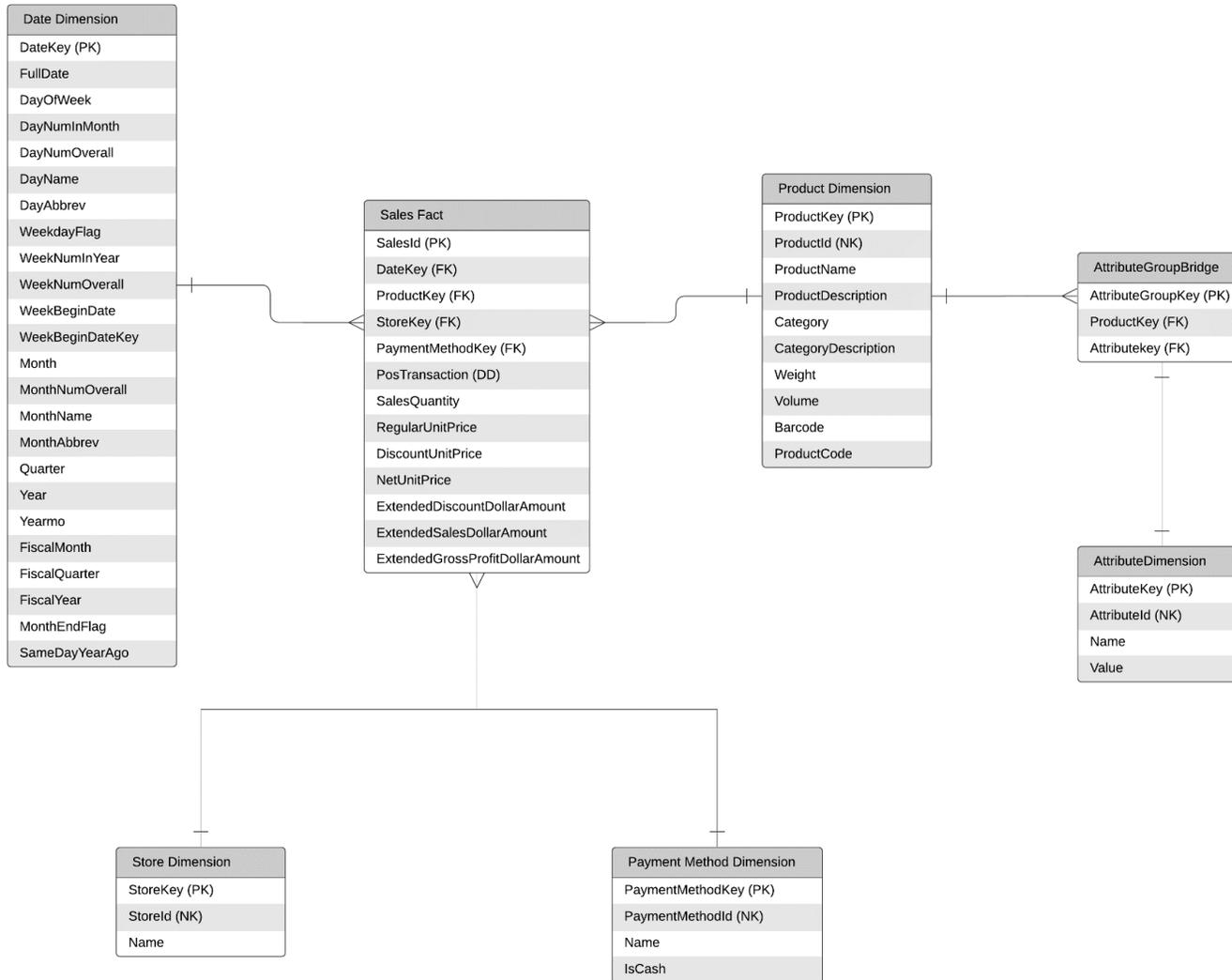


Imagen 4 Modelado dimensional

## **Descripción de la tecnología a utilizar**

### **Amazon S3**

Amazon S3 (Simple Storage Service) es el servicio de almacenamiento de objetos de Amazon Web Services (AWS) de tipo PaaS (Platform as a Service). Es la solución más común para almacenar datos en la nube de una forma segura, eficiente y escalable.

S3 es uno de los servicios más conocidos y usados de AWS desde sus inicios en el año 2006. Además, está totalmente integrado con la mayoría de los otros servicios ofrecidos y es muy sencillo de usar.

Los datos en S3 se almacenan como objetos dentro de los llamados Buckets S3. Un objeto es la unidad básica de almacenamiento en S3, consiste en un fichero con un identificador y metadatos asociados. Un Bucket en Amazon S3 no es más que un directorio lógico de alto nivel en el que se encuentran los objetos, cada uno de ellos identificado con una clave.

Al usar S3, es importante tener en cuenta las políticas de acceso, y evitar en lo posible dejar los buckets públicos. Es posible crear reglas y ACLs (Listas de Control de Acceso) para definir quién tiene acceso a los buckets y los objetos almacenados. También hacer uso de la transferencia de ficheros con el protocolo SSL.

### **Redshift**

Amazon Redshift es el servicio de Data Warehouse en la nube disponible en AWS que nos permite realizar consultas analíticas en SQL para big data de manera eficiente.

Es una base de datos MPP (Procesamiento paralelo masivo) que analiza datos usando múltiples nodos en paralelo. Cada uno de estos nodos se encarga de ejecutar un subconjunto de la consulta sobre los datos.

Nos proporciona las funcionalidades necesarias para implementar las consultas de alto rendimiento que alimenten procesos de BI, analítica en tiempo real y entrenamiento de modelos de machine learning con SQL.

Redshift es uno de los cloud data warehouses más populares. Aporta una gran potencia analítica de forma flexible y muy fácil de usar, integrado completamente con otros servicios del ecosistema de AWS como Amazon S3 o AWS Glue.

La configuración y la gestión de algunas actividades como la replicación, el backup y las actualizaciones son automáticas. Muchas tareas administrativas también se pueden automatizar a través de scripts. Redshift ofrece APIs con las que se puede interactuar a través de llamadas o de la consola de AWS, por ejemplo, para escalar el clúster de manera programática. También es compatible con Terraform como herramienta para automatizar los despliegues.

## **Talend**

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. De hecho, TOS (Talend Open Studio) te permite hacer grandes cosas de manera sencilla gracias a esta variedad de servicios.

## **PowerBi**

Power BI es el nombre colectivo para una variedad de aplicaciones y servicios basados en la nube que ayudan a las organizaciones a recopilar, administrar y analizar datos de una variedad de fuentes, a través de una interfaz fácil de usar. Power BI reúne los datos y los procesa, convirtiéndolos en información inteligible, a menudo utilizando gráficos y tablas visualmente convincentes y fáciles de procesar. Esto permite a los usuarios generar y compartir instantáneas claras y útiles de lo que está sucediendo en su negocio. Se conecta a una variedad de fuentes de datos, desde hojas de cálculo básicas de Excel hasta bases de datos, y aplicaciones tanto en la nube como en los servidores de nuestra empresa.

Power BI es un término general y puede referirse a una aplicación de escritorio de Windows llamada Power BI Desktop, un servicio en línea SaaS (Software como Service) llamado Power BI Service, o aplicaciones móviles de Power BI disponibles en teléfonos y tabletas con Windows, como, así como para dispositivos iOS y Android.

Se utiliza para ejecutar informes basados en los datos de una empresa. Power BI puede conectarse a una amplia gama de conjuntos de datos, y «ordena» la información que se alimenta para que pueda ser mejor digerida y entendida. Los informes y visuales generados a partir de estos datos se pueden compartir con otros usuarios.

## **Git**

Git es un sistema de control de versiones, un software que sirve básicamente para gestionar las versiones por las que va pasando el código de los proyectos.

Git es el más popular de los sistemas de control de versiones en la actualidad y una de las herramientas más indispensables para el desarrollo de proyectos.

Aunque un sistema de control de versiones sirva justamente para controlar los estados por los que ha pasado un código, la herramienta git facilita mucho el desarrollo de los proyectos en equipo, así como otras operaciones relacionadas con el flujo de trabajo de los desarrolladores.

## **AWS IAM**

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

IAM le ofrece las siguientes características:

- Permisos detallados
- Acceso seguro a los recursos de AWS para aplicaciones que se ejecutan en Amazon EC2
- Multi-Factor authentication (MFA)
- Información de identidad para realizar un control
- Integración con muchos servicios de AWS

IAM, al igual que muchos otros servicios de AWS, brinda consistencia final. IAM ofrece una alta disponibilidad, ya que replica datos entre varios servidores ubicados en centros de datos de Amazon de todo el mundo.

## Diagrama arquitectónico de la solución

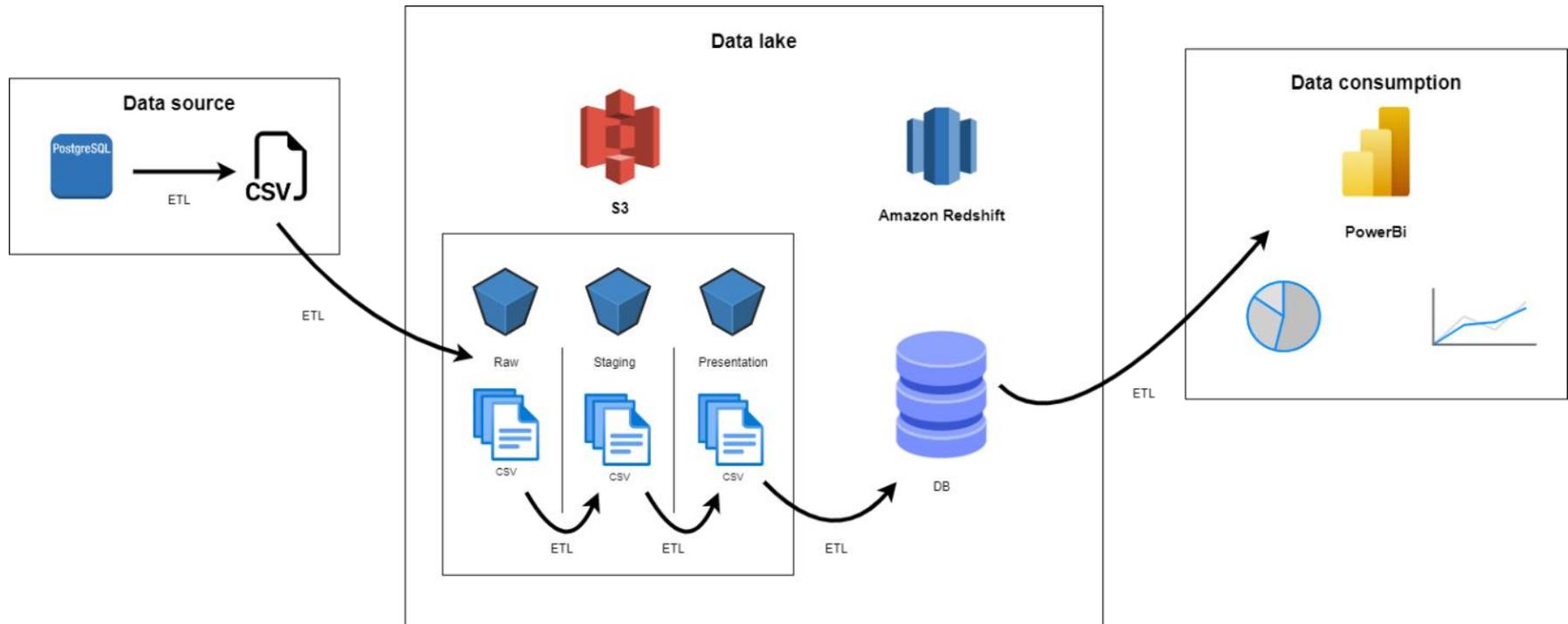


Imagen 5 Diagrama arquitectónico de la solución

## Descripción de cada componente de la solución

### Data Source

Nuestra base de datos relación se encuentra en PostgreSQL, en la versión 9.2, aquí al almacenamos todos los datos de nuestro sistema de ventas para la librería el pacífico, posteriormente, por medio de un proceso ETL, extraemos los datos de la base de datos y los almacenamos en archivos de tipo CSV, con sus respectivas columnas tal y como la base de datos la tiene.

### Data Lake

Aquí tenemos dos fases:

1. S3: Nos ayuda a almacenar los archivos que vamos transformando, para eso tenemos 3 diferentes zonas:
  - a. Raw: En esta zona almacenamos los archivos CSV que vienen de la base de datos tal y como los hemos extraídos utilizando el ETL, tenemos un archivo por cada tabla por cada vez que se ejecuta el ETL, los nombres de los archivos se componen del nombre de la tabla y de la fecha de creación.
  - b. Stage: En esta zona almacenamos los datos ya limpios que vienen de la zona de raw, esto se hace por medio de un ETL que se encarga de limpiar los datos, se tiene un archivo de tipo CSV por cada una de las tablas de la base de datos transaccional.
  - c. Presentation: En esta zona tenemos los datos ya transformados para crear una base de datos para el Data Warehouse, ya hemos utilizado varios ETL para limpiar y

transformar los datos, en este caso tenemos un archivo CSV por cada una de las tablas de hechos y atributos que hemos analizado para el Data Warehouse.

2. Redshift: Construimos una base de datos para nuestro Data Warehouse, aquí cargamos los datos que están en nuestra zona de Presentation en S3, de esta manera tenemos la base de datos lista para ser consultada por nuestra herramienta de visualización.

### **Data consumption**

En este componente hemos utilizado PowerBi como el programa que consume los datos de nuestro Data Warehouse, con la idea de crear visualizaciones que nos ayuden a resolver las métricas solicitadas por la librería el Pacífico, podemos crear todo tipo de gráficos que ayuden a la visualización rápida y fácil de los datos extraídos para esta solución.

## Capítulo III: Estrategia de implementación de propuesta de solución

### Estrategia de implementación

La estrategia de implementación será como un producto nuevo, debido a que la empresa Librería del Pacífico no posee ninguna solución de Data Warehouse implementada.

Los pasos a seguir para que implementar la solución de Data Warehouse son los siguientes:

#### *Instalar el software Talend Open Studio*

Previo a la instalación de Talend Open Studio, debemos asegurarnos de tener JDK al menos en su versión 8 instalado, de no ser así, podemos descargarlo desde el siguiente enlace:

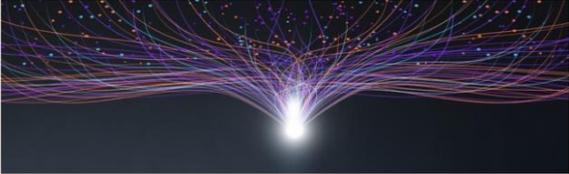
[Java Archive Downloads - Java SE 8 \(oracle.com\)](#)

URL para la descarga de Talend Open Studio:

[Talend Open Studio: Open-source ETL and Free Data Integration | Talend](#)

Al poner un correo, la información de descarga llegará en unos instantes:

If you are having trouble viewing this email in your browser, [click here](#) to view it online.



## Talend Open Studio for Data Integration

Thank you for your interest in Talend Open Studio for Data Integration! To get started, simply access the application for your OS ([Linux](#), [MAC](#), [Windows](#))

Need a little help getting off on the right foot? We have you covered:

- Read the [Getting Started Guide](#)
- Check out [Talend Community](#)

[ACCESS LINUX SOFTWARE/ZIP FILE INSTALL](#) →

[ACCESS MAC SOFTWARE](#) →

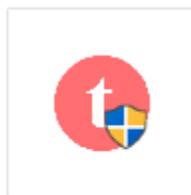
[ACCESS WINDOWS SOFTWARE](#) →

Talend.com [Privacy](#) [Email Preferences](#) [Unsubscribe](#)

© 2020 Talend All rights reserved.

