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

ANALISIS DE LAS VENTAS DE LA EMPRESA IBW COMUNICACIONES

PRESENTADO POR:

FRANCISCO ARNOLDO ORELLANA REYES
OSCAR ERNESTO MAYEN DE LA CRUZ
RAFAEL ENRIQUE AVILES CORNEJO

PARA OPTAR AL TÍTULO DE:
INGENIERO DE SISTEMAS INFORMÁTICOS

CIUDAD UNIVERSITARIA, DICIEMBRE 2023

UNIVERSIDAD DE EL SALVADOR

RECTOR:

M.SC. JUAN ROSA QUINTANILLA

SECRETARIO GENERAL:

LIC. PEDRO ROSALIO ESCOBAR CASTANEDA
FACULTAD DE INGENIERÍA Y ARQUITECTURA

DECANO:

ING. LUIS SALVADOR BARRERA MANCIA

SECRETARIO:

ARQ. RAUL ALEXANDER FABIAN ORELLANA
ESCUELA DE INGENIERÍA DE SISTEMAS INFORMÁTICOS

DIRECTOR:

ING. CESAR AUGUSTO GONZALEZ

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

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

Título:
ANALISIS DE LAS VENTAS DE LA EMPRESA IBW COMUNICACIONES

Presentado por:
FRANCISCO ARNOLDO ORELLANA REYES
OSCAR ERNESTO MAYEN DE LA CRUZ
RAFAEL ENRIQUE AVILES CORNEJO

Trabajo de Graduación Aprobado por:
Docente Asesor:
ING. RENE FABRICIO QUINTANILLA GOMEZ

SAN SALVADOR, DICIEMBRE 2023

Trabajo de Graduación Aprobado por:

Docente Asesor:

Ing. Rene Fabricio Quintanilla Gómez

INDICE

Introducción

La ingeniería de datos juega un papel crucial en la actualidad, donde la información y la tecnología son elementos fundamentales para el eficiente funcionamiento empresarial. En este escenario, IBW Comunicaciones, empresa especializada en servicios de Internet, enfrenta desafíos significativos, especialmente porque su sistema transaccional ERPNext no proporciona una información óptima para el análisis y la generación de informes gerenciales.

En el presente documento se propone abordar esta problemática mediante la propuesta y desarrollo de un modelo dimensional específico para IBW Comunicaciones. Nos sumergiremos en la situación actual de la empresa, analizando los desafíos existentes y explorando las razones detrás de la necesidad de implementar un enfoque más robusto en la gestión de datos.

La propuesta de solución abarcará tanto el análisis como el diseño detallado de la estrategia a seguir. Presentaremos una metodología de trabajo que dirigirá el desarrollo e implementación del data Waterhouse, destacando el papel crucial de herramientas clave como Talend para la ejecución de Extract, Transform, Load (ETLs).

Adicionalmente, exploraremos el uso de Amazon Redshift en AWS como componente fundamental para el almacenamiento eficiente de grandes volúmenes de datos que formara parte de la solución para nuestro data Waterhouse.

El documento también profundizará en la implementación de dashboards a través de Power BI, ofreciendo una visión clara y accesible de los análisis gerenciales. Exploraremos los elementos gráficos y funcionales de los dashboards, destacando cómo dicha solución contribuirá a la agilización y mejora en la toma de decisiones.

Capítulo I: Especificación del proyecto

a. Situación Actual

La situación actual de la empresa IBW en relación a la gestión de datos es que carece de un Data Warehouse, lo que afecta significativamente la eficiencia de sus operaciones y la toma de decisiones a nivel gerencial. Actualmente, la generación de informes y análisis se realiza directamente a partir de la base transaccional de la empresa, lo que conlleva varios desafíos.

Uno de los principales problemas radica en que la base transaccional contiene una gran cantidad de información debido a la operación de la empresa de telecomunicaciones. Esto significa que cada vez que se necesita realizar un análisis, se requiere un proceso de extracción y transformación de datos, lo que resulta en una carga significativa para el sistema y, en última instancia, en tiempos de respuesta más lentos. Esta situación afecta la capacidad de los equipos gerenciales para acceder a información crítica de manera oportuna.

Para abordar esta problemática, se ha propuesto la implementación de un Data Warehouse que se centrará en la consolidación y almacenamiento de datos relacionados con las facturas de venta. Este enfoque permitirá a IBW contar con un repositorio centralizado y optimizado de datos que servirá como una fuente única de información para la generación de informes y análisis.

La empresa IBW, que brinda servicios de internet en el sector de las telecomunicaciones, reconoce la importancia de este paso hacia la mejora de su infraestructura de datos. La implementación del Data Warehouse no solo acelerará la generación de informes, sino que también permitirá un análisis más profundo y significativo de los datos relacionados con las facturas de venta. Esto, a su vez, contribuirá a una toma de decisiones más informada y eficaz en todos los niveles de la organización.

En resumen, la situación actual de la empresa IBW se caracteriza por la ausencia de un Data Warehouse, lo que ralentiza la generación de informes y análisis debido a la dependencia de la base transaccional abrumada por la cantidad de datos. La implementación de un Data Warehouse centrado en las facturas de venta se considera una solución esencial para mejorar la eficiencia operativa y la toma de decisiones en la empresa.

i. Antecedentes

La empresa IBW, con una sólida trayectoria en el sector de las telecomunicaciones, ha experimentado un crecimiento constante en los últimos años. Su compromiso con la provisión de servicios de internet de alta calidad ha llevado a un aumento significativo en la base de clientes y, por lo tanto, a una mayor cantidad de datos generados a diario.

A medida que IBW ha continuado expandiendo sus servicios y su presencia en el mercado, se ha vuelto evidente que la gestión de datos es un componente crítico para el éxito de la empresa. Sin embargo, a lo largo de su historia, IBW ha dependido en gran medida de la base transaccional para la generación de informes y análisis. Esta dependencia ha generado desafíos considerables en términos de eficiencia y escalabilidad.

La base transaccional, que antes era suficiente para las necesidades de la empresa, ahora está abrumada por la creciente cantidad de datos generados por las operaciones diarias. Esto ha resultado en tiempos de respuesta lentos, dificultades en la extracción de información específica y un alto grado de complejidad en la generación de informes para el análisis gerencial.

Dado el crecimiento continuo y la importancia de tomar decisiones estratégicas basadas en datos precisos y oportunos, IBW ha decidido abordar estos desafíos implementando un Data Warehouse. Este enfoque permitirá una gestión de datos más eficiente y brindará a la empresa la capacidad de realizar análisis más profundos y significativos sobre los datos relacionados con las facturas de venta.

La implementación de un Data Warehouse representa un paso estratégico en la evolución de IBW, permitiéndole aprovechar al máximo la riqueza de datos generados por sus operaciones y mejorar la toma de decisiones en todos los niveles de la organización. Este cambio marca un paso importante en la búsqueda de la excelencia operativa y el liderazgo en el mercado de las telecomunicaciones.

ii. Descripción del problema

La ingeniería de datos es un área que muchas empresas siguen subestimando a la hora de convertir sus datos en valor añadido. La falta de utilización de la ingeniería de datos en las empresas puede dar lugar a diversas problemáticas que afectan su eficiencia y capacidad para tomar decisiones informadas.

No utilizar en las empresas ingeniería de datos puede generar una serie de desafíos que afectan la capacidad de la empresa para competir, innovar y operar eficientemente en un entorno empresarial cada vez más impulsado por los datos. La implementación de prácticas sólidas de ingeniería de datos es crucial para superar estas dificultades y aprovechar al máximo el potencial de los datos en el mundo empresarial actual.

Algunas de las principales dificultades:

- ✓ **Problemas de escalabilidad:** A medida que una empresa crece, la cantidad de datos también aumenta exponencialmente. La falta de una infraestructura adecuada para la gestión de grandes volúmenes de datos puede generar cuellos de botella, afectando el rendimiento y la escalabilidad de los sistemas.

- ✓ **Seguridad y cumplimiento normativo:** Sin una sólida ingeniería de datos, la seguridad de los datos puede estar comprometida, lo que podría resultar en violaciones de privacidad y problemas de cumplimiento normativo. La gestión inadecuada de los datos también puede exponer a la empresa a riesgos de ciberseguridad.

- ✓ **Ineficiencia en el procesamiento de datos:** La ingeniería de datos incluye la optimización de procesos para el procesamiento eficiente de datos. La falta de esta optimización puede dar lugar a tiempos de respuesta lentos, lo que afecta la capacidad de la empresa para obtener información en tiempo real.

- ✓ **Desafíos en la toma de decisiones estratégicas:** La toma de decisiones informadas se ve obstaculizada cuando los datos no están disponibles o no son confiables. Esto puede afectar la capacidad de la empresa para identificar oportunidades, anticipar tendencias del mercado y competir eficazmente.

iii. Planteamiento de problema

La empresa IBW, destacada en el competitivo mercado de las telecomunicaciones en El Salvador y reconocida por su excelencia en servicios de Internet, enfrenta un desafío crítico en su proceso de toma de decisiones y gestión empresarial. La ausencia de una solución centralizada y eficiente para el análisis de datos limita su capacidad para comprender a fondo el rendimiento de sus ventas, obstaculizando la toma de decisiones informadas que impulsen el crecimiento y la eficiencia operativa.

El problema fundamental radica en la carencia de una plataforma de inteligencia empresarial que permita a IBW recopilar, almacenar, procesar y analizar datos de ventas de manera efectiva. La empresa se enfrenta a desafíos significativos, entre ellos:

Falta de Visibilidad en Tiempo Real con Analisis Limitado: A pesar de contar con un sistema transaccional, ERPNext, para análisis, la velocidad de este proceso a nivel gerencial se ve afectada, limitando la capacidad de IBW para tomar decisiones ágiles y basadas en datos que optimicen estrategias de ventas y marketing.

Informes Dependientes del Sistema Transaccional: En lugar de contar con análisis aislados, IBW depende de informes proporcionados por el sistema transaccional, lo que ralentiza la obtención de información valiosa y precisa.

Con el objetivo de superar estos desafíos y maximizar su potencial de mercado, IBW ha decidido estratégicamente implementar un Data Warehouse integral. Este almacén de datos permitirá a la empresa centralizar sus datos de ventas, mejorará la calidad de los datos, ofrecerá una visión unificada en tiempo real, y facilitará análisis más avanzados.

La implementación de este Data Warehouse proporcionará a IBW la capacidad de realizar análisis de ventas detallados, identificar tendencias, mejorar la segmentación de clientes y, en última instancia, tomar decisiones estratégicas informadas que impulsen el crecimiento y la eficiencia operativa. Con esta solución, IBW estará mejor posicionada para mantener su liderazgo en el competitivo mercado de las telecomunicaciones en El Salvador, al tiempo que sigue ofreciendo servicios de Internet de alta calidad a sus clientes.

b. Objetivos

Objetivo general:

Desarrollar e implementar un modelo dimensional que fortalezca la capacidad de IBW para gestionar y analizar eficazmente sus datos, potenciando la toma de decisiones informadas y estratégicas.

Objetivos específicos:

- ✓ Diseñar e implementar un almacén de datos centralizado, robusto y escalable que albergue información proveniente de todas las fuentes de la empresa.
- ✓ Integrar fuentes de datos heterogéneas, incluyendo bases de datos SQL, sistemas de registro, archivos CSV y aplicaciones de terceros, para asegurar la uniformidad y accesibilidad de la información en un único repositorio.
- ✓ Establecer un proceso automatizado de extracción, transformación y carga (ETL) que garantice la actualización periódica y precisa de los datos almacenados, optimizando la consistencia y la disponibilidad de la información.
- ✓ Implementar Power BI como herramienta de visualización de datos, proporcionando a los usuarios de la empresa la capacidad de crear informes y paneles personalizados para un análisis eficiente y significativo.
- ✓ Facilitar la toma de decisiones ágiles y basadas en datos, empoderando a los diferentes niveles organizativos de IBW con herramientas analíticas que impulsen la eficacia operativa y la competitividad en el mercado de las telecomunicaciones en El Salvador.

c. Alcances y Limitaciones

Alcances

- ✓ Verificar que el modelo dimensional desarrollado cumpla de manera efectiva con los requisitos específicos de negocio de IBW, garantizando que la estructura y los datos respondan a las necesidades estratégicas y operativas de la empresa.

- ✓ Establecer procesos ETL claramente definidos y documentados, que abarquen la extracción, transformación y carga de datos desde diversas fuentes hasta el modelo dimensional. La transparencia en estos procesos es esencial para garantizar la integridad y la consistencia de la información.

- ✓ Ejecutar pruebas rigurosas junto con los usuarios finales para validar la precisión y consistencia de los datos en el modelo dimensional. Esto incluirá casos de uso específicos y escenarios diversos para asegurar la confiabilidad de la información que se utilizará en procesos de toma de decisiones.

- ✓ Lograr una integración efectiva del modelo dimensional con herramientas de visualización de datos, enfocándose especialmente en facilitar la interpretación y presentación de resultados. Esto permitirá a los usuarios explorar y comprender de manera intuitiva la información contenida en el almacén de datos.

- ✓ Generar documentación detallada que abarque diagramas claros del modelo dimensional, definiciones precisas de dimensiones y medidas, así como una explicación exhaustiva de los procesos ETL implementados. Esta documentación servirá como recurso esencial para la comprensión y el mantenimiento continuo del sistema.

Estos alcances se orientan a asegurar no solo la implementación técnica exitosa del modelo dimensional, sino también su alineación con las necesidades y objetivos específicos de negocio de IBW, promoviendo la confiabilidad y utilidad de la solución implementada.

Limitaciones

- ✓ La empresa IBW, por razones de seguridad, no proporcionó acceso directo a la base transaccional. En su lugar, se brindaron archivos CSV como fuente de datos, lo que podría impactar la velocidad y la automatización de los procesos ETL, así como la capacidad de abordar dinámicas en tiempo real de la información.

Esta limitación específica destaca la necesidad de adaptarse a las condiciones impuestas por la empresa IBW y ajustar los procesos ETL en consecuencia, resaltando la importancia de trabajar con datos suministrados de manera segura.

d. **Justificación**

IBW Comunicaciones, destacada en el sector de las telecomunicaciones en El Salvador, se enfrenta a desafíos críticos en la administración y análisis de sus datos empresariales. Se ha identificado que la consulta de reportes a nivel gerencial se ve afectada por su lentitud, lo que ha generado limitaciones significativas en la obtención de información detallada para análisis gerenciales. Esta demora en la ejecución de análisis ha provocado un problema en la capacidad de respuesta de la empresa ante las dinámicas cambiantes del mercado.

La implementación de un modelo dimensional se presenta como una estrategia integral para superar estas limitaciones. Este modelo dimensional no solo representa un almacenamiento centralizado de datos, sino también una oportunidad para integrar fuentes heterogéneas de información. La diversidad de datos se gestionará de manera eficiente, proporcionando a la empresa una visión unificada y completa de su panorama empresarial.

La automatización de procesos ETL constituye un componente clave de este proyecto. La ejecución automatizada de extracción, transformación y carga de datos garantizará la actualización regular y precisa del almacén de datos, mejorando la agilidad y confiabilidad de los análisis realizados.

El enfoque estratégico de dicho análisis no solo radica en abordar las limitaciones tecnológicas, sino en transformar la toma de decisiones en IBW Comunicaciones. Se busca mejorar la capacidad de la empresa para adoptar decisiones informadas y estratégicas, proporcionando análisis más detallados de las ventas. La visión integral que se obtendrá con este modelo contribuirá directamente al crecimiento y la competitividad de IBW Comunicaciones en el dinámico mercado de las telecomunicaciones en El Salvador, estableciéndola como líder innovador en la industria.

e. **Cronograma de actividades**

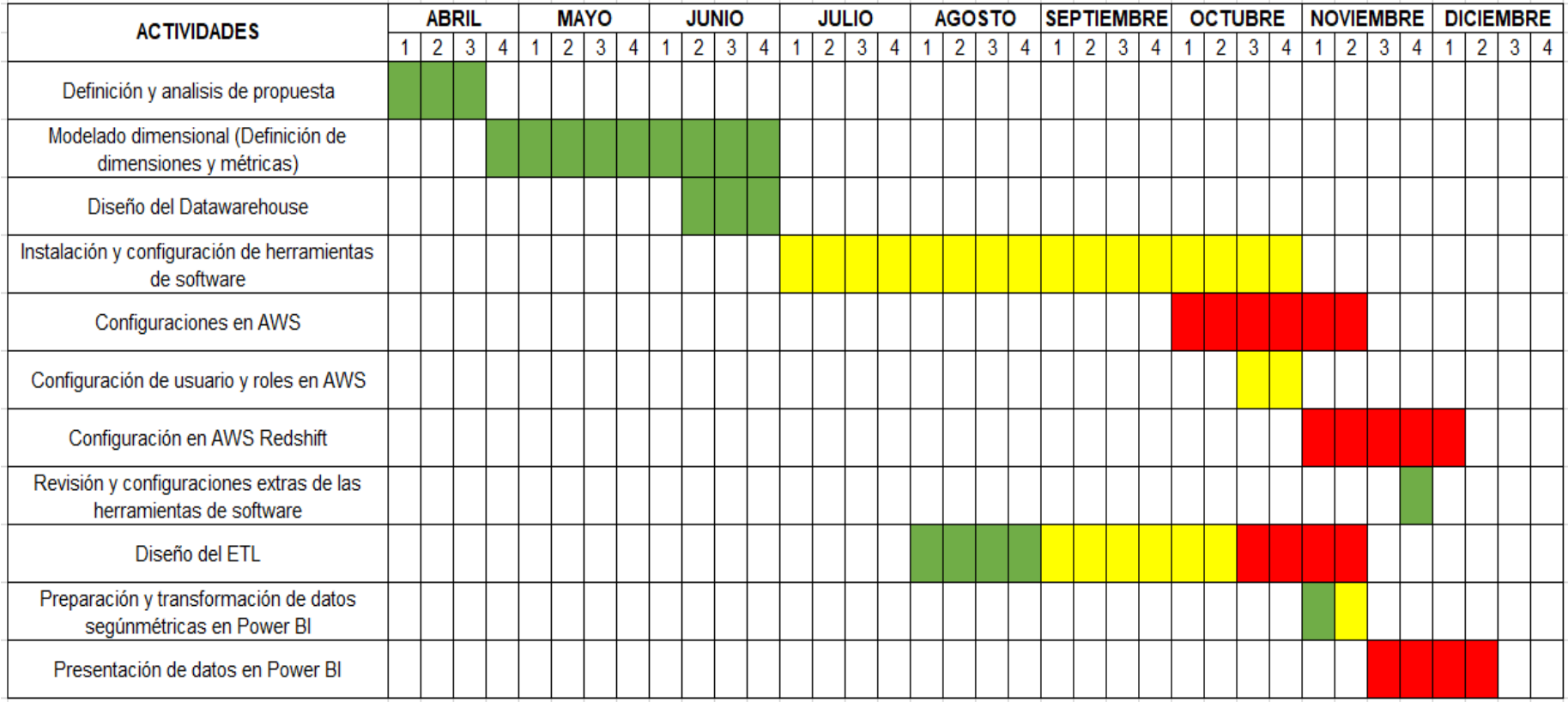

Imagen 1: Cronograma de actividades

f. Presupuesto

El proyecto se llevará a cabo en un período de 9 meses y se dividirá en varias fases, desde la planificación y diseño hasta la implementación y evaluación del Data Warehouse. Las actividades principales incluyen la identificación de requisitos, la selección de tecnologías, el diseño del esquema, la implementación y la validación de la solución.

1. **Recursos Humanos: (Este gasto no se incluirá Ver anexo 3)**
    - **Desarrolladores (3 personas):** $38,880.00 (15.00 por persona al día)

✓ **Subtotal Recursos Humanos:** $38,880.00

2. **Equipamiento y Software:**
    - **Servicio en AWS – S3 y Redshift:** $25.00
    - **Licencias de Software Power BI:** $0.0
    - **Hardware:** $100.00

✓ **Subtotal Equipamiento y Software:** $125.00

3. **Servicios:**
    - **Energía eléctrica:** $405.00
    - **Servicios de internet residencial:** $540.00

✓ **Subtotal Servicios:** $945.00

4. **Gastos Generales:**
    - **Viajes:** $50.00
    - **Misceláneos:** $30.00

✓ **Subtotal Gastos Generales:** $80.00

Costos Totales:

- ✓ **Recursos Humanos:** $0.00
- ✓ **Equipamiento y Software:** $125.00
- ✓ **Servicios:** $945.00
- ✓ **Gastos Generales:** $80.00

Costo Total del Proyecto: $1,150.00

Contingencia (10%): $115.00

Costo Total con Contingencia: $1,265.00

Este presupuesto proporciona una estimación detallada de los costos asociados con la implementación del proyecto de Data Warehouse para IBW. La contingencia del 10% se ha incluido para abordar posibles imprevistos durante la ejecución del proyecto. Se recomienda una revisión periódica del presupuesto para ajustar los costos según sea necesario a lo largo del desarrollo del proyecto.

Capitulo II: Análisis y diseño de la propuesta de solución

a. Metodología de trabajo

El proceso de desarrollo de la solución se ha estructurado en cuatro etapas:

En la etapa del origen de los datos, se abordó la identificación y comprensión de las diversas fuentes de información. En la etapa de creación del modelo dimensional, se llevó a cabo la construcción de un marco estructurado que permitiera una gestión eficiente de los datos. Los Procesos ETL se diseñaron minuciosamente para garantizar la calidad y actualización regular de la información, involucrando tanto a Talend como a las herramientas de AWS. Finalmente, en la etapa de presentación de los datos, se enfocó en la creación de dashboards atractivos y funcionales con Power BI.

Este enfoque estratégico garantiza una implementación coherente y efectiva de la solución, abordando cada fase con precisión y eficiencia, respaldado por las potentes capacidades de herramientas como Amazon Redshift, S3, Talend, y Power BI en el ecosistema AWS.

Origen de datos

La empresa IBW Comunicaciones utiliza con un sistema transaccional denominado ERPNext, cuya base transaccional se encuentra en MySQL. Dicho ERP abarca una amplia variedad de tablas a nivel transaccional, incluyendo campos específicos utilizados internamente por la plataforma. En el marco de nuestro estudio y desarrollo del Data Warehouse, se ha seleccionado un conjunto específico de tablas y campos relevantes, omitiendo aquellos que no son pertinentes para nuestros objetivos.

Es importante señalar que, debido a restricciones de acceso directo a la base transaccional, se proporcionó información crucial en forma de archivos CSV. Estos archivos contienen datos correspondientes a las tablas seleccionadas para la construcción del Data Warehouse. A partir de esta información y en colaboración con la empresa, se ha concebido un detallado diagrama de entidad-relación, que se presenta a continuación.