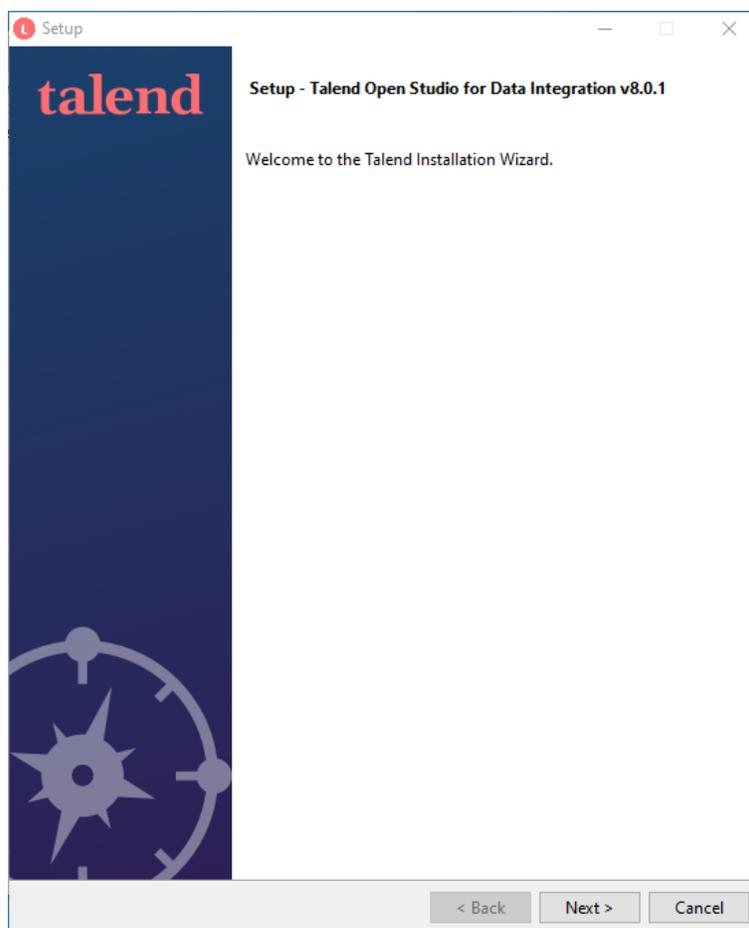
*Imagen 6 Correo de descarga Talend*

Nos descargará el archivo de instalación



TOS\_DI-20211109  
\_1610-V8.0.1-win  
dows-installer

*Imagen 7  
Archivo de instalación  
Talend*



*Imagen 8 Primer paso de instalación Talend*

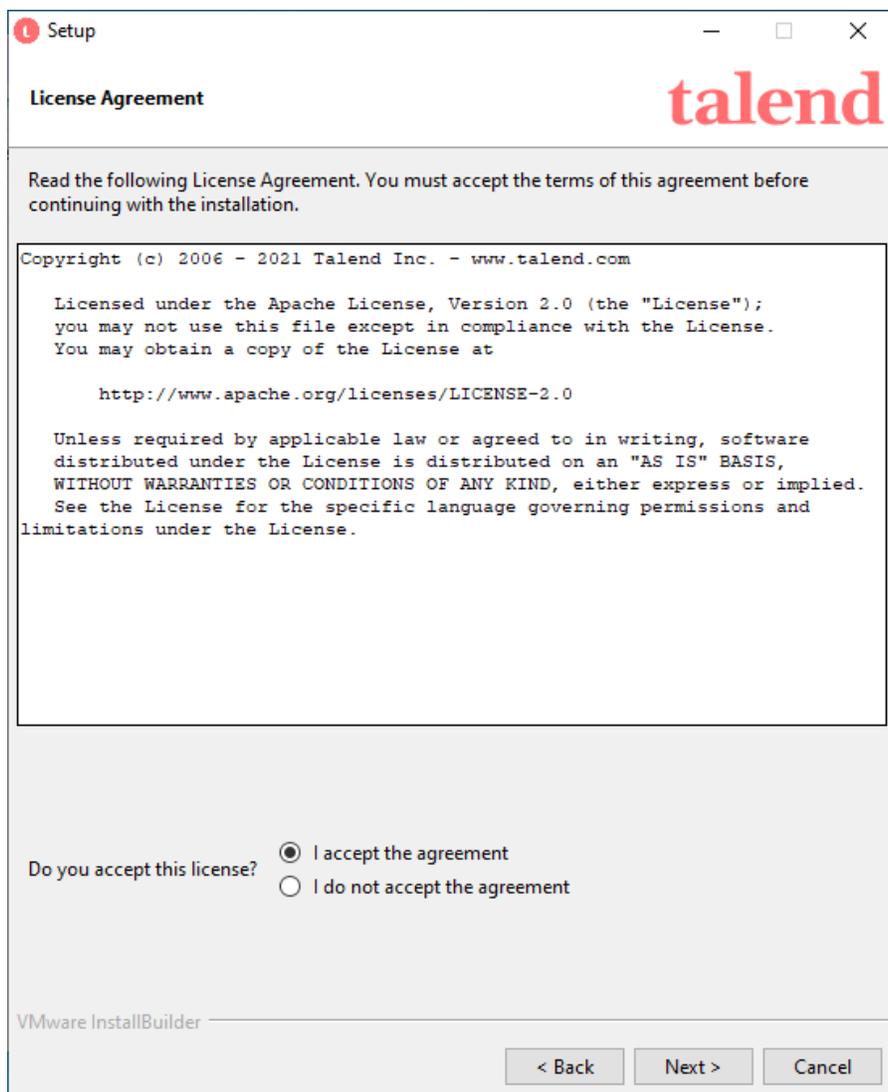


Imagen 9 Segundo paso instalación Talend

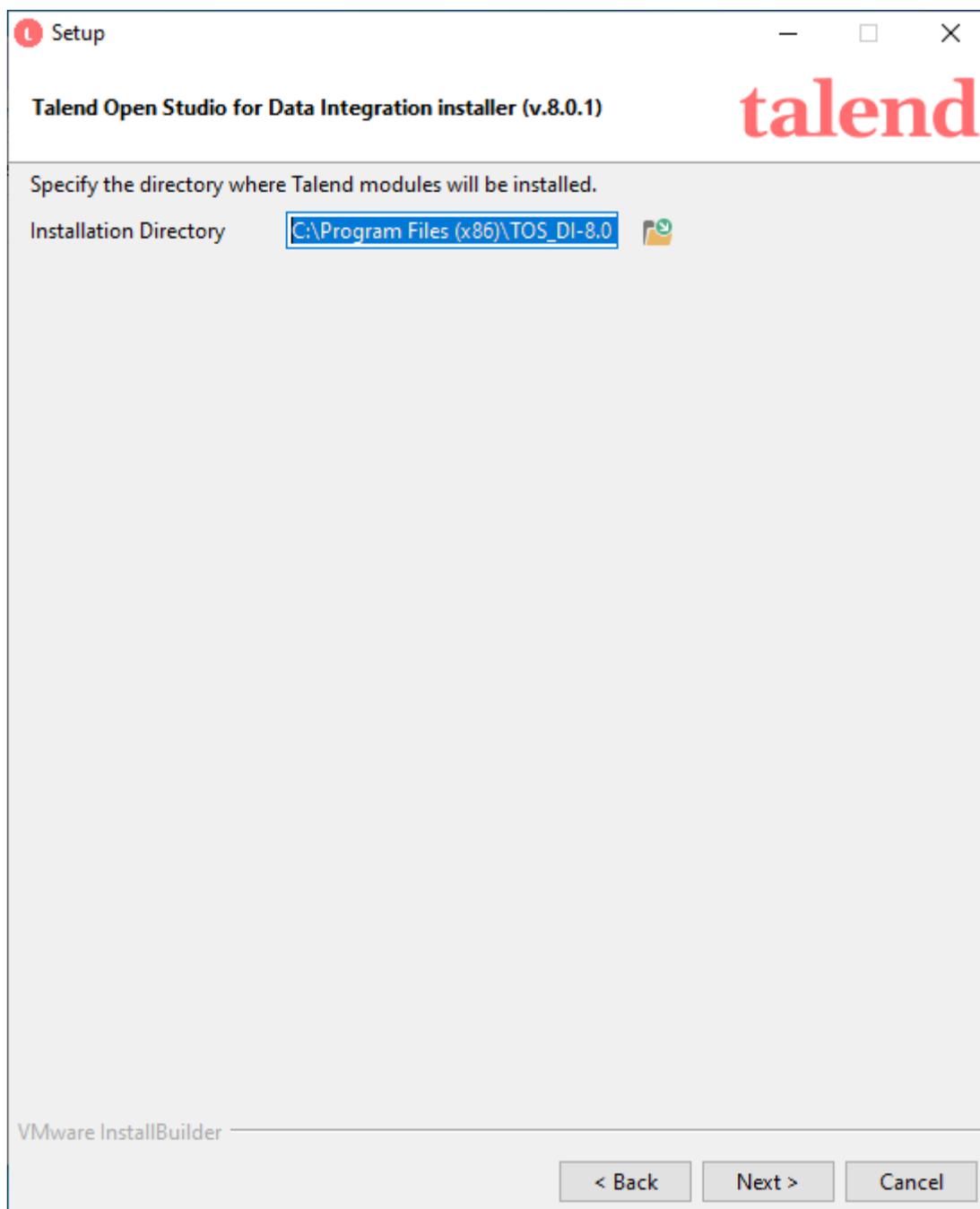
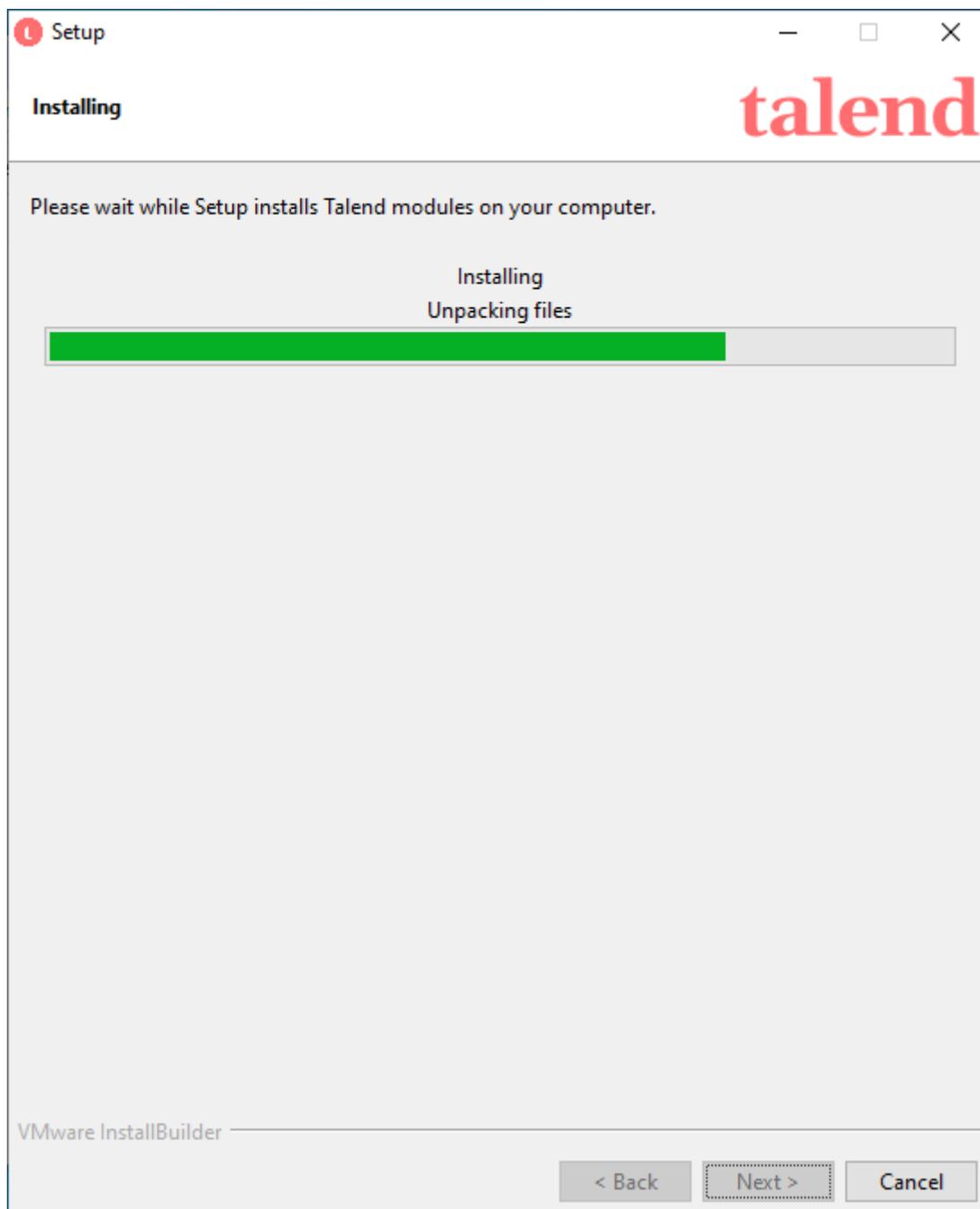
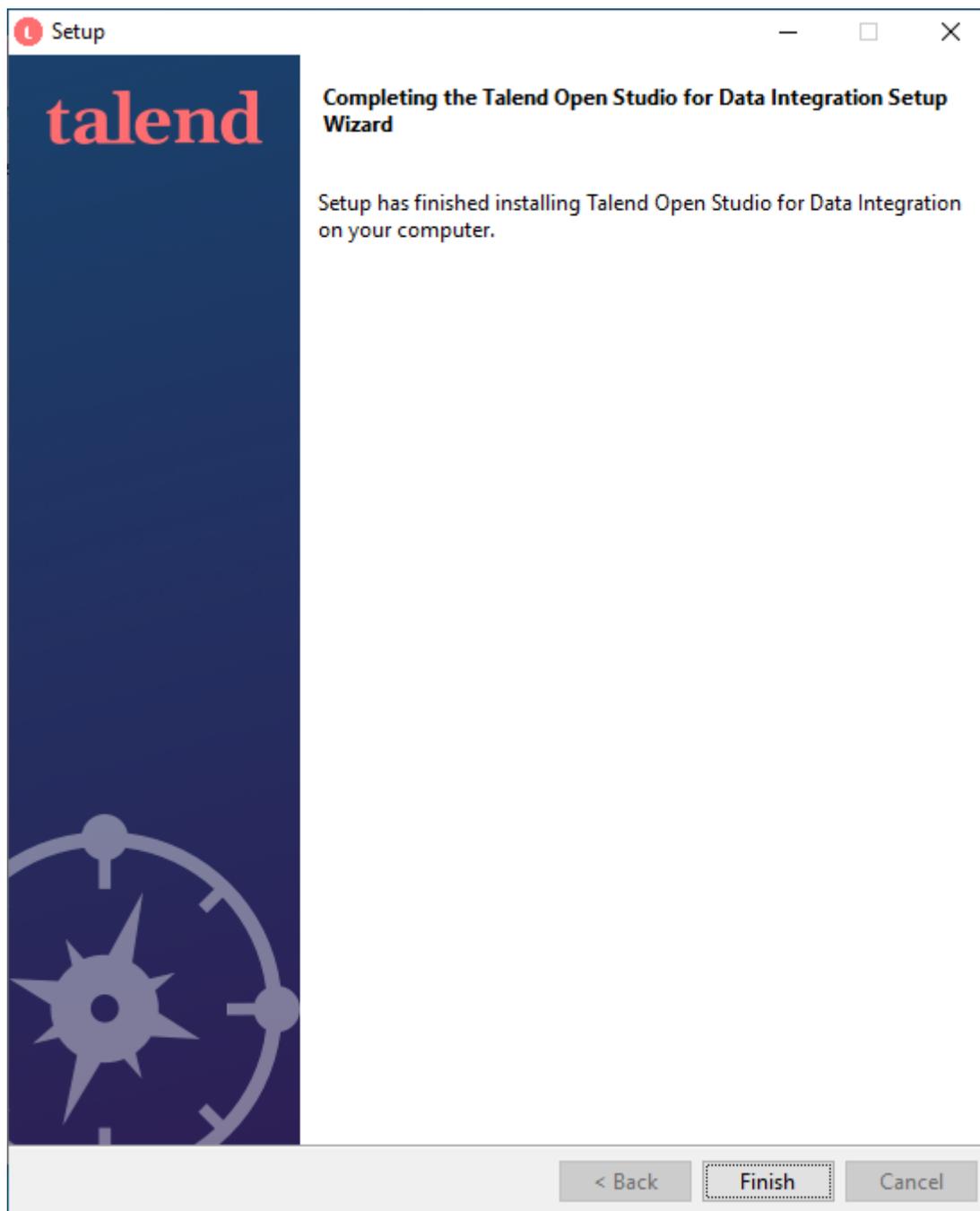


Imagen 10 Tercer paso instalación Talend

Esperamos que la instalación termine



*Imagen 11 Proceso instalación Talend*



*Imagen 12 Finalización de instalación Talend*

Una vez instalado podremos encontrar el icono en nuestro escritorio:



*Imagen 13 Acceso directo para Talend Open Studio*

### ***Clonar el repositorio del proyecto talend\_pacifico***

El repositorio completo del proyecto lo podemos encontrar en:

[https://github.com/lfelipeserrano/talend\\_pacifico](https://github.com/lfelipeserrano/talend_pacifico)

Para clonarlo debemos desde una consola o terminal del sistema ejecutar el siguiente comando estando en la ubicación dónde queramos almacenar el proyecto y todo su código:

**git clone [https://github.com/lfelipeserrano/talend\\_pacifico.git](https://github.com/lfelipeserrano/talend_pacifico.git)**

### ***Instalar Power BI***

Para el uso de PowerBi debemos de tener una cuenta en Microsoft, ya sea institucional o personal con la suscripción a Power Bi (el costo de esto ha sido detallado en los presupuestos de este documento)

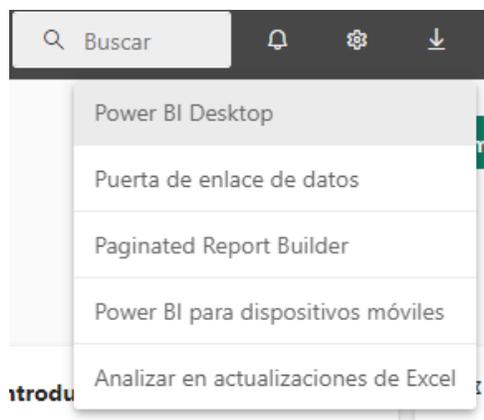
Para descargarlo debemos de ingresar al siguiente enlace:

[Power BI](#)

En la esquina superior derecha podremos encontrar el botón para descargar el software a nuestra computadora:

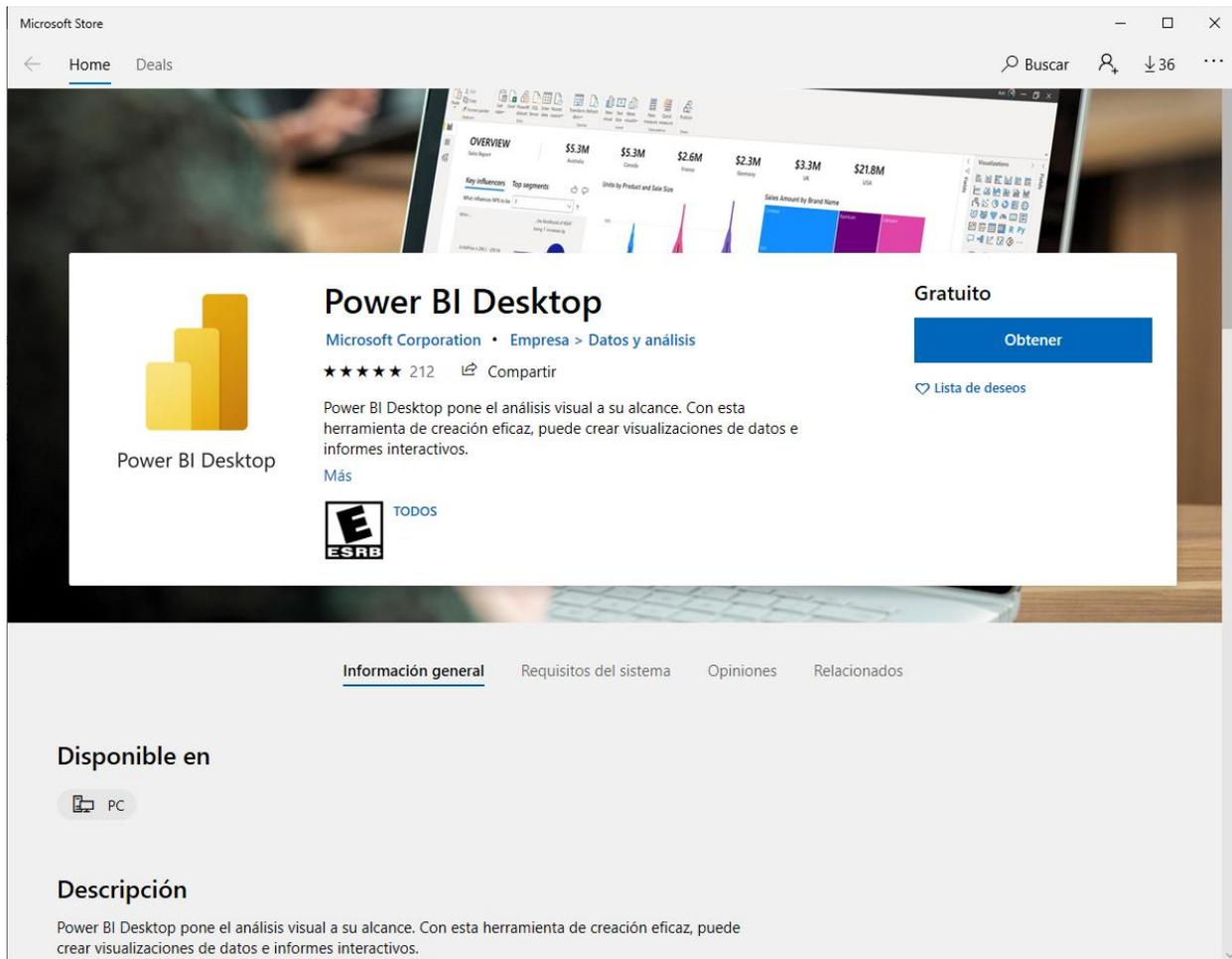


*Imagen 14 Acceso de descarga PowerBi*



*Imagen 15 Enlace de descarga PowerBi Desktop*

Esto nos llevará a la Microsoft Store, desde donde podremos obtener Power BI Desktop



Microsoft Store

Home Deals

Buscar 36

## Power BI Desktop

Microsoft Corporation • Empresa > Datos y análisis

★★★★★ 212 [Compartir](#)

Gratuito

[Obtener](#)

[Lista de deseos](#)

Power BI Desktop pone el análisis visual a su alcance. Con esta herramienta de creación eficaz, puede crear visualizaciones de datos e informes interactivos.

Más

 TODOS

[Información general](#) [Requisitos del sistema](#) [Opiniones](#) [Relacionados](#)

Disponible en

 PC

Descripción

Power BI Desktop pone el análisis visual a su alcance. Con esta herramienta de creación eficaz, puede crear visualizaciones de datos e informes interactivos.

Imagen 16 PowerBi en Microsoft Store

Al terminar la descarga ya estaremos listos para trabajar en PowerBI Desktop

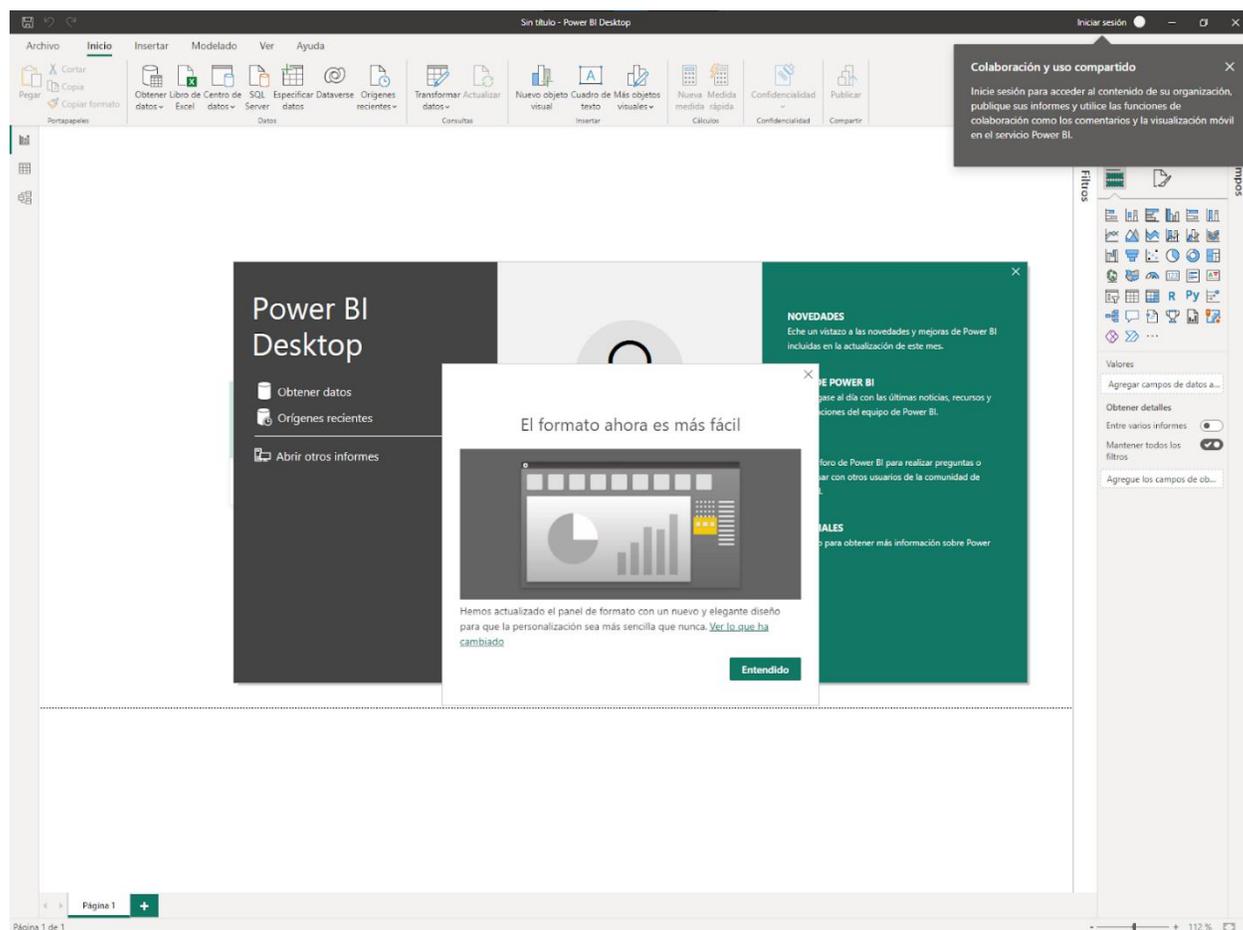


Imagen 17 PowerBi Desktop

## Configuración en AWS

- a. Crear un usuario IAM desde el usuario raíz, y utilizarlo para realizar los siguientes pasos.

Para la configuración de usuarios IAM debemos entrar al siguiente enlace:

[IAM Dashboard](#)

Al lado izquierdo podemos observar las opciones que tenemos, si presionamos la opción de Usuarios, podemos ver un listado de los IAM User y la opción de crear uno nuevo

---

## Identity and Access Management (IAM) ✕

---

 Search IAM

Dashboard

▼ Access management

User groups

**Users**

Roles

Policies

Identity providers

Account settings

▼ Access reports

Access analyzer

Archive rules

Analyzers

Settings

Credential report

Organization activity

Service control policies (SCPs)

---

*Imagen 18 Menú IAM*



*Imagen 19 Agregar nuevo usuario IAM*

Aquí podemos llenar y seguir cada uno de los pasos para crear un usuario

## Add user



## Set user details

You can add multiple users at once with the same access type and permissions. [Learn more](#)

User name\*

[+ Add another user](#)

## Select AWS access type

Select how these users will primarily access AWS. If you choose only programmatic access, it does NOT prevent users from accessing the console using an assumed role. Access keys and autogenerated passwords are provided in the last step. [Learn more](#)

- Select AWS credential type\*
- Access key - Programmatic access**  
Enables an **access key ID** and **secret access key** for the AWS API, CLI, SDK, and other development tools.
  - Password - AWS Management Console access**  
Enables a **password** that allows users to sign-in to the AWS Management Console.

Imagen 20 Paso 1 de creación usuario IAM

## Add user



▼ Set permissions

 Add user to group

 Copy permissions from existing user

 Attach existing policies directly

Add user to an existing group or create a new one. Using groups is a best-practice way to manage user's permissions by job functions. [Learn more](#)

## Add user to group

[Create group](#) [Refresh](#)

	Group ▼	Attached policies
<input type="checkbox"/>	Admin	AmazonRedshiftFullAccess and 3 more
<input type="checkbox"/>	Consultants	AmazonRedshiftReadOnlyAccess and 4 more

► Set permissions boundary

Imagen 21 Paso 2 de creación usuario IAM

## Add user



## Add tags (optional)

IAM tags are key-value pairs you can add to your user. Tags can include user information, such as an email address, or can be descriptive, such as a job title. You can use the tags to organize, track, or control access for this user. [Learn more](#)

Key	Value (optional)	Remove
<input type="text" value="Add new key"/>	<input type="text"/>	

You can add 50 more tags.

*Imagen 22 Paso 3 de creación usuario IAM*

## Add user



## Review

Review your choices. After you create the user, you can view and download the autogenerated password and access key.

## User details

<b>User name</b>	User1
<b>AWS access type</b>	Programmatic access - with an access key
<b>Permissions boundary</b>	Permissions boundary is not set

## Permissions summary

The user shown above will be added to the following groups.

Type	Name
Group	<a href="#">Admin</a>

## Tags

No tags were added.

*Imagen 23 Paso 4 de creación usuario IAM*

Una vez creada el usuario, podemos entrar a AWS utilizando este usuario y realizar los siguientes pasos

- b. Crear un bucket llamado libreriapacifico en amazon S3 y dentro del bucket crear las carpetas raw, stage y presentation

Para acceder a S3, podemos hacerlo por medio del siguiente enlace:

[Amazon S3](#)

Desde aquí podremos crear un nuevo bucket

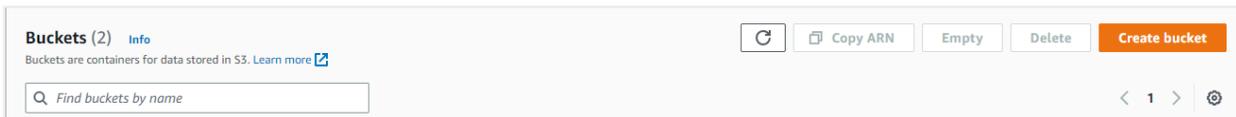


Imagen 24 Acceso a creación de bucket en S3

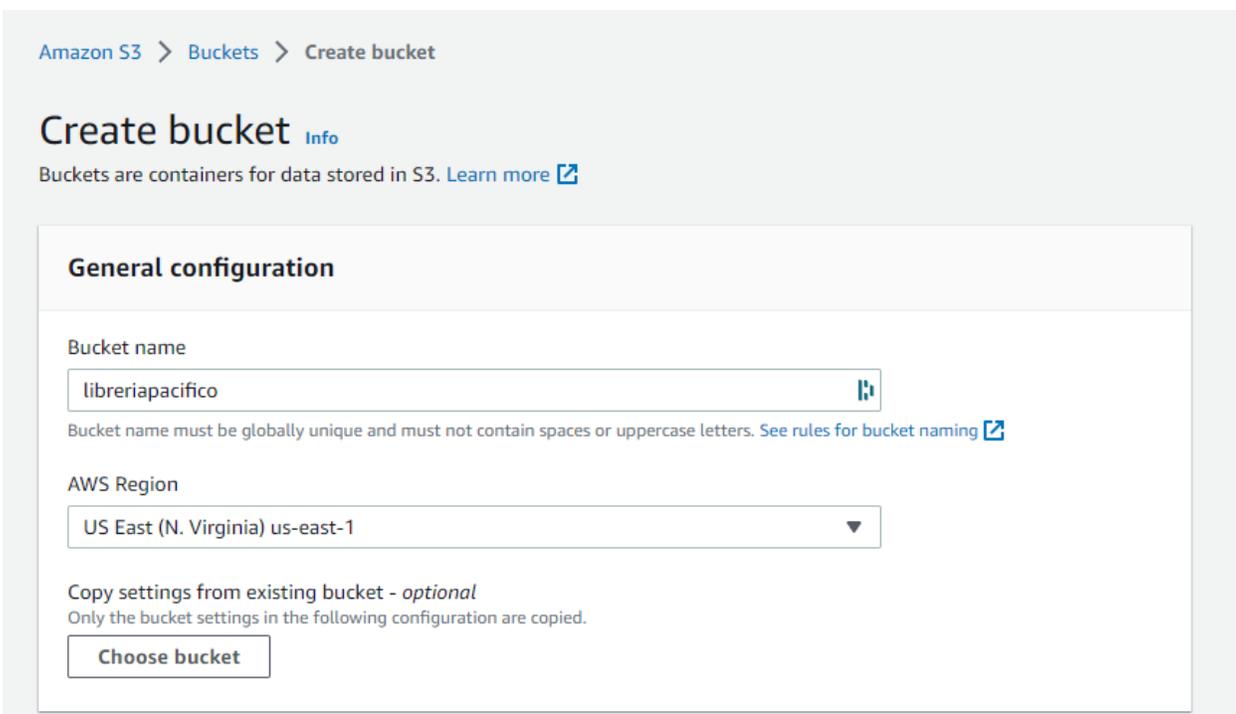


Imagen 25 Paso 1 de creación de bucket en S3

## Object Ownership [Info](#)

Control ownership of objects written to this bucket from other AWS accounts and the use of access control lists (ACLs). Object ownership determines who can specify access to objects.

- ACLs disabled (recommended)**  
All objects in this bucket are owned by this account. Access to this bucket and its objects is specified using only policies.

- ACLs enabled**  
Objects in this bucket can be owned by other AWS accounts. Access to this bucket and its objects can be specified using ACLs.

Object Ownership

Bucket owner enforced

## Block Public Access settings for this bucket

Public access is granted to buckets and objects through access control lists (ACLs), bucket policies, access point policies, or all. In order to ensure that public access to this bucket and its objects is blocked, turn on Block all public access. These settings apply only to this bucket and its access points. AWS recommends that you turn on Block all public access, but before applying any of these settings, ensure that your applications will work correctly without public access. If you require some level of public access to this bucket or objects within, you can customize the individual settings below to suit your specific storage use cases. [Learn more](#) [↗](#)

### **Block all public access**

Turning this setting on is the same as turning on all four settings below. Each of the following settings are independent of one another.

- Block public access to buckets and objects granted through *new* access control lists (ACLs)**  
S3 will block public access permissions applied to newly added buckets or objects, and prevent the creation of new public access ACLs for existing buckets and objects. This setting doesn't change any existing permissions that allow public access to S3 resources using ACLs.
- Block public access to buckets and objects granted through *any* access control lists (ACLs)**  
S3 will ignore all ACLs that grant public access to buckets and objects.
- Block public access to buckets and objects granted through *new* public bucket or access point policies**  
S3 will block new bucket and access point policies that grant public access to buckets and objects. This setting doesn't change any existing policies that allow public access to S3 resources.
- Block public and cross-account access to buckets and objects through *any* public bucket or access point policies**  
S3 will ignore public and cross-account access for buckets or access points with policies that grant public access to buckets and objects.

## Bucket Versioning

Versioning is a means of keeping multiple variants of an object in the same bucket. You can use versioning to preserve, retrieve, and restore every version of every object stored in your Amazon S3 bucket. With versioning, you can easily recover from both unintended user actions and application failures. [Learn more](#) [↗](#)

Bucket Versioning

- Disable  
 Enable

### Tags (0) - optional

You can use bucket tags to track storage costs and organize buckets. [Learn more](#)

No tags associated with this bucket.

[Add tag](#)

### Default encryption [Info](#)

Server-side encryption is automatically applied to new objects stored in this bucket.

Encryption key type [Info](#)

Amazon S3-managed keys (SSE-S3)

AWS Key Management Service key (SSE-KMS)

Bucket Key

When KMS encryption is used to encrypt new objects in this bucket, the bucket key reduces encryption costs by lowering calls to AWS KMS. [Learn more](#)

Disable

Enable

► **Advanced settings**

[i](#) After creating the bucket you can upload files and folders to the bucket, and configure additional bucket settings.