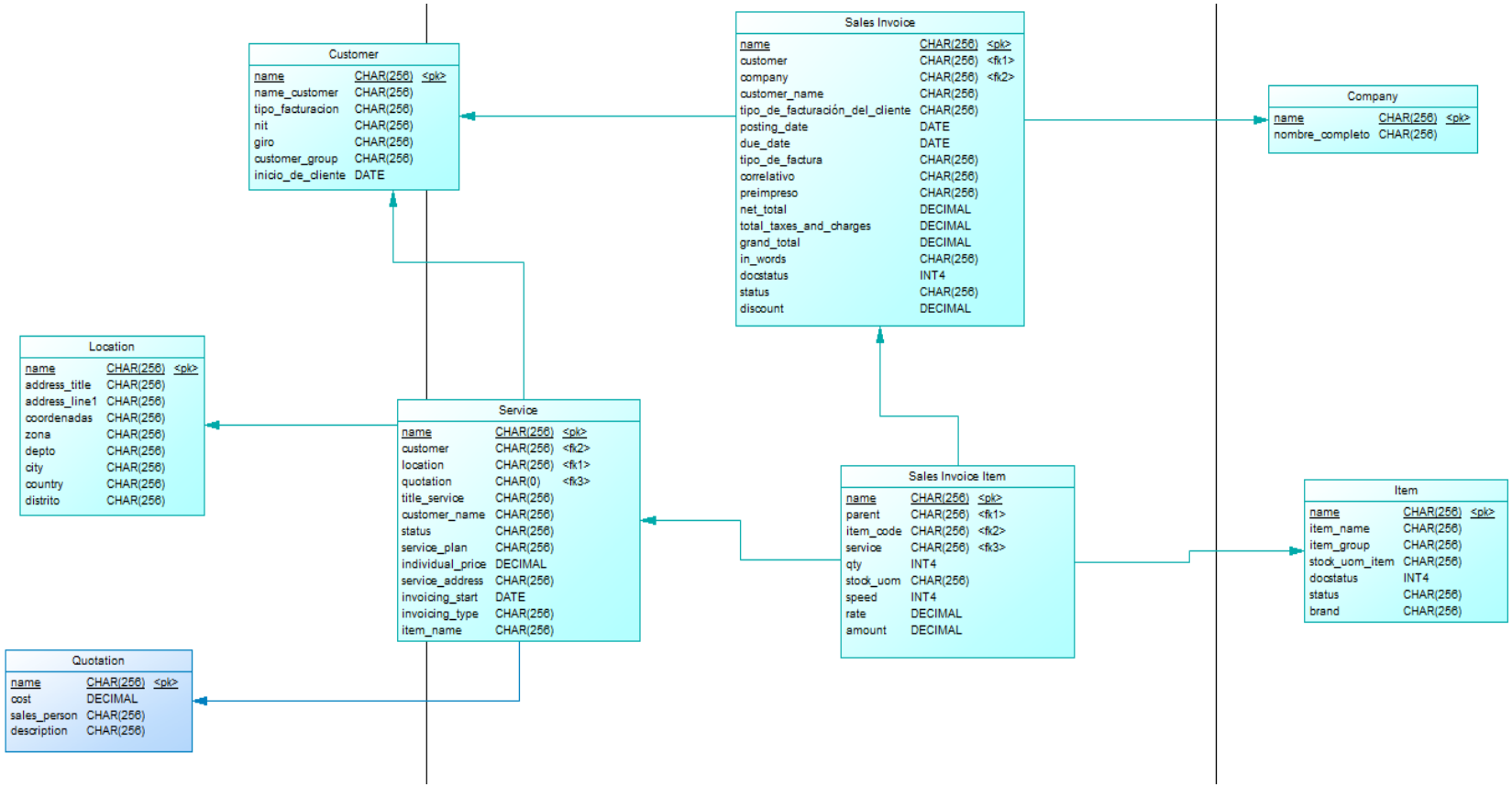

Imagen 2: Diagrama entidad relación

Este enfoque selectivo asegura la extracción y utilización eficiente de datos específicos, alineados con los objetivos del proyecto, y respalda el diseño integral del Data Warehouse a pesar de las limitaciones de acceso directo a la base transaccional

Creación del modelo dimensional

Se usó el modelo dimensional el modelo dimensional propuesto por Ralph Kimball que se basa en una metodología específica que busca simplificar y optimizar la estructura de los datos para facilitar la comprensión y consulta de la información. El proceso de diseño dimensional consta de cuatro pasos clave:

1. **Seleccionar el Proceso de Negocio**

   En este paso, se elige el proceso de negocio central que será el enfoque del modelo dimensional. Este proceso debe ser crítico para la organización y estar alineado con sus objetivos estratégicos.

2. **Definir la Granularidad**

   Este paso implica determinar el nivel de detalle o la granularidad de los datos en el modelo dimensional. Se decide qué nivel de detalle será necesario para analizar el proceso de negocio de manera efectiva.

3. **Identificar las Dimensiones**

   Seleccionar las dimensiones clave que proporcionarán contextos para el análisis. Estas dimensiones pueden incluir aspectos temporales, geográficos, categorías de productos, etc.

4. **Identificar las Métricas**

   En este paso, se eligen las métricas o medidas que se utilizarán para evaluar el rendimiento del proceso de negocio. Estas medidas son esenciales para obtener información cuantitativa relevante.

   Cada uno de estos pasos contribuye a la creación de un modelo dimensional sólido que simplifica la estructura de los datos y facilita el análisis efectivo del proceso de negocio seleccionado.

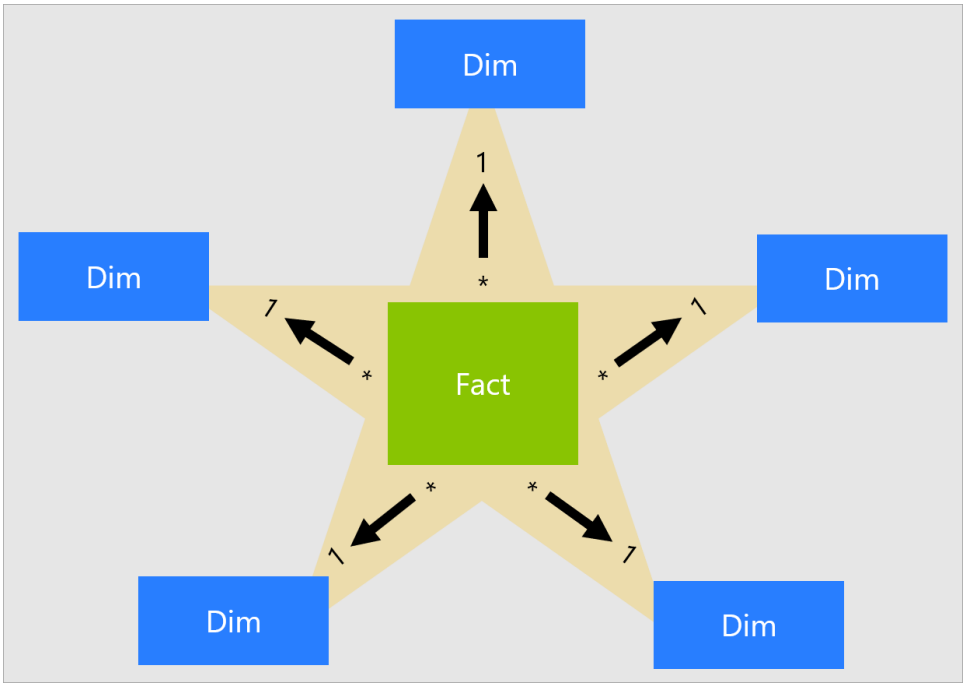

Imagen 3: Modelo Estrella

Procesos ETL

Los procesos ETL son esenciales para la construcción y mantenimiento del Data Warehouse. En este proyecto específico, se incorpora la herramienta Talend Open Studio para facilitar y optimizar las fases clave de estos procesos.

- ✓ **Extracción:**

La primera etapa involucra la extracción de datos desde los archivos CSV proporcionados por IBW Comunicaciones. Estos archivos actúan como la fuente primaria de datos y se almacenan inicialmente en la zona de "Raw". Talend despliega sus capacidades para facilitar una extracción eficiente y estructurada.

- ✓ **Transformación:**

En la fase de transformación, ubicada en la zona de "Staging", Talend desempeña un papel crucial al permitir la limpieza de datos, cambios en los formatos de origen, y la transformación de datos nulos, entre otras manipulaciones. La versatilidad de Talend garantiza una transformación ágil y precisa de los datos para prepararlos de manera óptima para su carga en el Data Warehouse.

- ✓ **Carga:**

Finalmente, los datos transformados y limpios se cargan en la zona de presentación o Data Warehouse. En este proyecto, se utiliza el servicio de Amazon Redshift en AWS para la carga de datos. Amazon Redshift proporciona un entorno altamente escalable y eficiente para almacenar grandes volúmenes de datos y facilita la rápida recuperación de información para análisis posteriores.

- ✓ **Presentación de los datos:**

La fase de presentación de datos se realizará a través de dashboards diseñados con Power BI, conectándose directamente a la base de datos de Amazon Redshift donde reside el Data Warehouse. Power BI permitirá la creación de paneles intuitivos y personalizados que reflejen el cumplimiento de métricas clave. Estos dashboards facilitarán la interpretación de datos, brindando a los usuarios finales la capacidad de analizar el rendimiento del proceso de negocio y tomar decisiones informadas de manera ágil. Power BI actúa como una herramienta integral para transformar los datos almacenados en análisis visuales, contribuyendo al crecimiento estratégico y la eficiencia operativa de IBW Comunicaciones.

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

1. Proceso de negocios

Para este ETL se ha definido un proceso de “Cálculo de ventas (facturas de venta) totales.”

2. Nivel de granularidad

a. ¿Qué detalle requiere el usuario del negocio?

Información sobre las ventas en función de productos, clientes, dirección, por año/mes

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

Obtener información de las ventas en función de todo lo que el usuario de negocio requiere

3. Identificar las dimensiones

Producto, Cliente, Fecha, Dirección

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

- Volumen (cantidad) de lo vendido sobre los servicios de internet. En donde se puede desglosar por monto facturado por zona, por clientes, por productos, entre otros.
- Cantidad de descuentos que se aplican a los servicios de internet.
- Margen de impuestos sobre las ventas generadas.

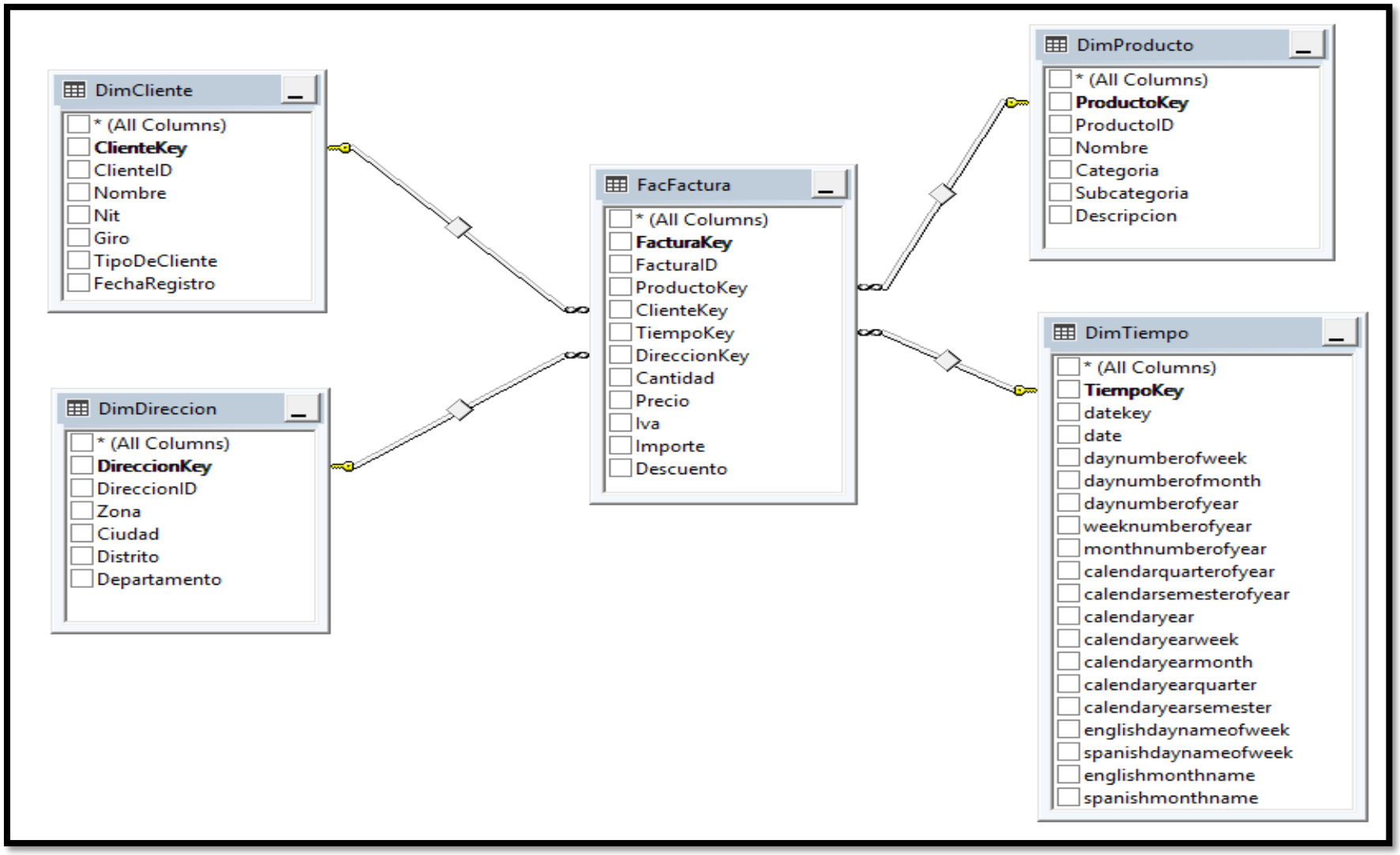

Imagen 4: Modelo dimensional estrella

c. Descripción de la tecnología a utilizar

Para el desarrollo de la solución se han utilizado diferentes herramientas tecnológicas las cuales las describiremos a continuación:

Talend Open Studio

Talend Open Studio es una plataforma de integración de datos de código abierto que proporciona herramientas para facilitar la extracción, transformación y carga (ETL) de datos. Esta herramienta es utilizada por profesionales de datos y desarrolladores para diseñar, implementar y gestionar flujos de trabajo de integración de datos.