[Cancel](#) [Create bucket](#)

Imagen 27 Paso 3 de creación de bucket en S3

Una vez creado, dentro el mismo bucket podemos crear las carpetas necesarias, en este caso son 3: raw, stage y presentation

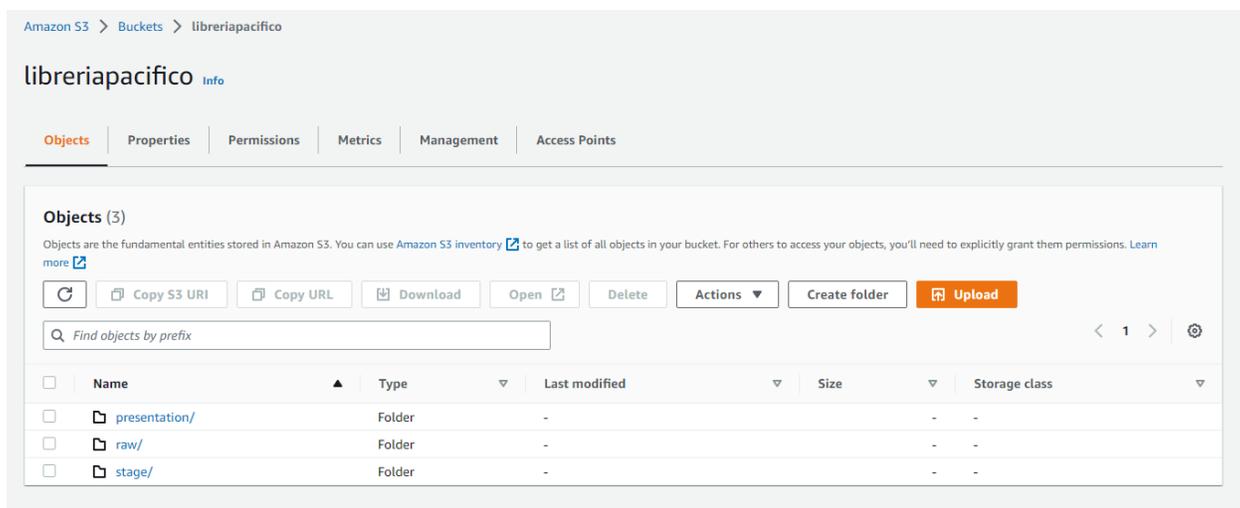


Imagen 28 Creación de carpetas en S3

Con esto tenemos listo nuestro bucket de S3 para comenzar a utilizarlo con Talend Open Studio

### c. Configuración de EC2

Para el correcto uso de Redshift, y que este pueda ser consultado por aplicaciones externas, para esto se necesita acceder al siguiente enlace:

[EC2 Dashboard](#)

A la izquierda debemos de buscar el apartado de Network & Security, y luego Security Groups

**EC2 Dashboard**

EC2 Global View

Events

Tags

Limits

**▼ Instances**

Instances

Instance Types

Launch Templates

Spot Requests

Savings Plans

Reserved Instances

Dedicated Hosts

Scheduled Instances

Capacity Reservations

**▼ Images**

AMIs

AMI Catalog

**▼ Elastic Block Store**

Volumes

Snapshots

Lifecycle Manager

**▼ Network & Security****Security Groups**

Elastic IPs

Placement Groups

Key Pairs

Network Interfaces

**▼ Load Balancing**

Load Balancers

Target Groups

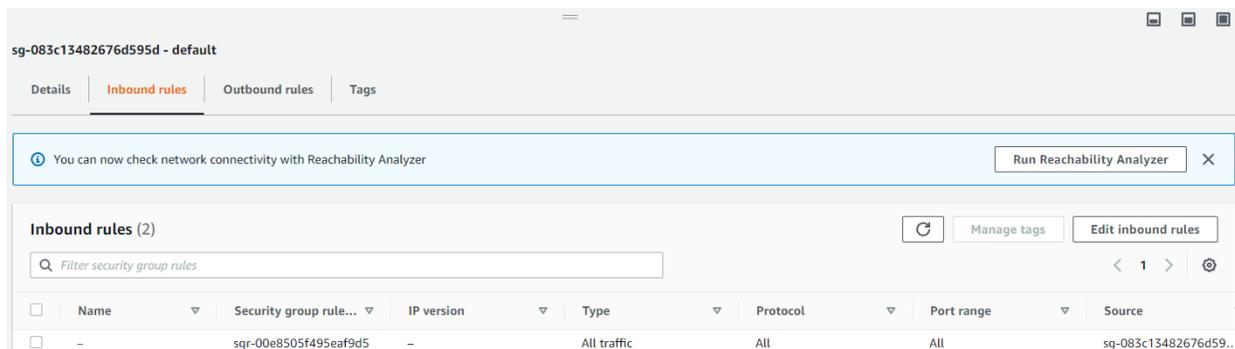
**▼ Auto Scaling**

Launch Configurations

Auto Scaling Groups

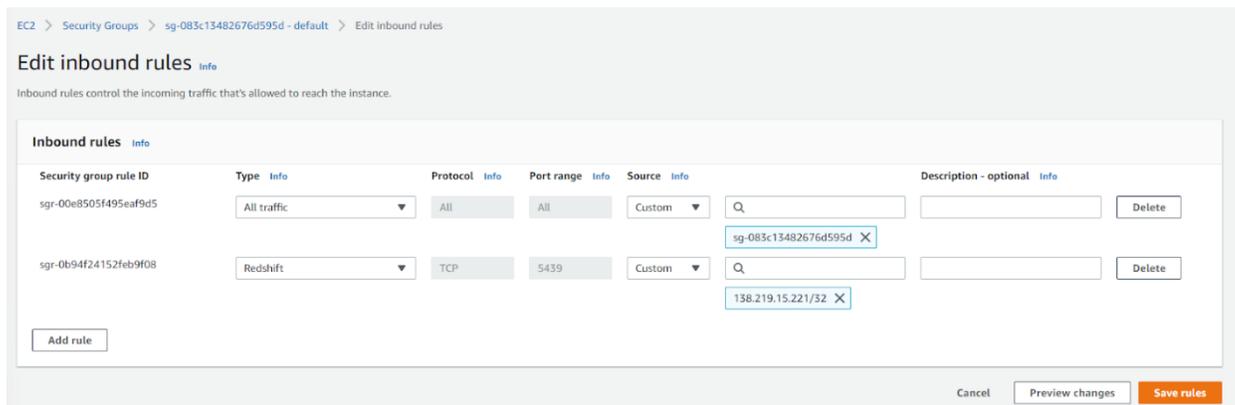
*Imagen 29 Menú de  
EC2*

Aquí podremos ver todos nuestros grupos de seguridad, y al seleccionar uno, en la parte de abajo, podemos ver nuestras reglas, seleccionaremos las inbound rules



*Imagen 30 Inbound Rules*

Presionamos el botón de Edit inbound rules y desde esa pantalla podemos agregar una nueva regla para Redshift



*Imagen 31 Edición de reglas*

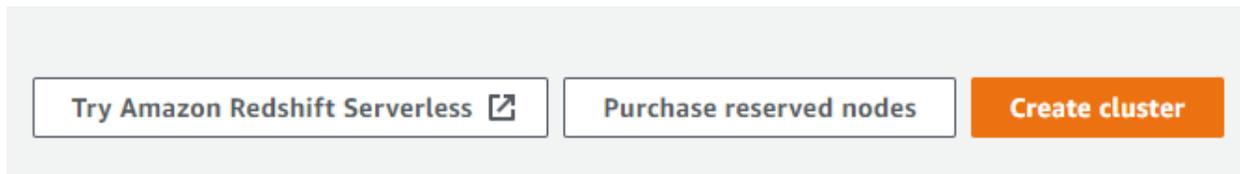
Guardamos y con esto nuestra instancia de Redshift se podrá comunicar con cualquier aplicación externa

#### d. Configuración de Redshift

Para crear nuestro clúster en Redshift, nos dirigimos al siguiente enlace:

[Amazon Redshift](#)

Aquí podemos crear nuestro nuevo clúster



*Imagen 32 Acceso a creación de cluster en Redshift*

Se nos pedirá un nombre, qué tipo de uso se le dará y un estimado del costo que tendrá, así como el nombre y contraseña del usuario administrador, entre otras cosas.

## Create cluster [Info](#)

### Cluster configuration

#### Cluster identifier

This is the unique key that identifies a cluster.

The identifier must be from 1-63 characters. Valid characters are a-z (lowercase only) and - (hyphen).

#### What are you planning to use this cluster for?

**Production**

Configure for fast and consistent performance at the best price.

**Free trial**

Configure for learning about Amazon Redshift. This configuration is free for a limited time if your organization has never created an Amazon Redshift cluster.

#### Choose the size of the cluster

**I'll choose**

Help me choose

#### Node type [Info](#)

Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.

#### Number of nodes

Enter the number of nodes that you need.

Range (2-32)

### Configuration summary [Info](#)

ra3.4xlarge | 2 nodes

**\$4,694.40/month**

**Estimated on-demand compute price**

Save more than 60% of your costs by purchasing reserved nodes.

[Learn more](#) [↗](#)

**256 TB**

**Max compressed storage**

RA3 stores data in Redshift managed storage. Each RA3 node gets up to 64 TB of compressed data capacity in managed storage to ensure optimal query performance.

**\$0.024/GB/month**

**Estimated storage price**

Pay only for the amount of data you store in managed storage when running an RA3 cluster.

Imagen 33 Paso 1 de creación de cluster en Redshift

## Database configurations

**Admin user name**  
Enter a login ID for the admin user of your DB instance.

  
The name must be 1-128 alphanumeric characters, and it can't be a [reserved word](#).

**Auto generate password**  
Amazon Redshift can generate a password for you, or you can specify your own password.

**Admin user password**  
Must be 8-64 characters long. Must contain at least one uppercase letter, one lowercase letter and one number. Can be any printable ASCII character except `"/", """, or "@"`.

  
 **Show password**

## Cluster permissions

**i** Create an IAM role as the default for this cluster that has the [AmazonRedshiftAllCommandsFullAccess](#) policy attached. This policy includes permissions to run SQL commands to COPY, UNLOAD, and query data with Amazon Redshift. The policy also grants permissions to run SELECT statements for related services, such as Amazon S3, Amazon CloudWatch logs, Amazon SageMaker, and AWS Glue.

## Associated IAM roles (0) [Info](#)

Create, associate, or remove an IAM role. You can associate up to 50 IAM roles. You can also choose an IAM role and set it as the default for this cluster.

<input type="checkbox"/>	<a href="#">IAM roles</a>	<input type="button" value="▼"/>	Status	<input type="button" value="▼"/>	Role type	<input type="button" value="▼"/>
<b>No resources</b> No associated IAM roles <input type="button" value="Associate IAM role"/>						

Imagen 34 Paso 2 de creación de cluster en Redshift

Una vez creado, y luego de una pequeña espera, podemos ver nuestro cluster en la lista de ellos

Cluster	Status	Cluster namespace	Availability Zone	Multi-AZ	Storage capacity us...	CPU utilization	Snapshots	Notificati...
redshift-cluster-1 dc2.large   1 node   160 GB	Available	3b831105-0641-4cad-...	us-east-1e	No	< 1%	13%	-	

Imagen 35 Listado de clusters en Redshift

Hacemos click en el nombre del clúster

**redshift-cluster-1**

**General information**

Cluster identifier redshift-cluster-1	Status Available	Node type dc2.large	Endpoint redshift-cluster-1.csmgpbaham.us-east-1.redshift.a...
Cluster namespace 3b831105-0641-4cad-a633-df5e960a8654	Date created January 07, 2023, 14:54 (UTC-06:00)	Number of nodes 1	JDBC URL jdbc:redshift://redshift-cluster-1.csmgpbaham.us-e...
Cluster configuration Production	Storage used 0.22% (0.35 of 160 GB used)		ODBC URL Driver={Amazon Redshift (x64)}; Server=redshift-clust...
	Multi-AZ No		

**Cluster performance** | Query monitoring | Schedules | Maintenance | Properties

**Recommendations (0)**  
To improve performance and decrease operating costs, the Amazon Redshift Advisor provides recommendations.

**Alarms (0)**  
CloudWatch alarms are triggered when a metric threshold is met.

**Events (1)**  
Amazon Redshift tracks events that occur on your cluster.

**Cluster metrics (11/11)**

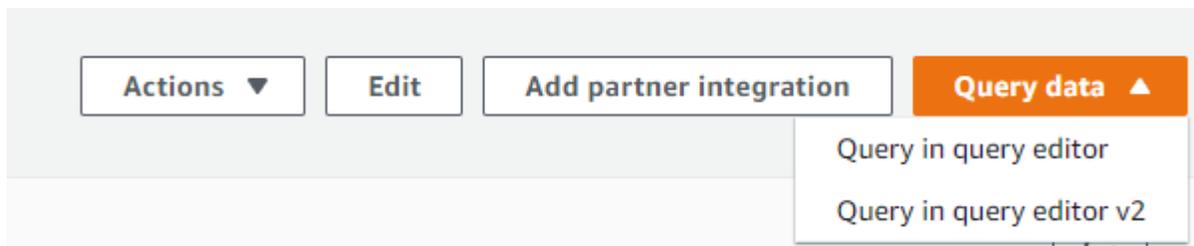
**CPU utilization**  
The percentage of CPU utilization.

**Percentage disk space used**  
The percent of disk space used.

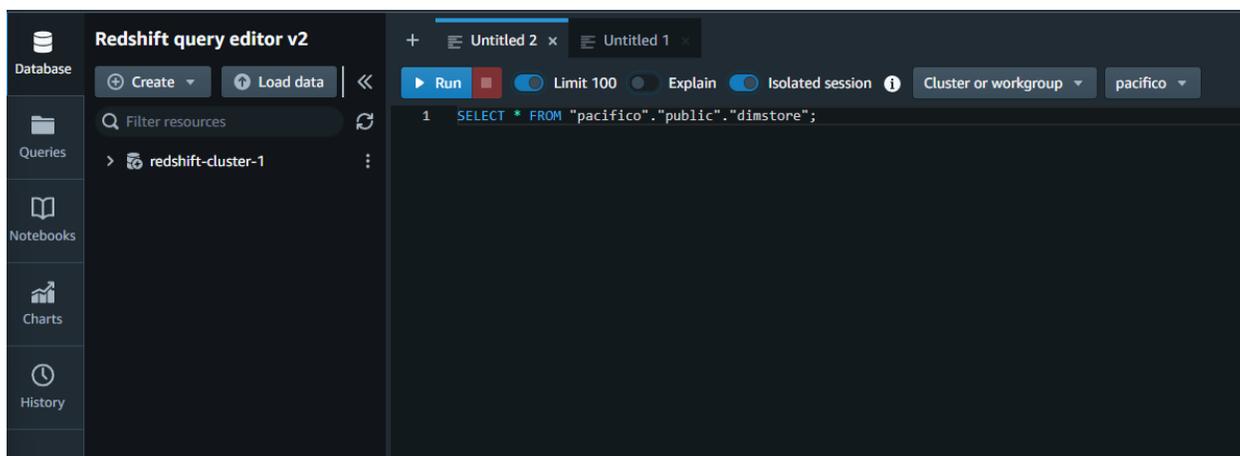
Imagen 36 Vista general de creación de cluster

Podemos ver información general del clúster, así como gráficos del uso que se le ha dado.

Procedemos al Query Editor, en este caso hemos decidido usar la versión 2:

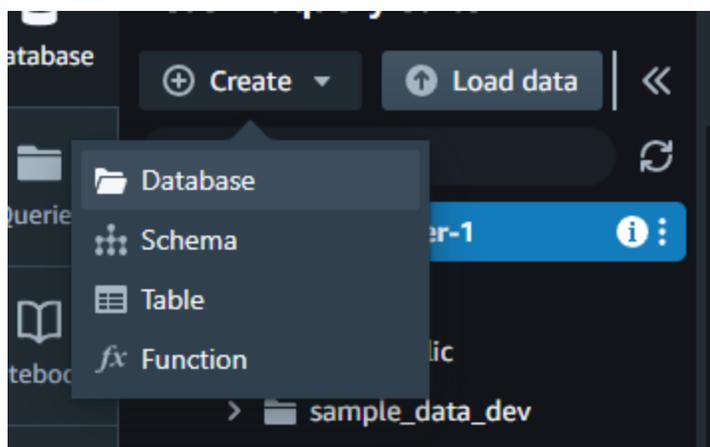


*Imagen 37 Acceso a Query Editor*



*Imagen 38 Query Editor*

Creamos una nueva base de datos:



*Imagen 39 Creación de bases de datos*

**Create database** [X]

Cluster or workgroup  
redshift-cluster-1 ▾

Database  
pacifico

The name consists of 1-127 UTF-8 characters (except control characters), and it can't be a reserved word.

Users and groups (optional)  
Database user ▾

Grant users and groups access to the database that was created. Use the SQL user interface to create new users.

Create using a datashare

Select a datashare ▾

To start querying data in a datashare, choose a datashare and create a database from it.

Cancel Create database

Imagen 40 Formulario creación de base de datos

Ya la podremos observar a la izquierda de la pantalla

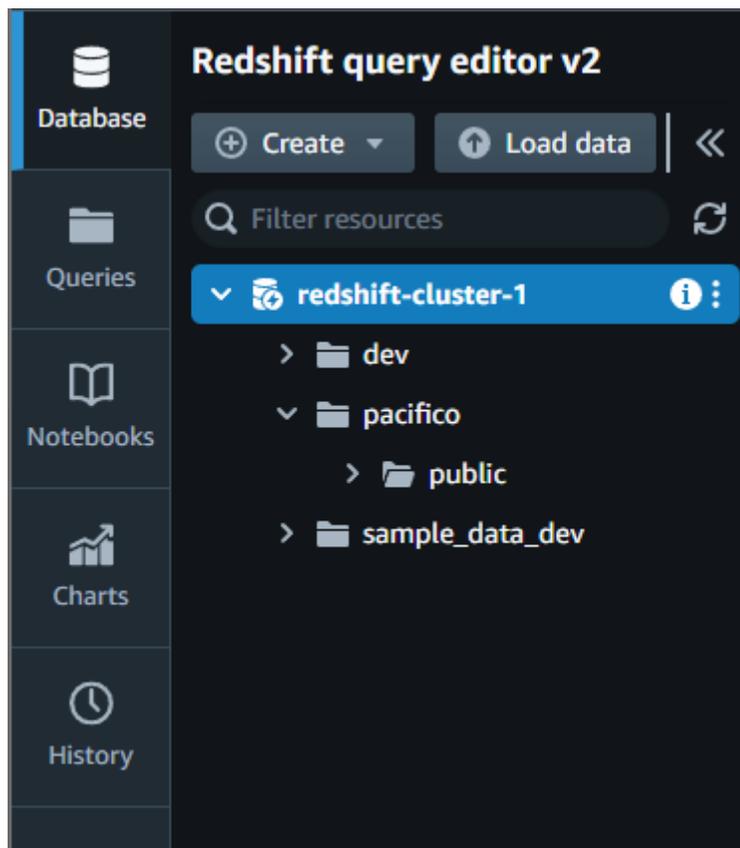


Imagen 41 Vista general de base de datos

Utilizamos el script para la creación de las tablas, este se puede encontrar en [Anexo 1.](#), debemos asegurarnos que se creen en la base de datos correcta:

```
▶ Run  Limit 100  Explain  Isolated session ⓘ redshift-cluster-1 pacifico
1 CREATE TABLE "public"."dimattribute"(attribute_key integer NOT NULL encode az64,
2   attribute_id integer encode az64,
3   name character varying(256) encode lzo,
4   value character varying(256) encode lzo,
5   CONSTRAINT dimattribute_pkey PRIMARY KEY(attribute_key));
```

Imagen 42 Ejemplo de script para creación de tablas

Al ejecutar cada uno de los scripts de creación de tablas, al lado izquierdo podemos ir verificando que las tablas hayan sido creadas de manera correcta:

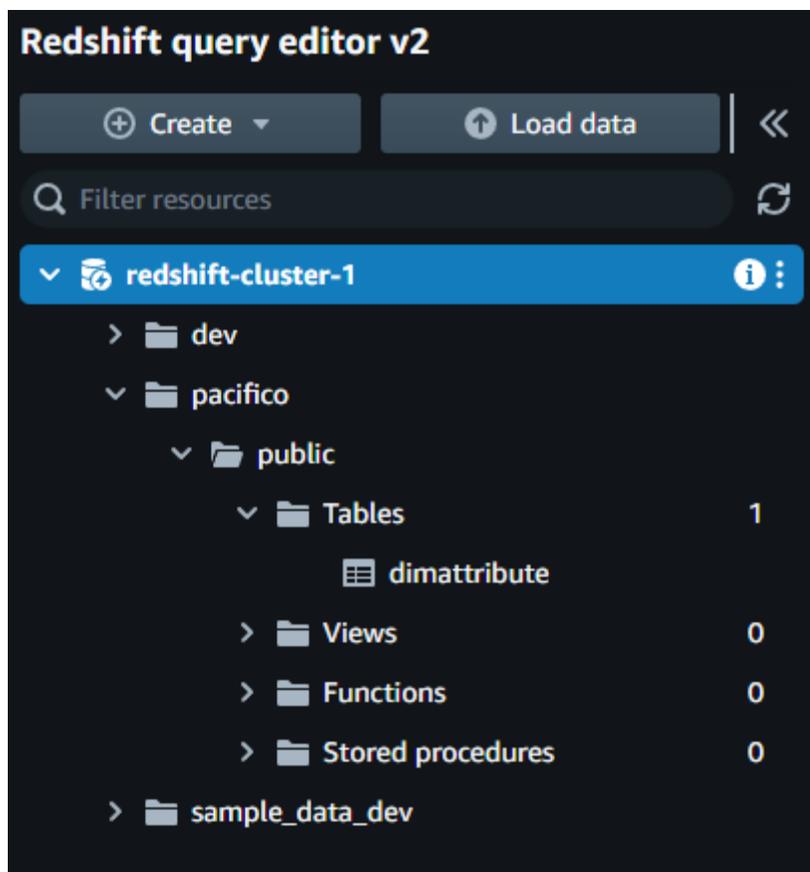


Imagen 43 Lista de tablas en base de datos

Luego de tener todas las tablas, Redshift está listo para ser utilizado por los procesos ETL y por PowerBI

### ***Configuración de Talend***

Crear una la conexión a la base de datos transaccional utilizando los siguientes parámetros:

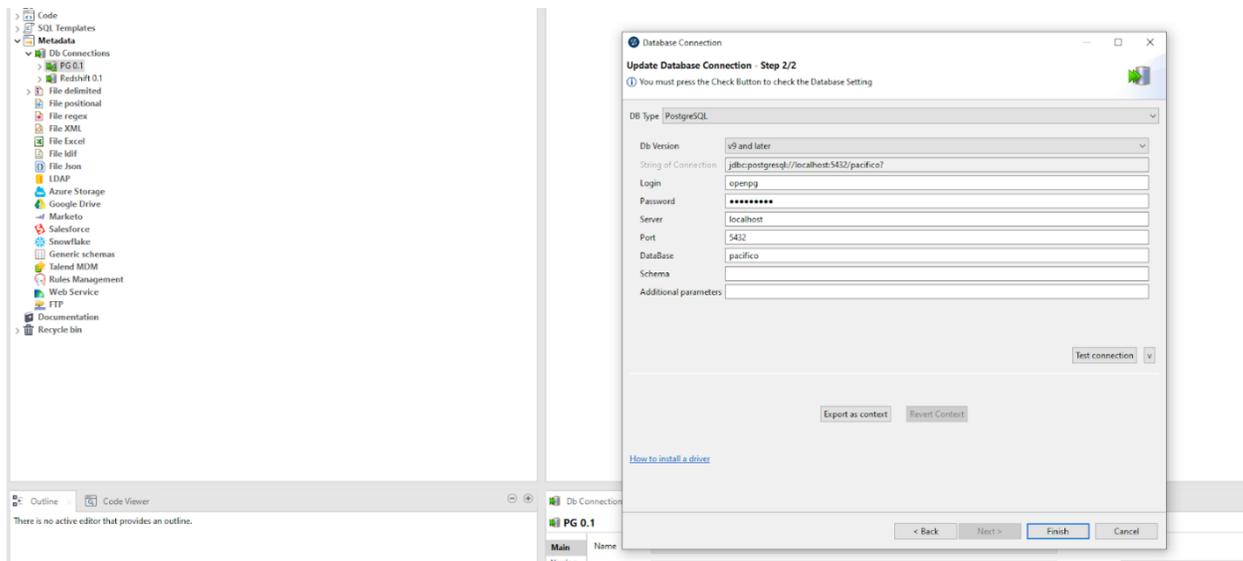


Imagen 44 Configuración de postgresql en Talend

Crear la conexión a la base de datos en redshift, utilizando como parámetros los que fueron creados durante la configuración del clúster en redshift.

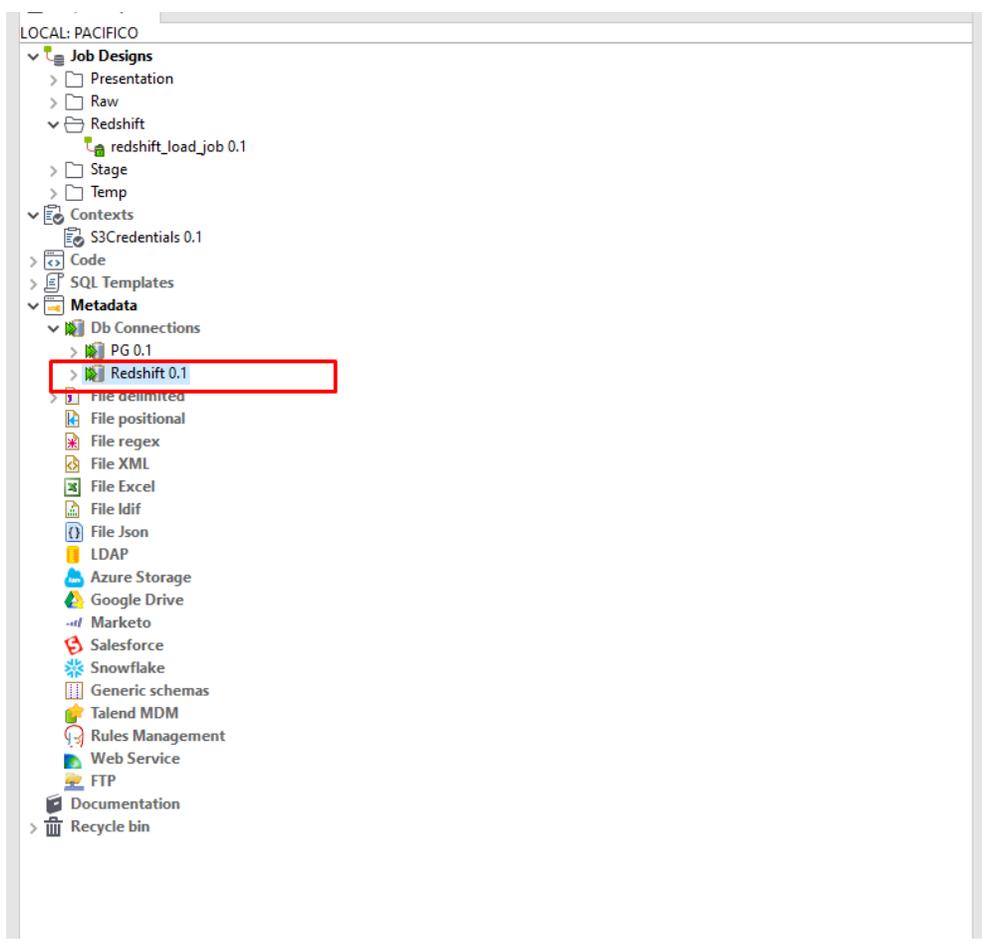


Imagen 45 Vista de conexiones a base de datos en Talend

Luego debemos llenar los campos con los datos previamente obtenidos desde Amazon Redshift y presionar dar click en finish.

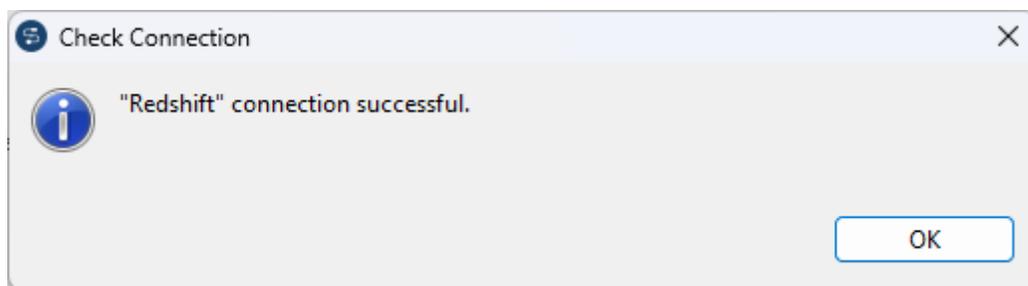
The screenshot shows a 'Database Connection' window titled 'Update Database Connection - Step 2/2'. The window contains the following fields and controls:

- DB Type:** A dropdown menu set to 'Redshift'.
- String of Connection:** A text field containing 'jdbc:redshift://redshift-cluster-1.csmpgpbaharn.us-east-1.redshift.amazonaws.com:5439/pacifico?'.
- Login:** A text field containing 'admin'.
- Password:** A text field with masked characters '••••••••'.
- Server:** A text field containing 'redshift-cluster-1.csmpgpbaharn.us-east-1.redshift.amazonaws.com'.
- Port:** A text field containing '5439'.
- DataBase:** A text field containing 'pacifico'.
- Schema:** A text field containing 'public'.
- Additional parameters:** An empty text field.
- Test connection:** A button with a dropdown arrow.
- Export as context:** A button.
- Revert Context:** A button.
- Navigation buttons:** '< Back', 'Next >', 'Finish', and 'Cancel'.

A link for 'How to install a driver' is visible at the bottom left of the dialog.

Imagen 46 Configuración de Redshift en Talend

Podemos hacer una prueba de la conexión para verificar que todos los datos estén correctos:



*Imagen 47 Prueba de conexión a Redshift desde Talend*

### ***Actualizar las credenciales de S3.***

Para ello primero se debe obtener el AccessKey y SecretAccessKey desde su usuario de IAM, siguiendo los pasos descritos a continuación:

1. Entrar con su usuario IAM a AWS.
2. Entrar a la consola IAM
3. Desde el menú de navegación hacer click en Users.

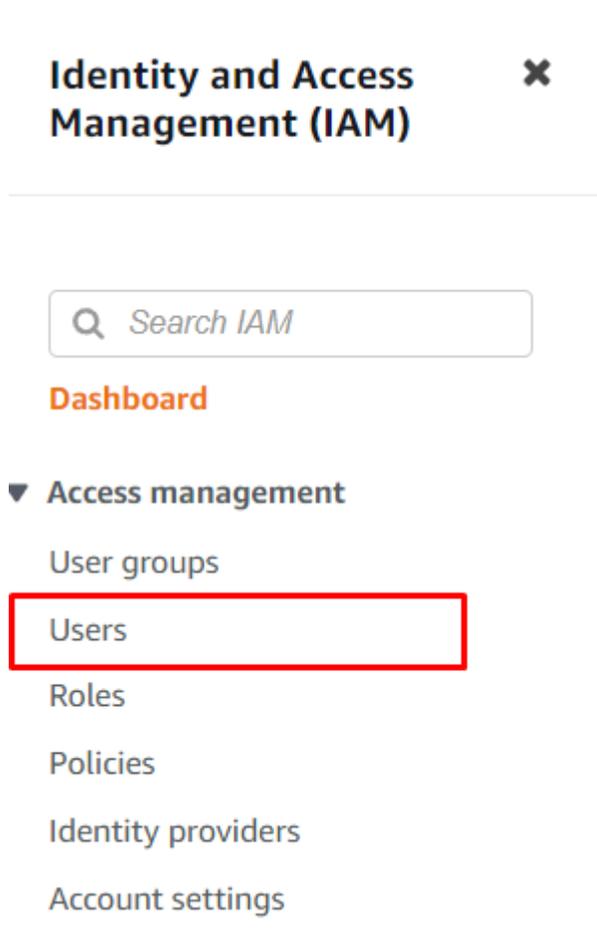


Imagen 48 Menú de IAM en AWS

#### 4. Hacer click en su usuario IAM

<input type="checkbox"/>	User name	Groups
<input type="checkbox"/>	para-volar	Consultants
<input type="checkbox"/>	pf15002	Admin

*Imagen 49 Lista de usuarios IAM*

#### 5. Dirigirse al apartado SecurityCredentials y hacer click en Create Access Key

Summary Delete user

User ARN: [arn:aws:iam::370254142437:user:para-volar](#)  
 Path: /  
 Creation time: 2022-11-09 08:29 CST

Permissions | Groups (1) | Tags | **Security credentials** | Access Advisor

Sign-in credentials

Summary • Console sign-in link: <https://370254142437.signin.aws.amazon.com/console>

Console password: Enabled (last signed in 58 days) | [Manage](#)

Signing certificates: None

Multi-factor authentication (MFA)

Use MFA to increase the security of your AWS environment. Signing in with MFA requires an authentication code from an MFA device. You can assign a maximum of 6 MFA devices. [Learn more](#)

[Manage](#) [Assign MFA device](#)

No MFA assigned.

Access keys

Use access keys to make programmatic calls to AWS from the AWS CLI, Tools for PowerShell, AWS SDKs, or direct AWS API calls. You can have a maximum of two access keys (active or inactive) at a time.

For your protection, you should never share your secret keys with anyone. As a best practice, we recommend frequent key rotation.

**If you lose or forget your secret key, you cannot retrieve it. Instead, create a new access key and make the old key inactive.** [Learn more](#)

[Create access key](#)

Access key ID	Created	Last used	Status
No results			

SSH keys for AWS CloudCommit

*Imagen 50 Apartado de credenciales de seguridad*

#### 6. Descargar en un el csv generado

Create access key ✕

**Warning**  
Never post your secret access key on public platforms, such as GitHub. This can compromise your account security.

**Success**  
This is the **only** time that the secret access keys can be viewed or downloaded. You cannot recover them later. However, you can create new access keys at any time.

[Download .csv file](#)

Access key ID	Secret access key
AKIAVMNG53PS3BUK3N2J	***** <a href="#">Show</a>

[Close](#)

Imagen 51 Clave de acceso para S3

Luego de obtener las credenciales de S3, en talend nos dirigimos a **Contexts->S3Credentials**

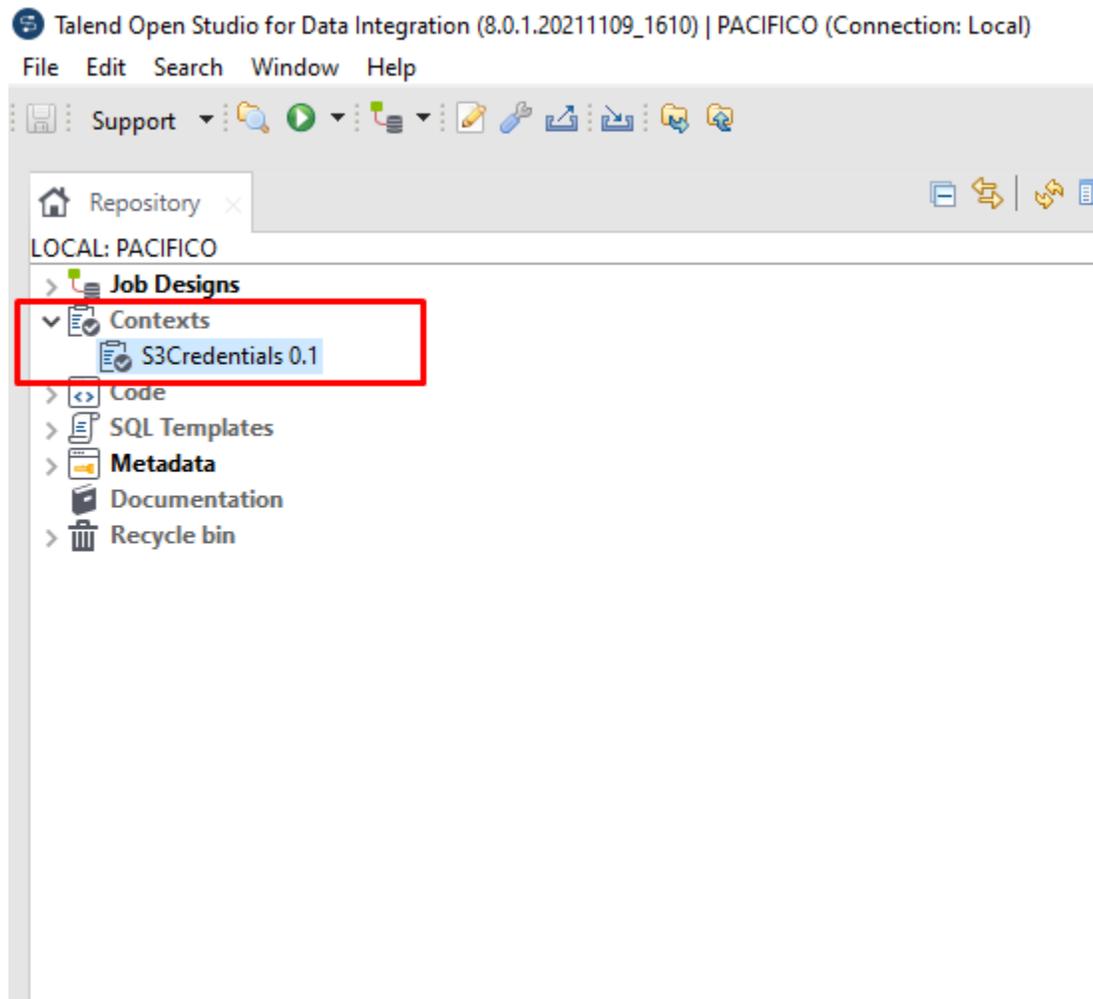
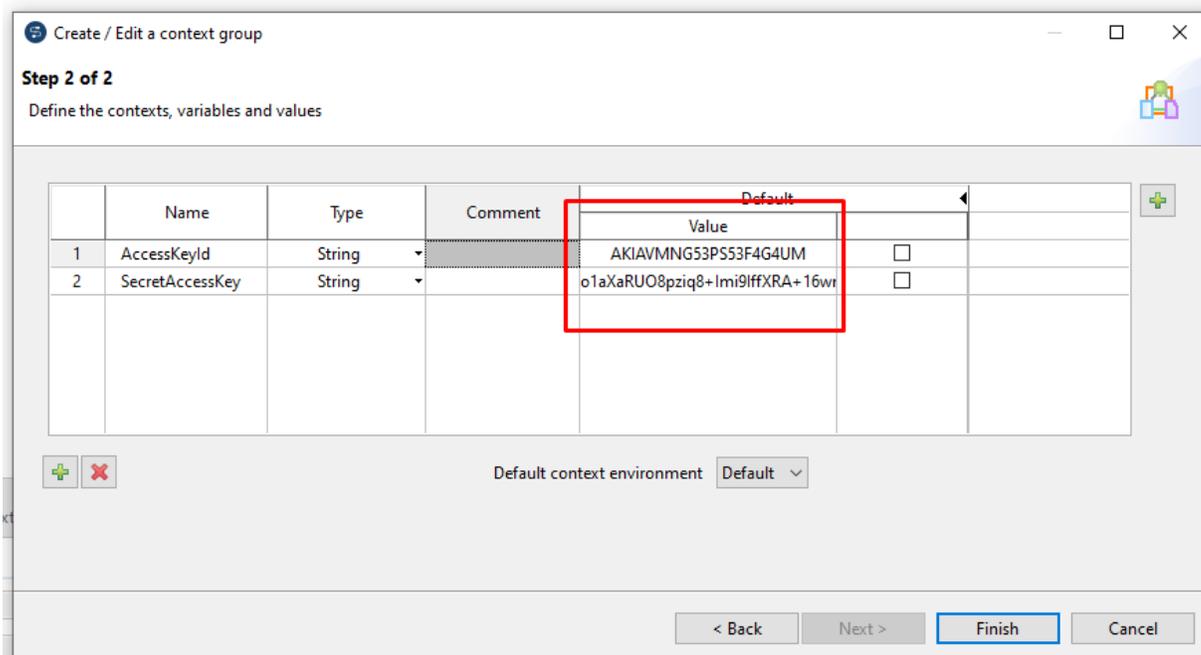