En resumen, es una herramienta integral que permite a los profesionales de datos y desarrolladores diseñar, implementar y gestionar procesos de integración de datos de manera eficiente y efectiva. Su enfoque visual, conectividad versátil y capacidades de transformación de datos lo convierten en una opción popular en el ámbito de la integración de datos.

Imagen 5: Logo de Talend Open Studio

Amazon Web Services (AWS)

Amazon Web Services (AWS) es una plataforma de servicios en la nube ofrecida por Amazon.com. Es uno de los proveedores líderes en el espacio de servicios en la nube y proporciona una amplia variedad de servicios y soluciones para satisfacer las necesidades de computación, almacenamiento, análisis, aprendizaje automático, inteligencia artificial, desarrollo de aplicaciones, Internet de las cosas (IoT) y más. Los servicios que se utilizaron para el desarrollo son S3, Redshift.

Imagen 6: Logo de Amazon Web Services (AWS)

Amazon Simple Storage Service (Amazon S3)

Amazon Simple Storage Service (Amazon S3) es un servicio de almacenamiento de objetos que ofrece escalabilidad, disponibilidad de datos, seguridad y rendimiento líderes en el sector. Clientes de todos los tamaños y sectores pueden almacenar y proteger cualquier cantidad de datos para prácticamente cualquier caso de uso, como los lagos de datos, las aplicaciones nativas en la nube y las aplicaciones móviles. Gracias a las clases de almacenamiento rentables y a las características de administración fáciles de usar, es posible optimizar los costos, organizar los datos y configurar controles de acceso detallados para cumplir con requisitos empresariales, organizacionales y de conformidad específicos.

Imagen 7: Logo de Amazon Simple Storage Service(S3)

Amazon Redshift

Amazon Redshift es un producto de almacenamiento de datos que forma parte de la plataforma más grande de computación en la nube Amazon Web Services. Está construido sobre la tecnología de la empresa de almacenamiento de datos de procesamiento paralelo masivo (MPP) ParAccel, para manejar conjuntos de datos a gran escala y migraciones de bases de datos. Redshift se diferencia de la otra oferta de bases de datos alojadas de Amazon, Amazon RDS, en su capacidad para manejar cargas de trabajo analíticas en conjuntos de datos de gran tamaño almacenados mediante un principio DBMS orientado a columnas. Redshift permite hasta 16 petabytes de datos en un clúster en comparación con el tamaño máximo de 128 terabytes de Amazon RDS Aurora.

Imagen 8: Logo de Amazon Redshift

Power BI

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

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.

Imagen 9: Logo de Power BI

d. Diagrama arquitectónico de la solución

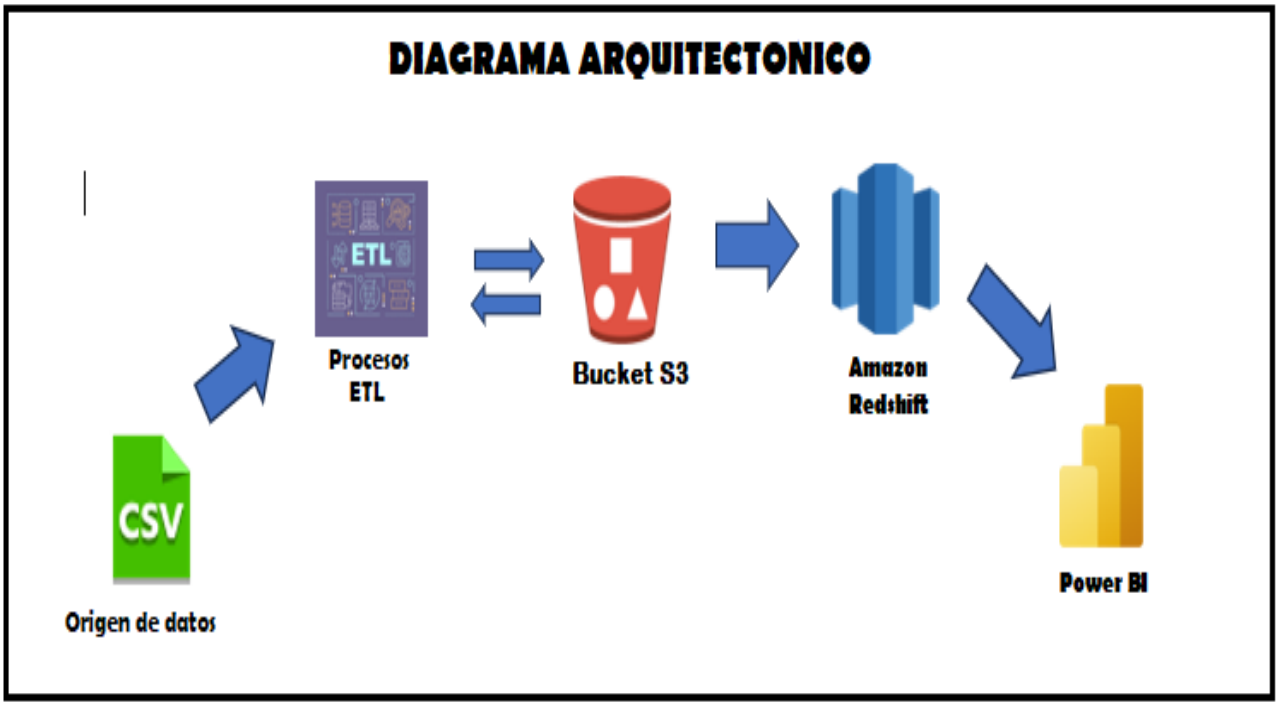

Imagen 10: Diagrama Arquitectónico de la solución

e. Descripción de cada componente de la solución

- **Origen de datos:** La empresa IBW Comunicaciones ha proporcionado archivos CSV que encapsulan información transaccional vital. Estos archivos, caracterizados por su estructura consistente, contienen datos relevantes sobre ventas y clientes, etc, actuando como un catalizador esencial para la planificación y diseño de los procesos ETL. Esta aportación no solo agiliza la creación del modelo dimensional, identificando dimensiones y tabla de hechos clave, sino que también juega un papel estratégico al garantizar la calidad y relevancia de los datos almacenados en nuestro proyecto de Data Warehouse.

- **Procesos ETL:** Con utilización de Talend Studio permitió no solo la extracción eficiente de datos, sino también la aplicación de procesos de transformación necesarios para adaptar la información a los requisitos específicos del modelo dimensional, Una vez que la información fue tratada adecuadamente, el siguiente paso crucial fue la carga de los datos en Amazon S3. Este proceso englobó las operaciones de Extracción, Transformación y Carga (ETL), garantizando que los datos procesados y enriquecidos estuvieran disponibles en un entorno de almacenamiento seguro y altamente escalable.

- **Bucket S3:** Este componen es el encargado de proporcionar una estructura organizativa y lógica para el almacenamiento de datos en la nube, es utilizado continuamente durante todas las fases del proceso de extracción, transformación y carga de los datos por medio de los procesos ETL construidos en Talend Studio. Los datos extraídos son llevados a la zona Raw tal y como fueron proporcionados en archivos CSV por IBW, al realizar algún proceso de transformación de los datos, estos son cargados a una zona intermedia llamada Staging y finalmente, los datos se adecuan al modelo dimensional propuesto y son llevados a la zona Presentation del lago de datos.

- **Amazon Redshift:** Este componente desempeña un papel crucial al servir como el alojamiento principal para los datos provenientes de la zona de Presentación del Bucket de S3. Su función es esencial para la estructura y organización de la información, ya que actúa como un depósito centralizado que almacena los datos conforme a la arquitectura definida en el modelado dimensional previamente construido. Este modelado no solo refleja la complejidad y la interrelación de las métricas definidas, sino que también proporciona una estructura lógica y optimizada que facilita la consulta y el análisis de datos.

- **Power BI:** Este componente es utilizado como una herramienta para realizar inteligencia de negocios a través de los datos alojados en Amazon Redshift, los cuales fueron mostrados de forma interactiva para los usuarios y de esa manera sea utilizado para la toma de decisiones.

Capitulo III: Estrategia de implementación de la propuesta de solución

a. Estrategia de implementación

Tras realizar un exhaustivo análisis, hemos identificado la solución óptima para la implementación del proyecto, satisfaciendo las necesidades específicas de la empresa IBW. El enfoque se centra en abordar su imperiosa necesidad de analizar volúmenes significativos de datos, presentados en forma de informes, con el fin de facilitar la toma de decisiones a nivel ejecutivo y gerencial.

En consonancia con las normativas y factores de seguridad establecidos por la empresa, se procederá a la extracción de información de la base transaccional, la cual estará en formato CSV. Este formato se ha seleccionado considerando las directrices previamente delimitadas. La información extraída constituirá el insumo principal para el desarrollo del Data Warehouse.

La solución propuesta se estructurará en una serie de módulos interrelacionados, cada uno diseñado para desempeñar un papel específico en el proceso integral. Estos módulos trabajarán de manera colaborativa, abordando eficazmente las etapas de extracción, transformación, carga y presentación de los datos. El objetivo final es ofrecer una representación clara y efectiva de la información, alineada con los requisitos y objetivos estratégicos de la alta gerencia de IBW.

Sistema transaccional y base de datos

La empresa IBW maneja un conjunto de sistemas y bases de datos, en nuestro caso nos centráremos en los sistemas transaccionales que manejan la parte operativa de las ventas;

Imagen 11: Logo de ERPnext

El Sistema el ERPNext: ERPNext es un sistema de planificación de recursos empresariales (ERP) de código abierto y basado en la nube. Desarrollado en Python y basado en el framework web Frappe, ERPNext ofrece una suite integral de aplicaciones empresariales que cubren diversas áreas funcionales, permitiendo a las organizaciones gestionar eficientemente sus procesos internos.

Base de datos MariaDB y DBeaver:

Imagen 12: Logo de base de datos MariaDB

MariaDB es un sistema de gestión de bases de datos (DBMS) de código abierto que se desarrolla como un reemplazo compatible con MySQL. Fue creado por el desarrollador original de MySQL. Utiliza la mayoría de los comandos y APIs de MySQL, lo que facilita la migración de aplicaciones y datos de MySQL a MariaDB sin cambios en el código. Ofrece características avanzadas, mejor rendimiento y mayor seguridad en comparación con versiones antiguas de MySQL. MariaDB se ha vuelto popular en la comunidad de código abierto y se utiliza en una variedad de aplicaciones, desde sitios web pequeños hasta grandes sistemas empresariales.