Imagen 52 Contexto en Talend para credenciales de S3

Luego dar doble click y editar los valores de las variables AccessKeyId y SecretAccessKey con los valores previamente obtenidos desde IAM.



*Imagen 53 Ejemplo de credenciales en ETL en Talend*

### ***Políticas de ejecución.***

Con todos los pasos anteriores los jobs deberán ser ejecutados en el siguiente orden:

1. Todos los jobs ubicados en la carpeta First Load de Raw (Este paso sólo debe realizarse la primera vez)
2. Todos los jobs ubicados en la carpeta Daily Jobs de Raw (Este paso se debe realizar las veces que sea necesario para actualizar la data)
3. Todos los jobs ubicados en la carpeta First Load de Stage (Este paso sólo debe realizarse la primera vez)

4. Todos los jobs ubicados en la carpeta Daily Jobs de Stage (Este paso se debe realizar las veces que sea necesario para actualizar la data)
5. Correr todos los jobs que se encuentran en las diferentes carpetas de Presentation.
6. Por último, correr el job que está ubicado en la carpeta Redshift.

Por último, existe un job creado específicamente para ser ejecutado periódicamente llamado Master Job, este se debe ejecutar bajo estas condiciones:

Diariamente y durante horas nocturnas, fuera de las horas laborales de todas las sucursales en el país, en el rango de las 10PM a 2AM, para evitar la carga en los servidores del sistema de ventas a la hora de extraer los datos de las ventas del día.

En caso de fallo de los procesos ETL o procesos de respaldo:

- En caso de fallo el proceso de respaldo de la base de datos transaccional, se deberá correr los procesos ETL para no perder datos importantes de un día en el datawarehouse y por lo tanto en las visualizaciones de las métricas a considerar.
- En caso de fallo en el proceso ETL, se deberá intentar correr de nuevo un máximo de tres veces, si en dado caso esto fallara, por cuestiones de tiempo se deberá suspender la ejecución de los mismos y reportarlo, para poder dar el soporte necesario y poder ejecutarlo al siguiente día.
- En caso de que ambos procesos fallen, se deberá reportar a los soportes correspondientes para poder ser ejecutado el siguiente día.

## Conexión para Redshift en Power Bi

Abrimos Power BI desktop y presionamos el botón **Get data**

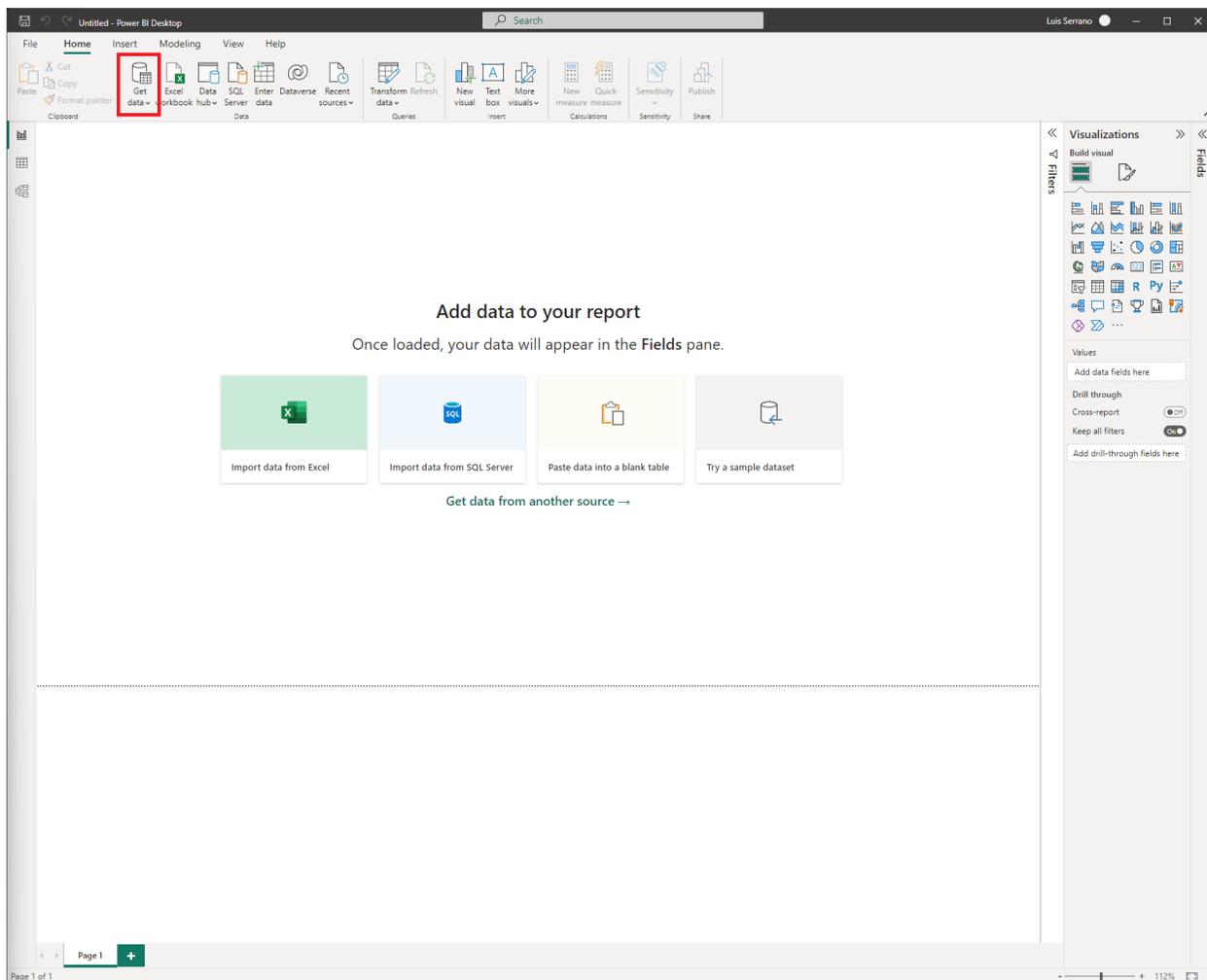


Imagen 54 Selección para obtener datos en PowerBi

Seleccionamos Database y luego Amazon Redshift

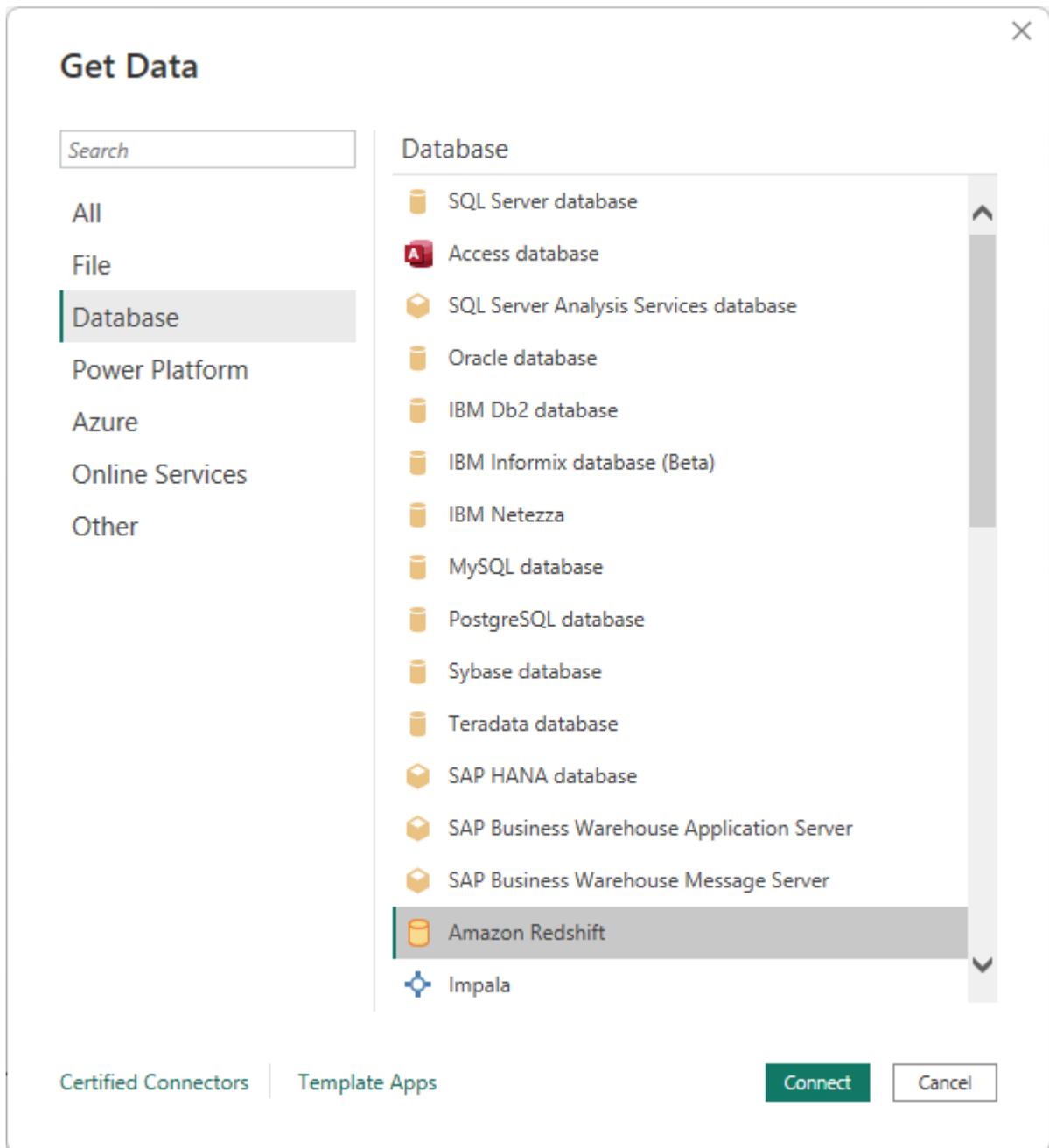
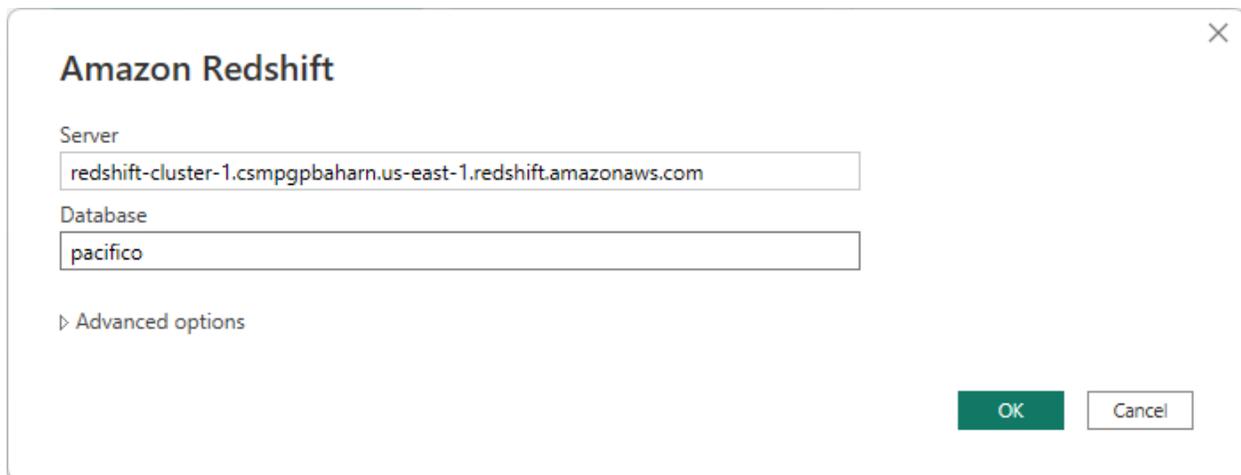


Imagen 55 Primer paso para la conexión de Redshift en PowerBi

Escribimos nuestra dirección de servidor y la base de datos a usar:



**Amazon Redshift**

Server  
redshift-cluster-1.csmppgbaharn.us-east-1.redshift.amazonaws.com

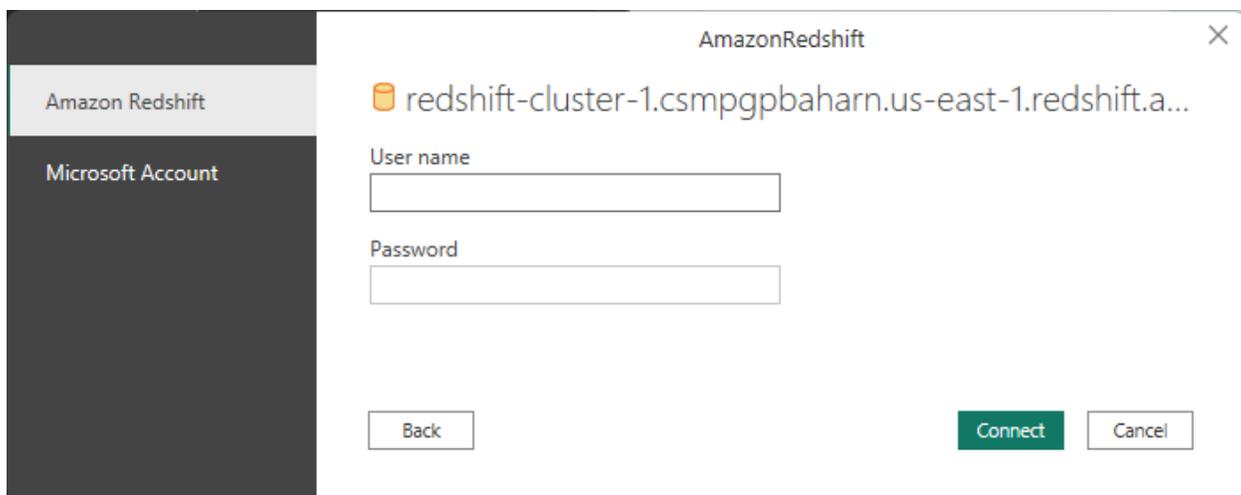
Database  
pacifico

▸ Advanced options

OK Cancel

*Imagen 56 Nombre del servidor y de base de datos en PowerBi*

Procedemos a poner nuestro usuario de Redshift antes establecido



**AmazonRedshift**

redshift-cluster-1.csmppgbaharn.us-east-1.redshift.a...