Imagen 13: Logo de la herramienta DBeaver

DBeaver es una herramienta de administración de bases de datos de código abierto que admite una amplia variedad de sistemas de gestión de bases de datos (DBMS). Proporciona una interfaz gráfica de usuario (GUI) que permite a los desarrolladores y administradores de bases de datos interactuar con sus bases de datos de manera eficiente.

Estructura de Carpetas en local

La estructura de carpetas en el sistema operativo local (SO Windows) es necesaria para almacenar archivos de datos, para dicha implementación es necesario crear las siguientes carpetas en la ruta indiada a continuación.

C:\Users\NombreDelSistema\Desktop\IBW-Proyecto

- IBW-Proyecto
  - Raw/
  - Staggin/
  - Presentation-access/

Dentro de la estructura se almacenarán los datos para los procesos ETL que usara la herramienta Talend Open Studio.

Amazon Web Services S3

Se decidió emplear el servicio de almacenamiento S3 (Simple Storage Service) de Amazon Web Services (AWS) para la integración de servicios en la nube. Esta elección se basó en la seguridad y disponibilidad que ofrece, así como en su atractivo desde el punto de vista económico, ya que su uso es gratuito hasta cierto límite dentro de la capa gratuita de AWS.

Para crear un bucket en S3, se deben seguir una serie de pasos como los siguientes:

1. Inicie sesión en el panel de administración de AWS y abra la consola de Amazon S3 en https://console.aws.amazon.com/s3/.

2. Seleccione la opción "Create bucket" (Crear bucket), lo que abrirá el asistente para la creación de buckets.

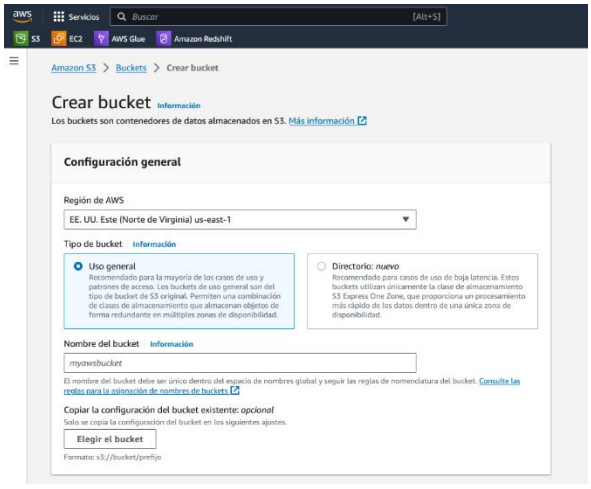

Imagen 14: Entorno de creación de Bucket en S3

3. En "Bucket name" (Nombre del bucket), ingrese un nombre compatible con DNS para el bucket. Para nuestro caso se elegio el siguiente nombre: “bucket-grupo2”

4. En "Region de AWS", elija la región de AWS en la que desea que se ubique el bucket. La cual pone por defecto -> EE. UU. Este (Norte de Virginia) us-east-1

5. En "Object Ownership" (Propiedad de objetos), puede desactivar o habilitar las ACL para controlar la propiedad de los objetos cargados en el bucket.

6. Finalmente, seleccione la opción "Create bucket" (Crear bucket).

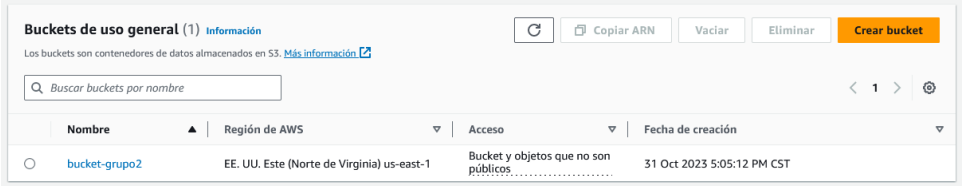

Imagen15: Información del bucket creado

Una vez creado el bucket, se requiere la siguiente estructura de carpetas dentro del bucket S3, con el nombre de bucket como raíz:

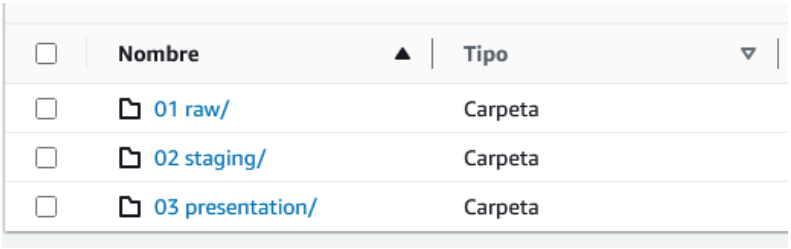

Imagen 16: Estructura básica de carpetas en S3

Esta estructura permitirá aprovechar la solución propuesta de manera efectiva.

Amazon Redshift

Una parte fundamental del proceso de implementación es la creación de un Clouster en AWS Redshift, el cual soportara la estructura del Data Werehouse para su futura carga de datos, como paso previo es necesario que el Bucket en S3 ya esté en funcionamiento, iniciamos la creación del Clouster accediendo a la consola de AWS, en la cual ingresaremos a Amazon Redshift.

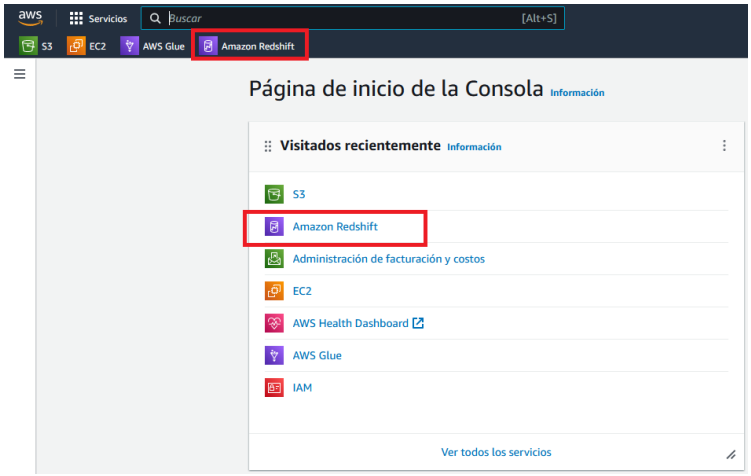

Imagen 17: Panel de AWS, ingresar a Redshift.

Una vez cargue el apartado de Redshift, se debe asegurar que la región seleccionada sea la adecuada, esto para no incurrir en costos y utilizar la capa gratuita por 2 meses, luego dará clic en el botón “Crear Clúster”, a continuación, se cargara la interfaz de creación donde se colocara los siguientes parámetros; ver imagen siguiente.

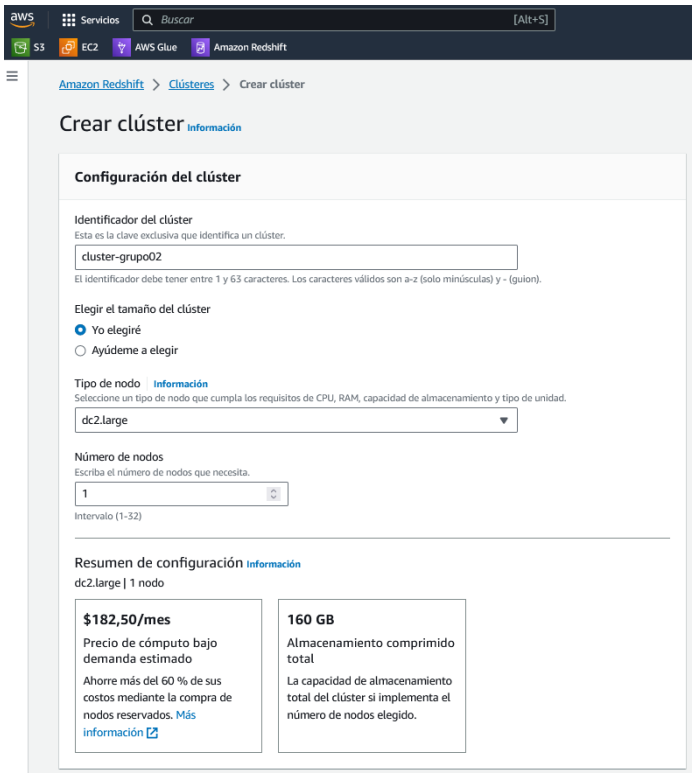

Imagen18: Configuración básica del clúster (capa gratuita)

Para la consecución de los siguientes apartados, se requiere la configuración detallada de diversos elementos. En primer lugar, es imperativo definir los Datos de muestra; en este contexto, optaremos por no seleccionar ninguna muestra, ya que no precisamos información de prueba en este momento. Seguidamente, nos sumergiremos en la Configuración de la base de datos, donde nos ubicaremos en la sección correspondiente al nombre de usuario del administrador, que, por defecto, Amazon designa como "awsuser". Aunque es posible modificarlo, en esta instancia lo conservaremos en su configuración por defecto. En cuanto a la contraseña de administrador, por razones de seguridad, optaremos por crear manualmente una contraseña robusta, seleccionando la opción "Añadir manualmente la contraseña del administrador".

En el apartado de "Permisos del clúster", se nos presentan dos alternativas: la elección de un rol IAM preexistente o la utilización del rol predeterminado. Cabe destacar que esta elección puede ser modificada después de la creación del clúster. En esta ocasión, se procederá a la creación rápida de un nuevo usuario, transfiriendo el rol predeterminado al usuario recién creado. Finalmente, se dejarán activadas las configuraciones adicionales antes de presionar el botón "Crear Clúster". Es importante señalar que este proceso demandará algunos minutos, ya que la creación no es instantánea. Al completarse, la

interfaz deberá reflejar un resultado similar al presentado en la siguiente imagen. Este procedimiento es esencial para establecer la base necesaria para los siguientes análisis y desarrollos en el marco de la tesis.

Imagen 19: Clúster creado y habilitado.

En la fase final, procederemos a ingresar al clúster que acabamos de crear. Una vez dentro, nos dirigiremos hacia el botón ubicado en la esquina superior derecha, identificado como "Datos de consulta". Este botón nos presentará dos opciones, ambas destinadas al mismo propósito. Optaremos por acceder a la función "Consultar en el editor de consulta v2".

Esta acción nos permitirá ingresar al editor de consultas avanzado, proporcionando un entorno donde podremos ejecutar consultas de manera eficiente. Este paso es crucial, ya que nos brindará las herramientas necesarias para analizar y extraer datos relevantes del clúster creado. El acceso al "Editor de consulta v2" representa un punto fundamental en nuestro proyecto, ya que nos facilitará la interacción y manipulación de los datos almacenados en el clúster, sentando las bases para la posterior realización de análisis detallados.

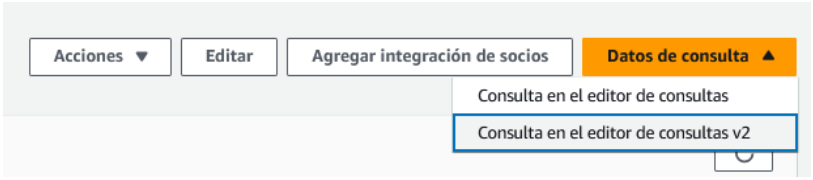

Imagen 20: Consultar en el editor de datos v1 y v2.

AWS Es una herramienta que nos proporciona un entorno completo para el desarrollo de la solución propuesta, ejemplo de ello es el editor que nos da Redshift para hacer consultas, para culminar solo falta ingresar el usuario y la clave que se crearon al inicio. Esto nos desplegara el esquema y una hoja para ingresar consultas sql, en nuestro caso solo cargaremos el script desarrollado donde está el modelo estrella, el cual es el resultado del análisis de los datos.

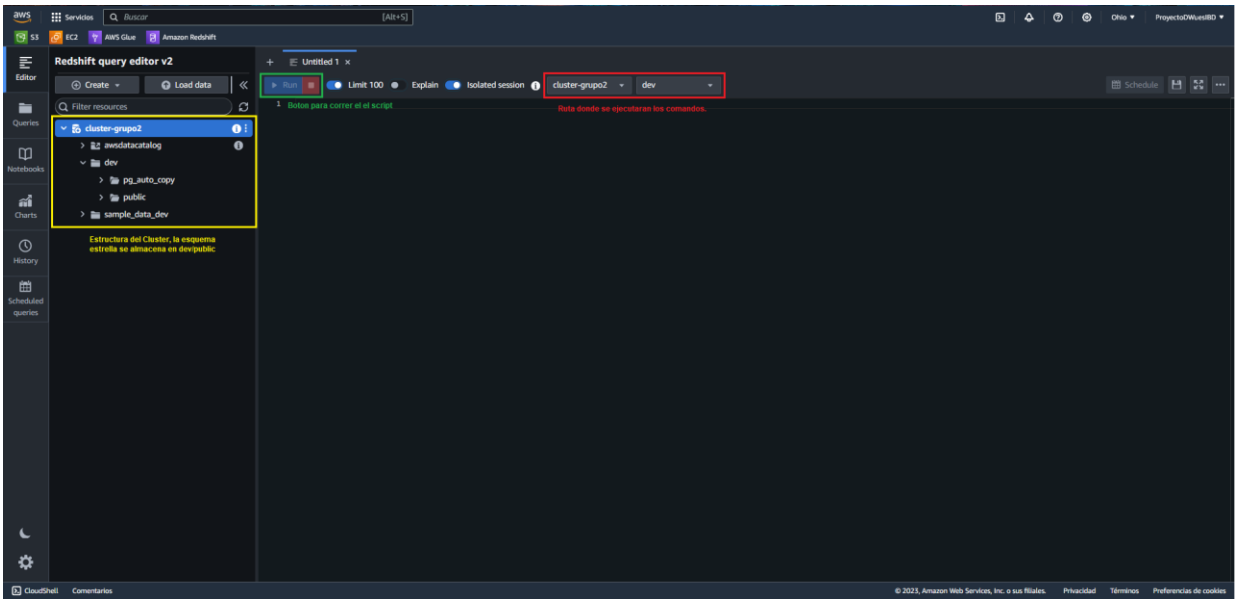

Imagen 21: Editor de consultas v2

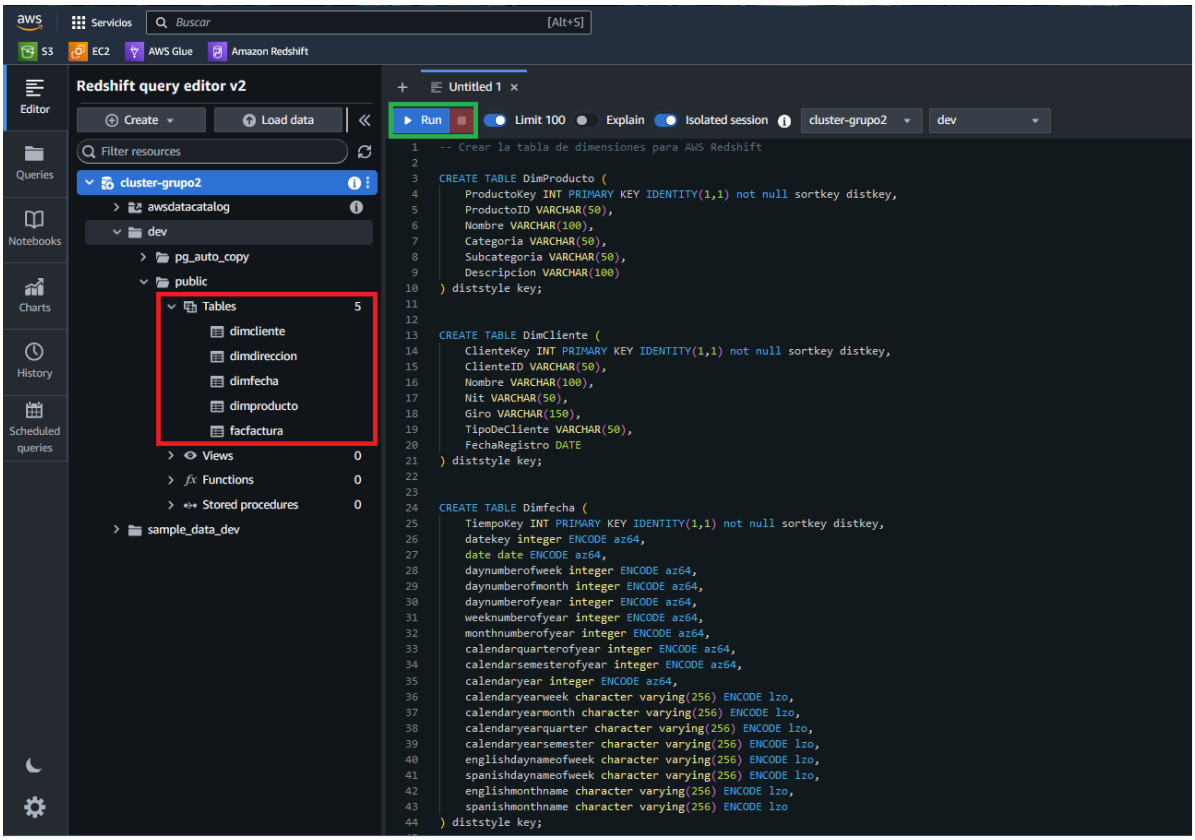

Imagen 22: Esquema estrella (ver anexo1 para consultar script)

Nota: es importante que el cluster tenga los parámetros de acceso abiertos para la conexión a Power Bi, esto se hace otorgándole los permisos necesarios al rol IAM y habilitando la configuración de accesibilidad publica, al ingresar al clúster creado en el apartado Acciones.

Usuario IAM de AWS

Después de haber configurado el Bucket con su correspondiente estructura de carpetas, el siguiente paso crítico implica la creación de un usuario IAM (Identity and Access Management) y la asignación de los permisos necesarios para acceder al bucket en cuestión. En situaciones donde tanto el usuario IAM como el bucket de S3 son parte de la misma cuenta de AWS (Amazon Web Services), es posible conceder al usuario acceso a una carpeta específica dentro del bucket mediante la implementación de una política IAM.

Cuando se establece la relación entre el usuario de IAM y el bucket de S3 dentro de la misma cuenta de AWS, se tiene la flexibilidad de otorgar acceso a carpetas específicas mediante políticas de IAM sin necesidad de actualizar explícitamente la política de bucket. En otras palabras, siempre y cuando la política de bucket no prohíba de manera explícita el acceso del usuario a la carpeta en cuestión, no será necesario realizar modificaciones en la política de bucket si la política IAM ya concede los permisos necesarios. Este enfoque simplifica la gestión de permisos y garantiza un flujo eficiente de acceso a los recursos de S3.

Es importante destacar que la aplicación de la política IAM puede llevarse a cabo de manera individual, permitiendo la adición de dicha política a usuarios de IAM específicos. Alternativamente, para una administración más centralizada, es posible asociar la política IAM a un rol de IAM, lo que permite que varios usuarios tengan la capacidad de cambiar entre roles según sea necesario. Esta flexibilidad en la asignación de políticas IAM brinda una mayor adaptabilidad a las necesidades específicas de acceso dentro de un entorno AWS, facilitando la gestión de la seguridad y simplificando la administración de permisos a lo largo del tiempo.

Procesos ETL en Talend Open Studio 8

Talend Open Studio es una suite de software de integración de datos de código abierto que permite a las organizaciones conectar, acceder y gestionar datos de diversas fuentes y formatos. Se utiliza comúnmente en proyectos de integración de datos, migración de datos, transformación de datos y carga (ETL), así como en la creación de data warehouses, entre otras aplicaciones relacionadas con la gestión de datos.

Para la implementación del proyecto desarrollado en Talend es necesario preparar el entorno de trabajo cambiando la ruta del workspace que trae por defecto la herramienta, en nuestro caso se trasladara a la carpeta ETL/ del repositorio clonado de github “DWPROJEX-GRUP02”, desde la opcion “Manage Connections” esto creara las siguientes carpetas:

.Java/

.JETEmitters/

.metadata/

Temp/

Una vez trasladado el workspace, seleccionaremos el proyecto desde la opcion “Import an existing project”, en la ventana emergente nos desplazaremos a la parte inferior y precionaremos el boton “Import several projects”, con la opcion selecionada de Select root directory, daremos clieck en Browse... y buscaremos la ruta del proyecto \DWPROJEX-GRUP02\ETL\IBWIMPORTARDATOSAWS una vez cargada la ruta eliminaremos el cheque de la casilla “Copy projects into workspace”, para concluir daremos click en “Finish”

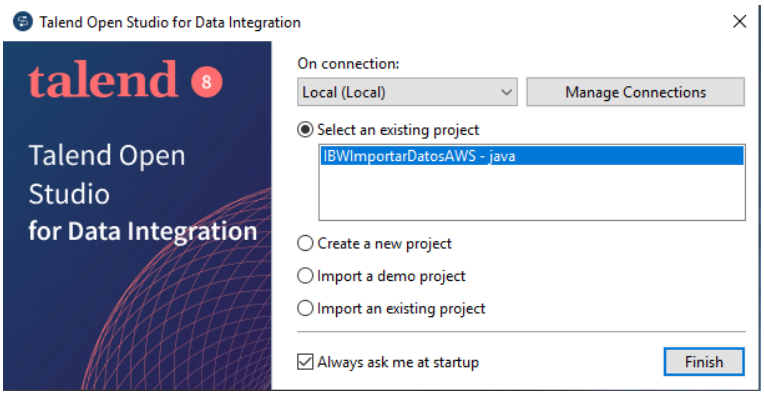

Imagen 23: Proyecto listo para ejecutar.

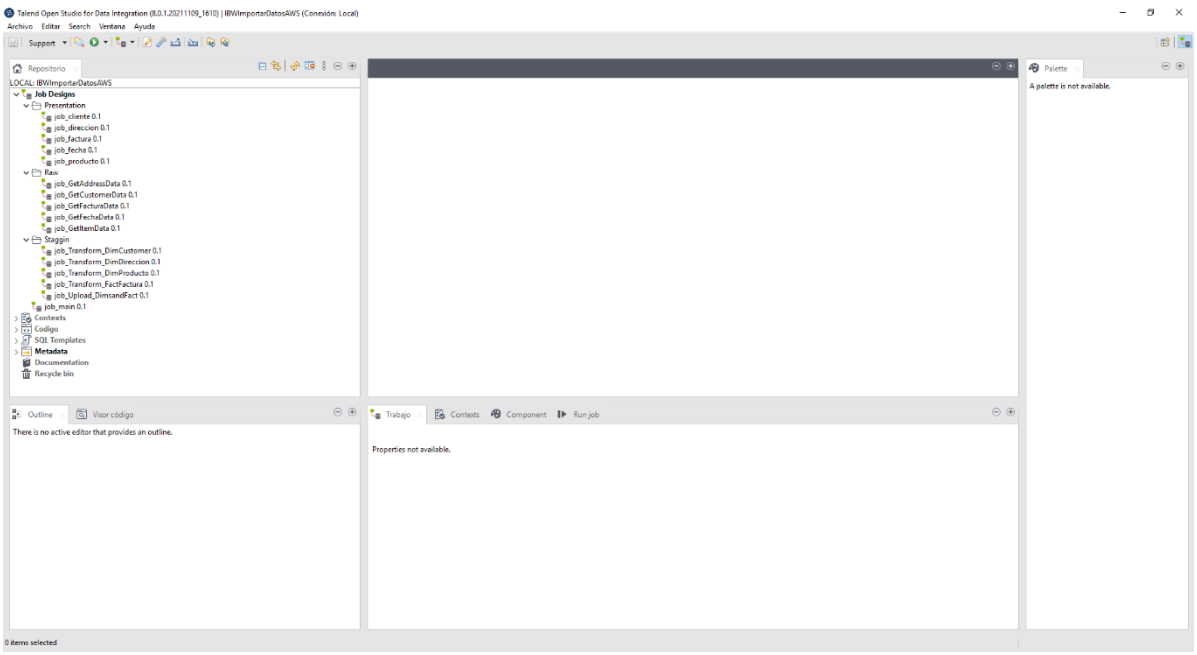

Imagen 24: Pantalla principal de Talend para el proyecto IBW

Antes de correr el proyecto se debe de configurar los parámetros de rutas para los archivos delimitados en el apartado Metadata. Luego seleccionado cada archivo y dando clic derecho en el apartado Edit file delimited.