User name

Password

Back Connect Cancel

*Imagen 57 Nombre de usuario y clave para Redshift en PowerBi*

Podremos ver y seleccionar las tablas a utilizar en nuestro proyecto de PowerBi

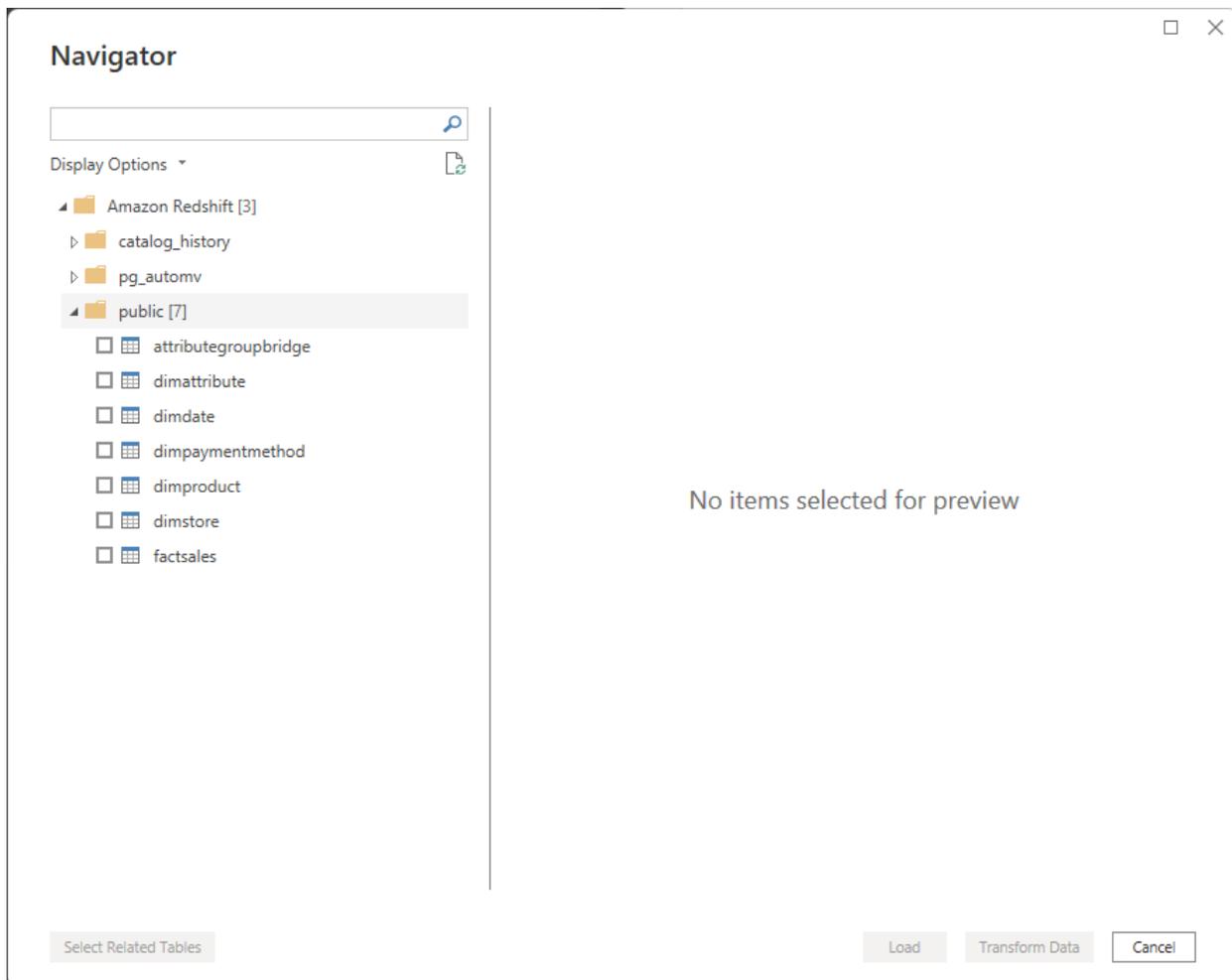
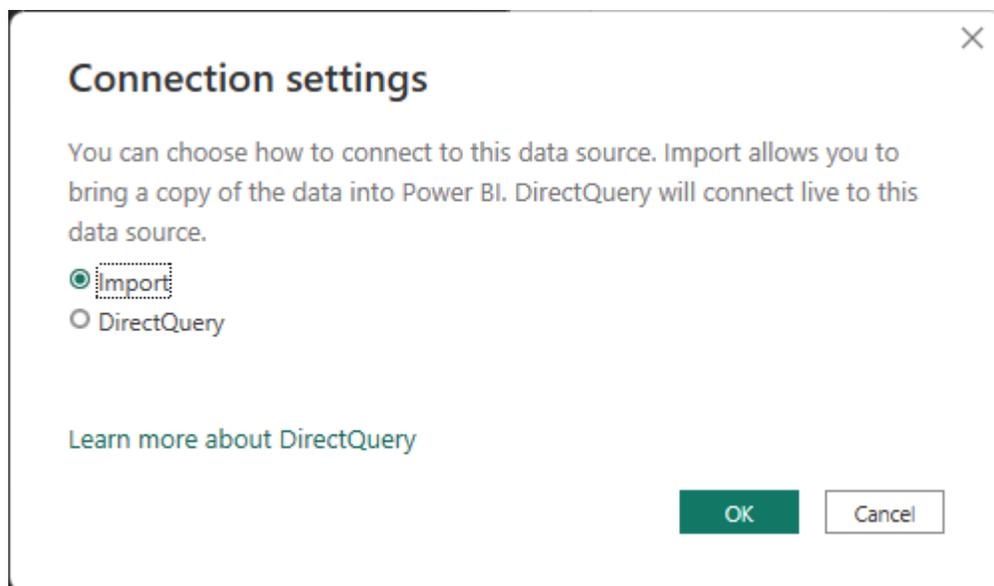


Imagen 58 Listado de tablas en PowerBi

Por cuestiones de costos, se ha decidido usar la opción de importar para cargar los datos de la base de datos



*Imagen 59 Selección de tipo de conexión a Redshift en PowerBi*

## Presupuesto de implementación

Para que la solución empiece a operar y se mantenga operando es necesario tomar en cuenta lo siguientes costos. Los costos se calcularon de manera mensual.

### Costo de S3

	<b>Cantidad</b>	<b>Costo</b>	<b>Total</b>
Almacenamiento de S3 Estándar	160 GB	\$0.023 x GB	\$3.68
Solicitudes PUT, COPY, POST y LIST a S3 Standard	2000 solicitudes	\$0.000005 x Solicitud	\$0.01
Solicitudes GET, SELECT y todas las demás desde S3 Estándar	2000 solicitudes	\$0.0000004 x Solicitud	\$0.0008
Datos devueltos por S3	10 GB	\$0.0007 x GB	\$0.007
<b>Costo mensual de S3</b>			<b>\$3.70</b>

### Costo de Amazon Redshift.

Por razones de optimizar los recursos económicos, se utilizará solamente un nodo y solo 10 horas al mes, esto quiere decir que se deberá tener encendido el nodo máximo 2.5 horas por cada semana del mes.

**Costo mensual de redshift = Cantidad de nodos X \$4.8 por hora X cantidad de horas activas**

**Costo de redshift = 1 X \$4.80/hora X 10 hora**

**Costo mensual redshift = \$48.00**

### **Costo de Power Bi.**

Para la creación de visualizaciones se necesita una licencia de Power Bi Pro, la cual cuesta \$9.99 mensuales por un usuario, como se piensa utilizar solamente por una persona, esto cubriría la necesidad de la empresa.

### **Costo de Soporte.**

Adicionalmente a los costos de Amazon Redshift y S3, se necesita de un ingeniero de datos con cualidades para dar mantenimiento a la solución y soporte a la empresa, así como la construcción inicial de las visualizaciones en Power Bi, se podría requerir de sus servicios aproximadamente 10 horas al mes con un costo de \$10 por hora.

**Costo de soporte = \$10/hora \* 10 hora**

**Costo de soporte = \$ 100.00**

### **Costo total mensual.**

Por lo tanto, el costo de implementación y posterior funcionamiento es el siguiente.

**Costo Total Mensual = Costo de S3 + Costo de Redshift + Costo de Power Bi + Costo de Soporte**

**Costo Total Mensual = \$3.70 + \$48.00 + \$9.99 + \$100.00**

**Costo Total Mensual = \$161.69**

El costo total mensual para que la solución de Data Warehouse se mantenga operando es de aproximadamente **\$151.70**, cabe destacar que durante la implementación es necesario el costo

de soporte, pero una vez que la solución este operando queda a total criterio de la empresa cuando necesitara soporte y la cantidad de horas que se invertirán en dado caso.

### **Tamaño de la base de datos**

- Base de datos transaccional: A la hora de cargar los datos actuales de la base de datos se obtuvo un tamaño total de 62MB, esto en un tiempo de 1 mes, si el crecimiento se mantiene constante durante 24 meses se espera un tamaño total de 1.49GB al final de este período.
- Datawarehouse: La base de datos para nuestro datawarehouse después de las ventas de un mes tuvo un tamaño de 24MB, si el crecimiento se mantiene constante durante 24 meses, se espera un tamaño total de 0.58GB al final de este período, por lo cual, este peso se mantiene en el rango aceptado para la solución propuesta en Amazon Redshift.
- Cantidad de ventas: La empresa en un mes ha generado datos para 3,167 ventas, si se mantiene un crecimiento constante durante 24 meses, al final de este período se tendría un total de datos para 76,008 ventas.

Todas estas estimaciones pueden variar por diferentes razones, por ejemplo, meses de temporadas altas, movimientos en las sucursales, nuevos productos, publicidad, etc. Pero estos datos deberían de dar un estimado útil para tener noción del tamaño de cada una de las bases de datos que está siendo usada en la solución.

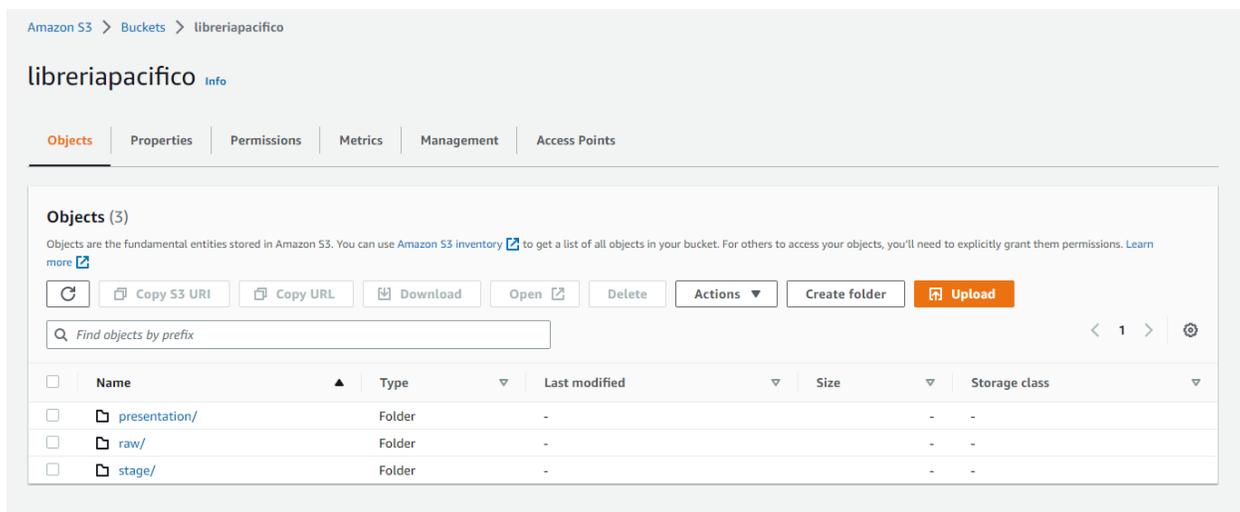
### **Requerimientos mínimos de hardware.**

Para la ejecución de los procesos que extraen, transforman y cargan los datos de las ventas en nuestro datawarehouse se necesitará una computadora con las siguientes especificaciones:

- Sistema Operativo: Windows 11 Home (\$139)
- CPU: Intel Core i5-12400 (\$182)
- Memoria RAM: 16GB (\$45)
- Motherboard: GIGABYTE H610M S2H (\$90)
- Almacenamiento: SSD 1TB (\$60)
- Fuente de poder: 550W 80+ Bronze (\$60)

## Análisis de resultados

En el bucket de S3 llamado libreriapacifico, se encuentran los 3 layers que están siendo utilizados para la solución del Data Warehouse, los datos están siendo cargados mediante los procesos ETL.



*Imagen 60 Layers utilizados en S3 para la solución del DW*

A continuación, se pueden observar las tablas que fueron creadas después de la ejecución del script en Redshift. [Anexo 1.](#)

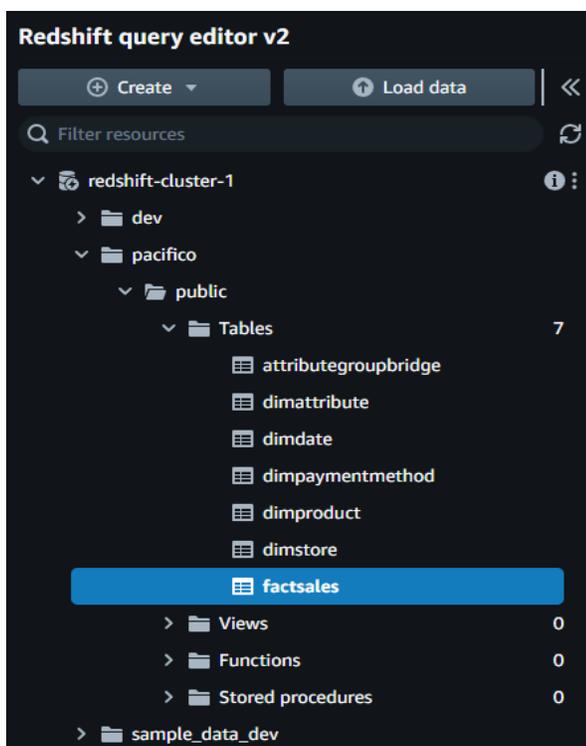


Imagen 61 Tablas creadas en la base de datos pacifico que se encuentra alojada en Redshift

En la tabla *factsales* se puede verificar los datos que obtenidos mediante la solución del Data Warehouse, estos también se encuentran disponibles en la capa de presentación en el bucket *libreriapacifico* en S3

The screenshot displays the Redshift query editor interface. On the left, a navigation pane shows the database structure for 'redshift-cluster-1', including schemas 'dev', 'public', and 'sample\_data\_dev'. The 'public' schema contains seven tables: 'attributegroupbridge', 'dimattribute', 'dimdate', 'dimpaymentmethod', 'dimproduct', 'dimstore', and 'factsales'. The main editor area shows a SQL query: 'SELECT \* FROM "public"."factsales";'. Below the query, the 'Run' button is visible, along with options for 'Limit 100', 'Explain', and 'Isolated session'. The results pane at the bottom shows a table with 20 rows and 8 columns. The columns are: 'sale\_id', 'date\_key', 'product\_key', 'store\_key', 'payment\_method\_key', 'pos\_transaction', and 'sales\_q'. The data represents individual sales transactions with their respective keys and quantities.

sale_id	date_key	product_key	store_key	payment_method_key	pos_transaction	sales_q
1	20220616	8	1	1	1	3
2	20220616	23	1	1	2	1
3	20220616	12	1	1	2	1
4	20220616	15	1	1	3	1
5	20220616	1	1	1	3	1
6	20220616	19	1	1	4	10
7	20220616	22	1	1	5	5
8	20220616	8	1	1	5	5
9	20220616	32	1	1	6	5
10	20220616	31	1	1	7	1
11	20220616	22	1	1	7	10
12	20220616	30	1	1	8	1
13	20220616	22	1	1	8	4
14	20220616	21	1	1	9	10
15	20220616	17	1	1	9	10
16	20220616	14	1	1	10	5
17	20220616	24	1	1	10	5
18	20220616	11	1	1	12	1
19	20220616	17	1	1	12	2
20	20220616	30	1	1	13	1

Imagen 62 Visualización en Redshift de datos obtenidos mediante los procesos ETL

En PowerBI, después de consumir los datos que tenemos en nuestro clúster de Redshift, hemos construido las siguientes visualizaciones basadas en las métricas solicitadas por Librería Pacífico:

- Cantidad de ventas por productos



Imagen 63 Métrica: Cantidad de ventas por productos

Se puede observar que el artículo más vendido es la silla de oficina, con más de 200 productos vendidos, también se puede apreciar por medio de los filtros de método de pago, que la gran mayoría de productos se ha vendido utilizando dinero en efectivo, esto debido a que, por medio de este método de pago, existía un descuento aplicado a productos, de igual manera, el producto menos vendido fue la fotocopiadora, con una cantidad de 9 ventas en este periodo de tiempo.

- Volumen de ventas por productos



Imagen 64 Métrica: Volumen de ventas por productos

La mesa para reuniones larga ha sido el producto con mayor volumen de venta, esto es debido a su precio y a la cantidad de artículos que se ha vendido, su volumen supera los \$120,000, en el otro lado, tenemos que el borrador y la caja de grapas son los artículos con menor volumen de ventas, estos son artículos que se han agregado hace poco al inventario de la librería por lo que no se esperaba un gran volumen de ventas para ellos.

- Costo por producto vendido



**Librería Pacífico**

### Costo por producto vendido

Costo por producto vendido

Nombre del producto	Costo del producto
Acetatos para fotocopiadora	\$4.50
Acoustic Bloc Screens	\$265.50
Agenda	\$13.50
Borrador	\$0.45
Caballote de mesa portatil	\$40.50
Cabinet with Doors	\$126.00
Caja de grapas	\$0.89
Caja de lapices	\$3.60
Caja de plumones	\$4.95
Cartuchos de tinta	\$18.00
Conference Chair	\$29.70
Conference Chair	\$35.46
Corner Desk Left Sit	\$76.50
Corner Desk Right Sit	\$132.30
Cuaderno de contabilidad	\$4.50
Customizable Desk	\$675.00
Customizable Desk	\$720.36
Desk Combination	\$405.00
Desk Stand with Screen	\$1,890.00
Drawer Black	\$22.50
Engrapadora	\$2.65
Flipover	\$1,755.00
Folder tama?o carta	\$3.59
Folder tama?o oficio	\$0.90

*Imagen 65 Métrica: Costo por producto vendido*

Se puede observar como los diferentes costos de productos varían de una gran manera entre categorías, como puede ser que una persona personal cueste más que artículos de librería, pero estas computadoras se venden por más y se tiene una ganancia mayor en comparación a artículos más pequeños y menos costosos.