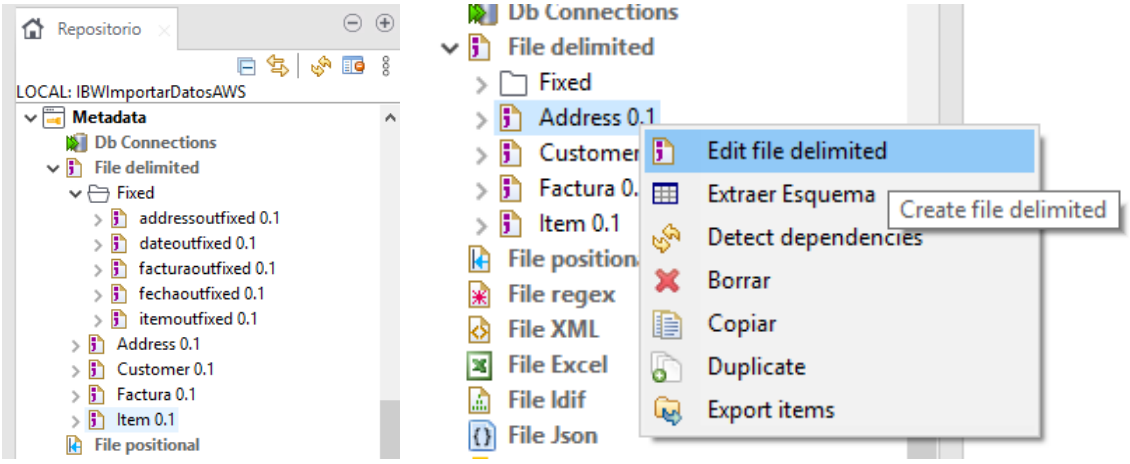

Imagen 25: Configuración de ruta de archivos delimitados

Una vez se despliegue la venta de parámetros presionaremos siguiente, a continuación, añadiremos el nombre de la computadora en **NombrePC** en las rutas destino C:/Users/**NombrePC**/Desktop/IBW-Proyecto/.... por último, seleccionaremos en el botón finalizar. Talend nos preguntara si queremos propagar los cambios realizados, le diremos que si, para que las rutas sean seteadas en los archivos de los diferentes job, para los archivos delimitados .csv que se encuentran en Raw, Stagging y Presentation.

Esto se deberá cambiar en los archivos tS3Get, tFileDelete de Raw así como los de tFileOutputDelimited de Staggin este paso es importante dado que estas rutas son las de salida y eliminación de los archivos preexistentes para mantener los más nuevos.

Power BI

Power BI es una herramienta de análisis de datos desarrollada por Microsoft que permite a los usuarios visualizar y compartir información de manera intuitiva y efectiva. Se utiliza para transformar datos brutos en informes interactivos y paneles de control visualmente atractivos, tomando como base esta breve introducción explicaremos los pasos necesarios para la implementación de la solución desarrollada con la herramienta.

Para acceder al informe en Power BI, es imperativo seguir un procedimiento específico. En primer lugar, se debe seleccionar la opción "Abrir informe" y hacer clic en el botón "Examinar informes". Este paso conducirá a la exploración de archivos, donde se deberá navegar hasta localizar el archivo con la extensión .pbix que corresponda a la solución elaborada. La identificación precisa y selección de este archivo es esencial para asegurar la correcta visualización y utilización del informe en Power BI.

Una vez que el informe se ha abierto, es posible que se requiera actualizar el origen de los datos para reflejar cambios o incorporar nueva información. Para llevar a cabo esta acción, se debe dirigir al menú inicio de Power BI, luego seleccionar "Transformar datos" y finalmente acceder a la opción "Configuración de origen de datos". Este proceso permitirá ajustar y sincronizar el origen de los datos de manera eficiente, asegurando la coherencia y la integridad de la información presentada en el informe.

Estas operaciones son fundamentales para garantizar la óptima utilización y actualización de los informes en Power BI, proporcionando a los usuarios la capacidad de interactuar con datos precisos y actualizados en el contexto de sus análisis y toma de decisiones.

b. Presupuesto de implementación

Recurso tecnológico

La configuración de la solución se ha diseñado estratégicamente al optar por herramientas de software de código abierto, lo que permite un despliegue económico y eficiente desde el punto de vista técnico. En una decisión que conjuga eficacia y economía, se ha seleccionado software de libre disponibilidad, minimizando así los costos asociados al desarrollo e implementación.

Adicionalmente, la elección de esta infraestructura se alinea con la maximización de recursos existentes en la organización. Se aprovechan licencias ya existentes para los sistemas operativos y herramientas, lo cual no solo reduce los gastos asociados a adquisiciones adicionales, sino que también optimiza la compatibilidad entre la solución implementada y la infraestructura tecnológica preexistente.

Es fundamental destacar que esta estrategia se ampara en la existencia de hardware adecuado ya disponible, eliminando la necesidad de inversiones adicionales en equipos. Por ende, se evitan desembolsos innecesarios, asegurando la viabilidad económica del proyecto. En resumen, la implementación de la solución se ha llevado a cabo de manera eficiente, aprovechando al máximo las herramientas disponibles sin incurrir en gastos tecnológicos adicionales.

Recurso humano

Bajo lo estipulado por la empresa IBW la implementación del proyecto lo llevara a cabo una sola persona del grupo de desarrollo, en un periodo no mayor a una semana, esto al ser un proyecto presentado por el equipo de tesis no generara un costo a la entidad.

Servicios básicos

La entidad IBW como una de las empresas líder en su mercado, dispone de los servicios básicos, así como una conectividad a Internet superior a muchas otras, lo cual constituye un recurso fundamental para la implementación del proyecto en cuestión. En términos de servicios, no se generan costos adicionales para llevar a cabo la implementación de la solución, dado que los elementos básicos requeridos están ya contemplados y disponibles sin implicar desembolsos económicos. Este enfoque estratégico contribuye a una gestión eficaz de los recursos, asegurando la viabilidad económica del proyecto desde su fase inicial.

c. Análisis de resultados

Carga de Datos

Se subieron los archivos CSV, brindados por la empresa IBW que contienen la información de las tablas transaccionales que nos ayudaran a crear el data warehouse, al bucket llamado **bucket-grupo2**, en la carpeta 01 raw

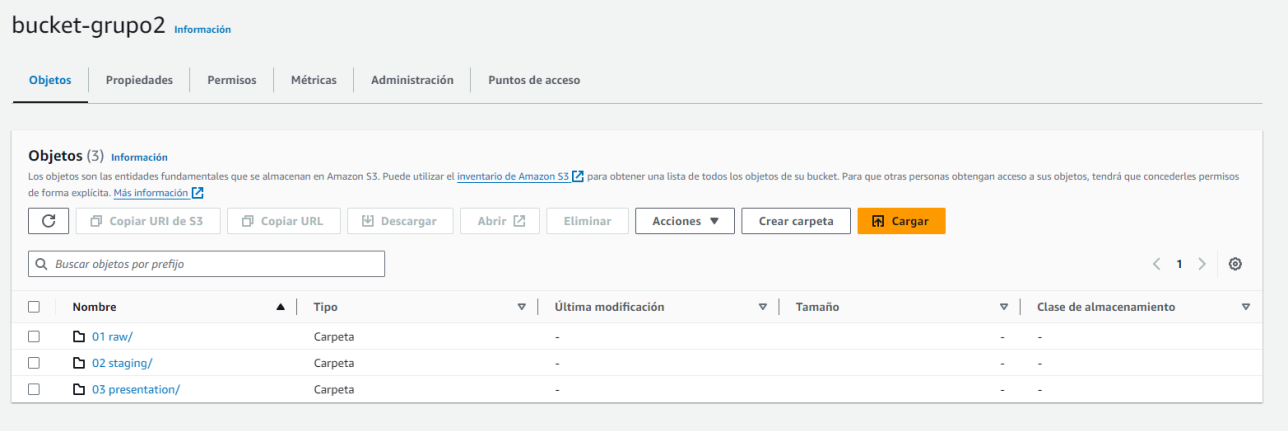

Imagen 26: Estructura de carpetas creada en bucket-grupo2

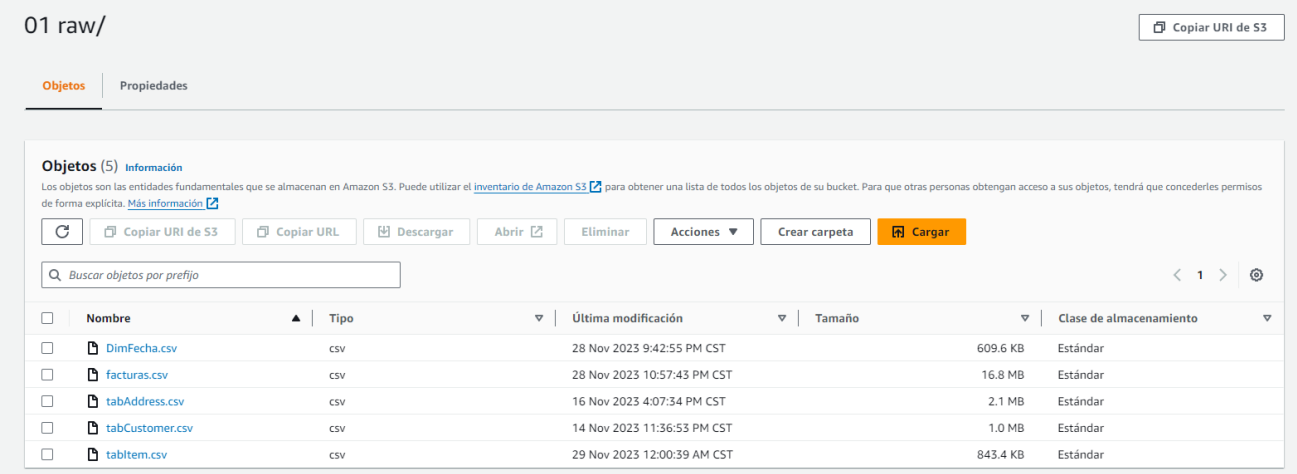

Imagen 27: Carpeta 01 raw en bucket-grupo2

Usando la herramienta talend se creó una carpeta llamada Raw, para poder obtener los archivos csv que compartió la empresa que se subieron a S3 de AWS.

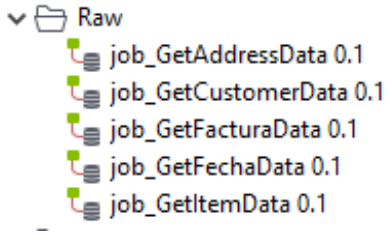

Imagen 28: Carpeta Raw creada en talend

Se crearon los diferentes Jobs para poder obtener los csv que se cargaron previamente en S3 en la carpeta raw, el job job_GetAddressData 0.1 ,que es el job que trae el archivo csv de la dirección, se creo el elemento tFileDelete_1, que borra un archivo csv en la maquina local , se hace esto con la idea de que se borre un archivo si ya existe para que no de error de que ya existe un archivo csv con ese nombre.