- Margen de ventas

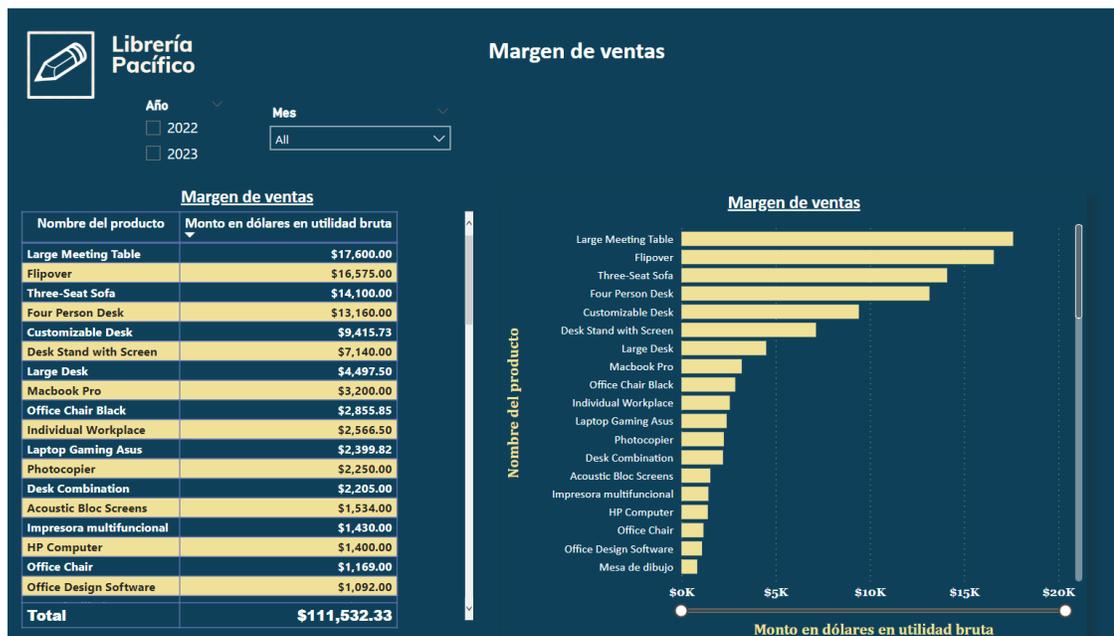


Imagen 66 Métrica: Margen de ventas

El artículo con mayor margen de ventas es la MacBook Pro, que supera por muchos a los demás artículos, esto debido a que su costo es bajo a comparación a lo que se vende, los demás productos se mantienen con un margen estándar.

- Monto de descuentos



Imagen 67 Métrica: Monto de descuentos

El producto que tuvo la mayor sumatoria de descuentos es el escritorio personalizable, seguidos por la silla de oficina y los espacios individual de trabajo, esto ya que la mayoría de ventas para estos productos fueron pagados con dinero en efectivo.

- Cantidad de impuesto pagado por producto

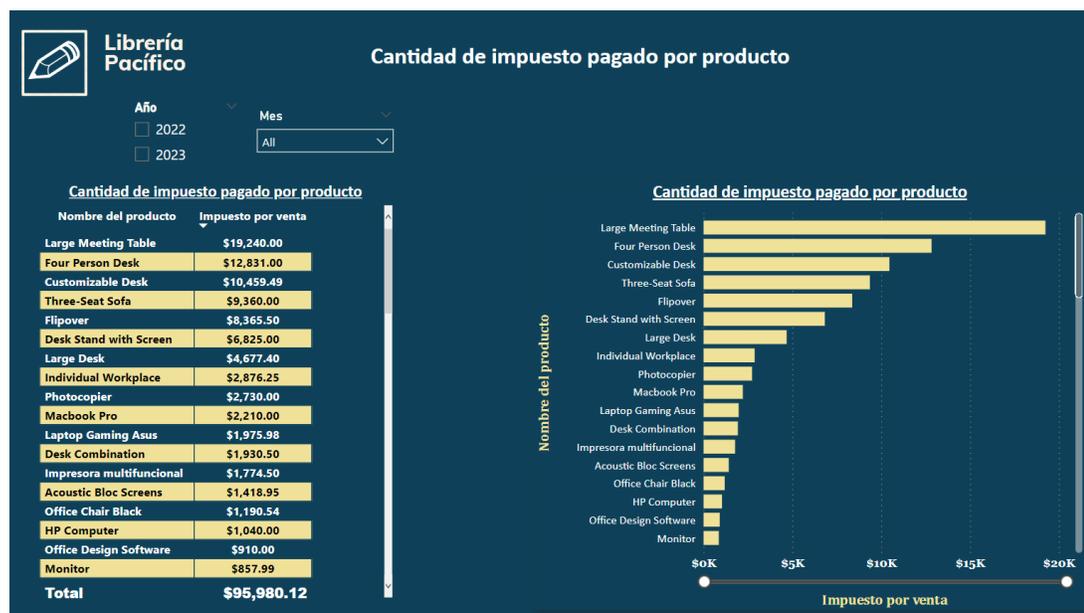


Imagen 68 Métrica: Cantidad de impuesto pagado por producto

Se pago un total \$71,296.03 en impuestos, siendo la mesa grande de reuniones fue el producto con el cual se debió pagar más impuesto, superando los \$15,000, los artículos que le siguen son el escritorio de cuatro personas, el escritorio personalizable y el sofá con 3 asientos.

## Conclusiones y recomendaciones

### Conclusiones

- Durante la elaboración de la solución de Data Warehouse para la empresa Librería del Pacífico que utiliza el software ERP Odoo, se aplicaron los conocimientos adquiridos durante la especialización de ingeniería de datos.
- Se diseñó el modelo dimensional haciendo uso de la metodología de construcción de un Data Warehouse, lo cual facilitó la creación de mismo. El modelado dimensional está basado en las necesidades de la empresa Librería del Pacífico.
- El análisis de datos mediante la herramienta DataCleaner, permitió conocer a detalle el estado y característica de los datos utilizados en el proceso de ventas de la empresa Librería del Pacífico. Así como permitió un mejor desarrollo de los procesos ETL.
- A través de la capa de presentación elaborada en la solución de Data Warehouse los datos son consumidos mediante una aplicación de BI, la cual permite presentar visualizaciones de los datos haciendo uso de informes interactivos y que son fácilmente legibles por los usuarios.

### Recomendaciones

- La implementación de la solución de Data Warehouse debe ser realizada de preferencia por un ingeniero de datos que tenga conocimientos en los servicios de AWS, Talend Open Studio, Postgresql y Power Bi.
- El master job debe de ser ejecutado como se describió en el apartado de estrategia de implementación en el punto **Políticas de ejecución**.

- Se recomienda seguir utilizando la metodología ocupada durante el desarrollo de la solución de Data Warehouse para que se acomode a las necesidades planteadas por la organización, esto previo a cualquier cambio que se realice en la solución.

## Bibliografía

*5 minutos para saber más sobre Microsoft Power BI.* (s.f.). Obtenido de Ibermatica365.com:

<https://www.ibermatica365.com/todo-lo-que-siempre-quisiste-saber-sobremicrosoft-power-bi>

*AWS Pricing Calculator.* (s.f.). Obtenido de AWS Pricing Calculator: <https://calculator.aws/#/>

*Cloud Data Warehouse - Amazon Redshift - Amazon Web Services.* (s.f.). Obtenido de Amazon

Web Services: [https://aws.amazon.com/redshift/?nc1=h\\_ls](https://aws.amazon.com/redshift/?nc1=h_ls)

*End of life information - Talend.* (16 de Julio de 2021). Obtenido de Help Talend:

<https://help.talend.com/r/en-US/8.0/release-notes/documentation>

*Git - Book.* (s.f.). Obtenido de git: <https://git-scm.com/book/en/v2>

*HOW TO GET ACCESS KEY ID AND SECRET ACCESS KEY OF AMAZON S3 ACCOUNT?*

(s.f.). Obtenido de Supsysitic: <https://supsysitic.com/documentation/id-secret-access-key-amazon-s3/>

*Introducción a Amazon Redshift.* (30 de Diciembre de 2022). Obtenido de Aprender BIG DATA:

<https://aprenderbigdata.com/redshift/>

*Introducción a Amazon S3.* (28 de Noviembre de 2022). Obtenido de Aprender BIG DATA:

<https://aprenderbigdata.com/amazon-s3/>

NGuerrero. (22 de Enero de 2021). *¿Qué es IAM? en AWS.* Obtenido de Programa en Linea:

<https://www.programaenlinea.net/que-es-iam-en-aws/>

Pedamkar, P. (s.f.). *Talend Open Studio.* Obtenido de Educba: [https://www.educba.com/talend-](https://www.educba.com/talend-open-studio/)

[open-studio/](https://www.educba.com/talend-open-studio/)

*Power BI - Quick Guide.* (s.f.). Obtenido de tutorialspoint:

[https://www.tutorialspoint.com/power\\_bi/power\\_bi\\_quick\\_guide.htm](https://www.tutorialspoint.com/power_bi/power_bi_quick_guide.htm)

*Qué es Power BI.* (s.f.). Obtenido de Arimetrics: <https://www.arimetrics.com/glosario-digital/power-bi>

Redacci[on KeepCoding. (23 de Mayo de 2022). *¿Qué es Talend Open Studio?* Obtenido de KeepCoding Tech Shcool: <https://keepcoding.io/blog/talend-open-studio/>

Ross, M. &. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling.* Wiley.

Sarad. (7 de Enero de 2022). *AWS Redshift 101: A Guide to Redshift Architecture & More.* Obtenido de HEV0: <https://hevodata.com/blog/amazon-redshift-data-warehouse/>

*Secure and resizable cloud compute - Amazon EC2.* (s.f.). Obtenido de Amazon Web Service: [https://aws.amazon.com/ec2/?nc1=h\\_ls](https://aws.amazon.com/ec2/?nc1=h_ls)

*What is Amazon S3? - Amazon Simple Storage Service.* (s.f.). Obtenido de Amazon AWS: <https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html>

*What is IAM? - AWS Identity and Access Management.* (s.f.). Obtenido de Amazon AWS: <https://docs.aws.amazon.com/IAM/latest/UserGuide/introduction.html>

## Anexos

### Anexo 1.

Script de la base de datos en Redshift, es necesario primero crear la base de datos en redshift y luego ejecutar el siguiente script SQL.

```

CREATE TABLE "public"."attributegroupbridge" (attribute_group_key
integer NOT NULL encode az64,
    attribute_key      integer encode az64,
    product_key        integer encode az64,
    CONSTRAINT attributegroupbridge_pkey PRIMARY
KEY(attribute_group_key),
    CONSTRAINT attributegroupbridge_attribute_key_fkey FOREIGN KEY
(attribute_key) REFERENCES dimattribute(attribute_key),
    CONSTRAINT attributegroupbridge_product_key_fkey FOREIGN KEY
(product_key) REFERENCES dimproduct(product_key));

CREATE TABLE "public"."dimattribute" (attribute_key integer NOT NULL
encode az64,
    attribute_id      integer encode az64,
    name              character varying(256) encode lzo,
    value             character varying(256) encode lzo,
    CONSTRAINT dimattribute_pkey PRIMARY KEY(attribute_key));

CREATE TABLE "public"."dimproduct" (product_key integer NOT NULL encode
az64,
    product_id        integer encode az64,
    product_name      character varying(256) encode lzo,
    product_description character varying(256) encode lzo,
    category          character varying(256) encode lzo,
    category_description character varying(256) encode lzo,
    weight            integer encode az64,
    volume            character varying(256) encode lzo,
    barcode           character varying(256) encode lzo,
    product_code      character varying(256) encode lzo,
    CONSTRAINT dimproduct_pkey PRIMARY KEY(product_key));

CREATE TABLE "public"."dimdate" (datenum integer NOT NULL encode az64,
    date              character varying(256) encode lzo,
    yearmonthnum      integer encode az64,
    calendar_quarter  character varying(256) encode lzo,
    monthnum          integer encode az64,
    monthname         character varying(256) encode lzo,
    monthshortname    character varying(256) encode lzo,
    weeknum           integer encode az64,
    daynumofyear      integer encode az64,
    daynumofmonth     integer encode az64,
    daynumofweek      integer encode az64,
    dayname           character varying(256) encode lzo,

```

```

    dayshortname      character varying(256) encode lzo,
    quarter           integer encode az64,
    yearquarternum    integer encode az64,
    daynumofquarter   integer encode az64,
    CONSTRAINT dimdate_pkey PRIMARY KEY(datenum)) distkey(datenum)
compound sortkey(datenum);
CREATE TABLE "public"."dimpaymentmethod"(payment_method_key integer NOT
NULL encode az64,
    payment_method_id integer encode az64,
    name                character varying(256) encode lzo,
    iscash              boolean,
    CONSTRAINT dimpaymentmethod_pkey PRIMARY KEY(payment_method_key));

CREATE TABLE "public"."dimstore"(store_key integer NOT NULL encode
az64,
    store_id           integer encode az64,
    name               character varying(256) encode lzo,
    CONSTRAINT dimstore_pkey PRIMARY KEY(store_key));

CREATE TABLE "public"."factsales"(sale_id integer NOT NULL encode az64,
    date_key           integer encode az64,
    product_key        integer encode az64,
    store_key          integer encode az64,
    payment_method_key integer encode az64,
    pos_transaction    integer encode az64,
    sales_quantity     integer encode az64,
    regular_unit_price numeric(18,0) encode az64,
    discount_unit_price numeric(18,0) encode az64,
    net_unit_price      numeric(18,0) encode az64,
    extended_discount_dollar_amount numeric(18,0) encode az64,
    extended_sales_dollar_amount    numeric(18,0) encode az64,
    extended_gross_profit_dollar_amount numeric(18,0) encode az64,
    CONSTRAINT factsales_pkey PRIMARY KEY(sale_id),
    CONSTRAINT factsales_payment_method_key_fkey FOREIGN KEY
(payment_method_key) REFERENCES dimpaymentmethod(payment_method_key),
    CONSTRAINT factsales_store_key_fkey FOREIGN KEY (store_key)
REFERENCES dimstore(store_key),
    CONSTRAINT factsales_product_key_fkey FOREIGN KEY (product_key)
REFERENCES dimproduct(product_key),
    CONSTRAINT factsales_date_key_fkey FOREIGN KEY (date_key)
REFERENCES dimdate(datenum));

```

**Anexo 2.**

Mapping de la fact table.

Table Name	Sales
Table Type	Fact
View Name	Sales
Display Name	Sales
Description	Sales captures each sales transaction at any point of sale

Target											Source						
Column Name	Display Name	Description	DataType	Size	Precision	Key?	FK To	ULL?	Default Value	Example Values	Source System	Source Schema	Source Table	Source Field Name	Source Datatype	Extraction/Transformation Rules	Comments
_key_date	date key	Key to date dimension	nt			K	DimDate.DateKey			,2,3	erived					Key lookup from	
uct_key	product key	Key to product dimension	nt			K	DimProduct.ProductKey			,2,3	erived					Key lookup from	
omer_key	customer Key	Key to customer dimension	nt			K	DimCustomer.CustomerKey			,2,3	erived					Key lookup from	
_key_store	store key	Key to store dimension	nt			K	DimStore.StoreKey			,2,3	erived					Key lookup from	
payment_method_key	payment Method key	Key to payment method dimension	nt			K	DimPayment.PaymentMethodKey			,2,3	erived					Key lookup from	

transaction_pos_	os Transaction	unique pos transaction number	nt			K											
sales_quantity	ales Quantity	old products amount	nt							do	ublic	os_order_line	ty	umeric	ap Null to zero	M	
regular_unit_price	egular Unit price	tandard price for this item	umeric							do	ublic	os_order_line	rice_unit	umeric	ap Null to zero	M	
discount_unit_price	iscount Unit price	iscount price for this item	umeric							erived					rice_unit * discount	P	
net_unit_price	et Unit price	et price for this item	umeric							erived	ublic				(unit_price + (unit_price * tax_amount) - unit_price * discount) * sales_quantity		
extended_discount_dollar_amount	xtended Discount Dollar Amount	otal discount	umeric							erived	ublic				(regular_unit_price - discount_unit_price) * sales_quantity		
extended_sales_dollar_amount	xtended Sales Dollar Amount	otal price	umeric							erived	ublic				et_unit_price * sales_quantity	n	
extended_gross_profit_dollar_amount	xtended Gross Profit Dollar amount	otal gross profit for this item	umeric							erived	ublic				(net_unit_price - unit_price) * sales_quantity		

### Anexo 3

Mapping de la tabla Product Dimension.

Table Name	ProductDimension
View Name	Product
Display Name	Product
Description	The product dimension includes all available products
Used in schemas	Sales

Target										Source								
Column Name	Display Name	Description	Attribute Group	DataType	Size	Precision	Key?	Key To	NULL?	Default Value	CD Type	Source System	Source Schema	Source Table	Source Field Name	Source Datatype	TL Rules	Comments
product_key	Product Key	Surrogate primary key		int			K											
product_name	Product Name	Product Name		varchar	50							do	public	product_template	name	varchar		
product_description	Product Description	Product Description		varchar	500							do	public	product_template	description_sale	text	ap NULL to empty string	
category	Category	Category		varchar	10							do	public	product_category	name	varchar		
category_description	Category Description	Category Description		varchar	100							do	public	product_category	complete_name	varchar	ap NULL to empty string	
weight	Weight	Weight		numeric								do	public	product_template	weight	numeric	ap NULL to zero value	



**Anexo 4.**

Mapping de la table Store Dimension

Table Name	Store Dimension
View Name	Store
Display Name	Store
Description	The store where the order was created
Used in schemas	Sale

Target												Source						
Column Name	Display Name	Description	Attribute Group	Data Type	Size	Precision	Key?	Key To	NULL?	Default Value	CD Type	Source System	Source Schema	Source Table	Source Field Name	Source Datatype	TL Rules	Comments
store_key	store Key	surrogate primary key		int			K											
name	name	store name		varchar	5							do	public	os_config	name	varchar		

**Anexo 5.**

Mapping de la table Payment Method Dimension

Table Name	Payment Method Dimension
View Name	Payment Method
Display Name	Payment Method
Description	The payment dimension includes all payment types that the company accepts
Used in schemas	Sales

Target													Source					
Column Name	Display Name	Description	Attribute Group	DataType	Size	Precision	Key?	K To	ULL?	Default Value	CD Type	Source System	Source Schema	Source Table	Source Field Name	Source Datatype	TL Rules	Comments
payment_method_key	Payment Method Key	Surrogate primary key		Int			K											
name	name	the name of the payment method		varchar	5							do	public	pos_payment_method	name	varchar		
is_cash	is Cash Count	Verify if the payment method is cash only		byte								do	public	pos_payment_method	is_cash_count	bool	ap Null to zero value	zero value means that is not cash method payment type