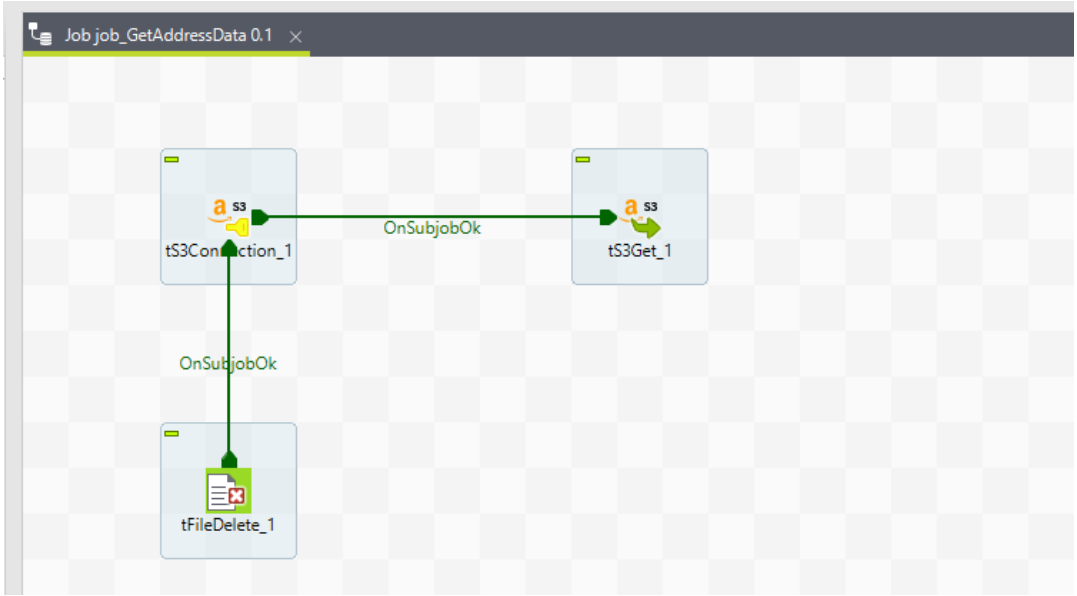

Imagen 29: Job que obtiene el archivo de dirección

Posteriormente se tiene el elemento tS3Connection, que se encarga de hacer la conexión a nuestro bucket en AWS y poder tener acceso a los archivos cargados previamente, se tiene que colocar el Access Key y Secret Key que los genera AWS al crear el bucket.

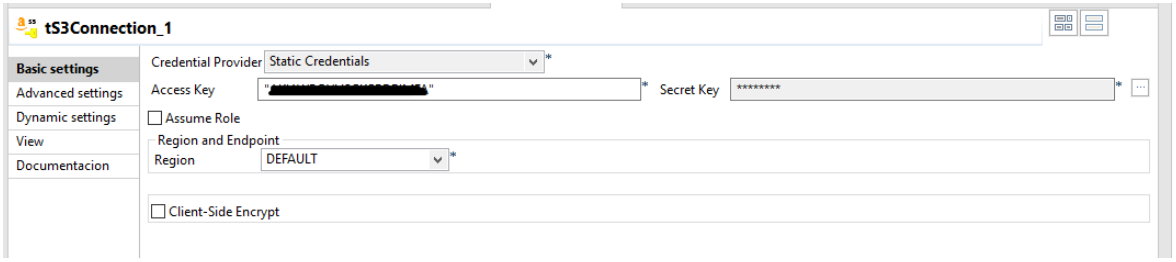

Imagen 30: Conexión al bucket de AWS

Y finalmente tenemos el elemento tS3Get_1, que es el que extrae el archivo csv del bucket y lo descarga en la computadora local para posteriormente poder transformar los datos. Se tiene que especificar el nombre del bucket creado en AWS en el campo de bucket en el campo key se debe especificar la ruta en donde se encuentra el archivo csv de dirección y en el campo file se coloca la dirección en donde se guardara el archivo csv de la dirección.

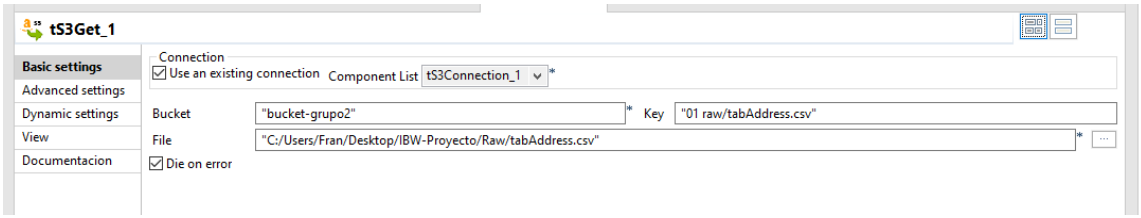

Imagen 31: Componente get que almacena el archivo csv en la computadora local

Transformación de Datos

Para la transformación de datos, se creó una carpeta en talend llamada Staggin.

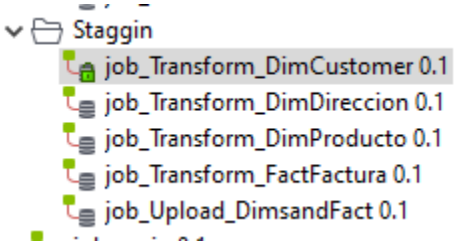

Imagen 32: Creacion de carpeta Staggin en talend

Transformación de archivo csv de dirección en la dimensión DimDireccion

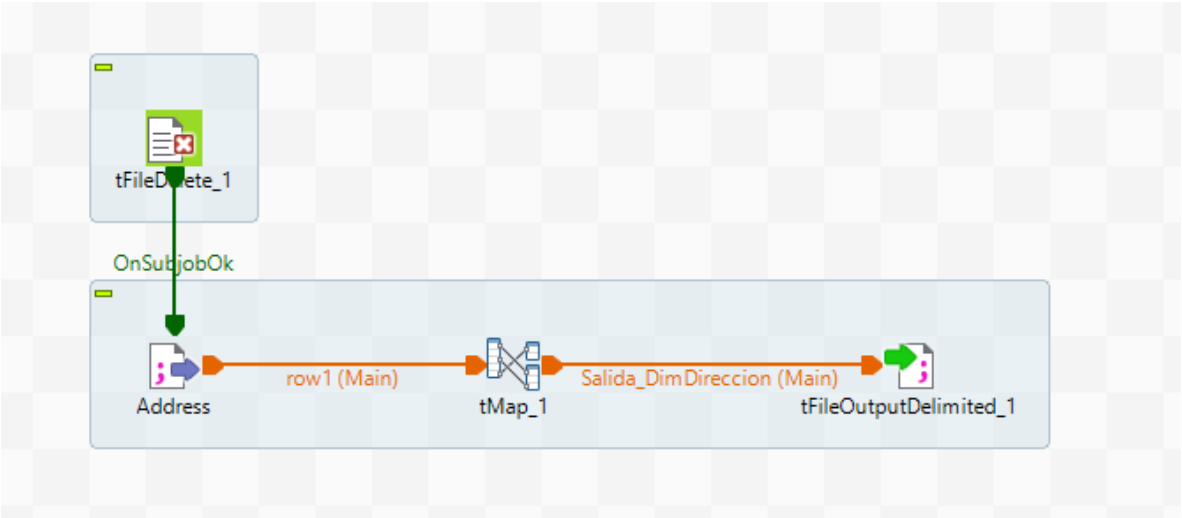

Imagen 33: Transformación y limpieza de dirección

Transformación de archivo csv de cliente en la dimensión DimCliente

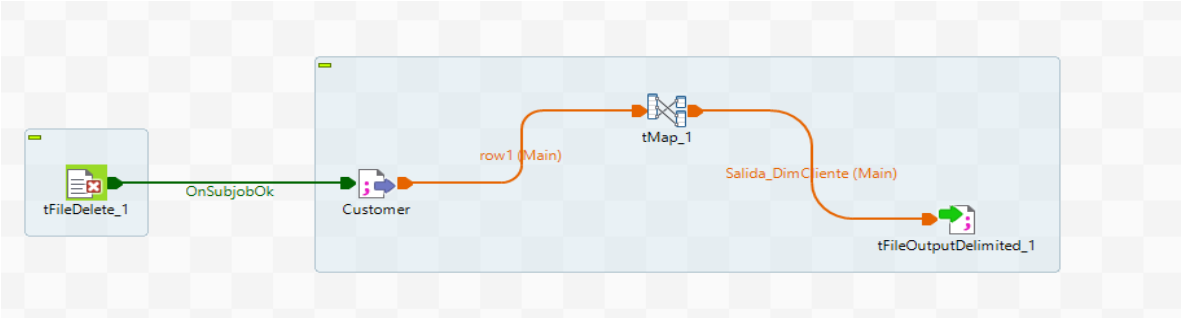

Imagen 34: Transformación y limpieza de cliente

Transformación de archivo csv de producto en la dimensión DimProducto

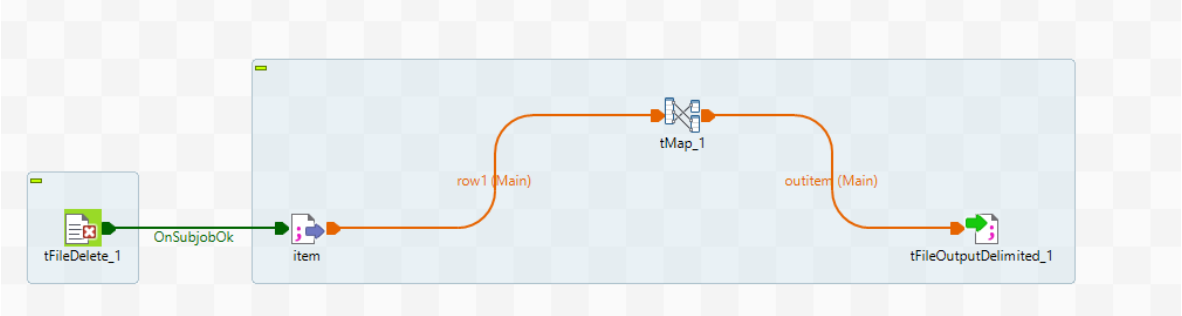

Imagen 35: Transformación y limpieza de producto

Transformación de archivo csv de factura, con las dimensiones de dirección, cliente, producto y fecha en la tabla de hechos FacFactura.

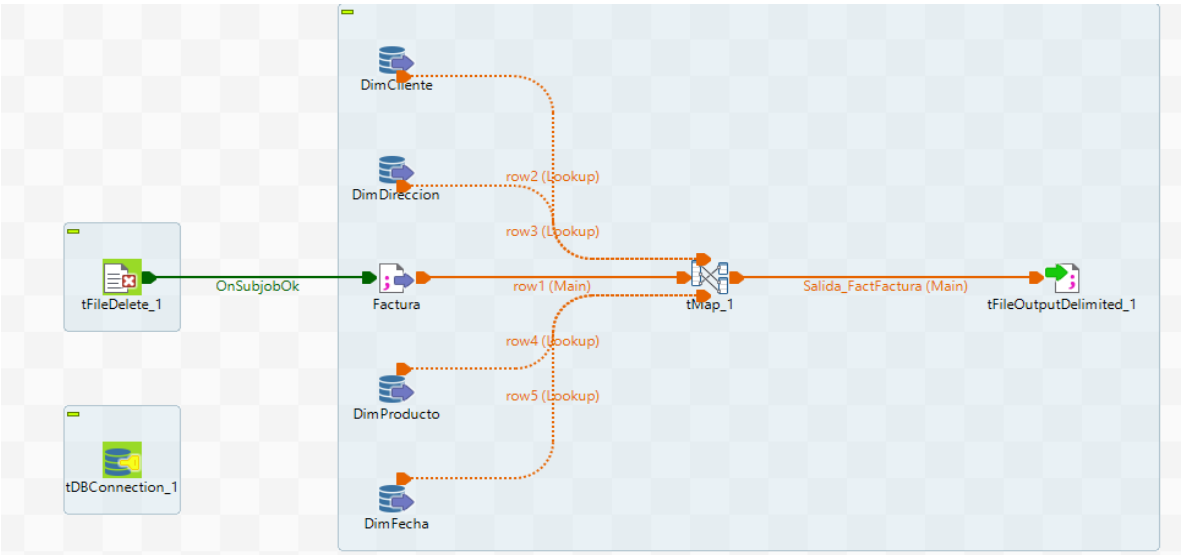

Imagen 36: Transformación y limpieza de factura de venta

Para ejecutar todos los Jobs anteriores se creó un job padre.

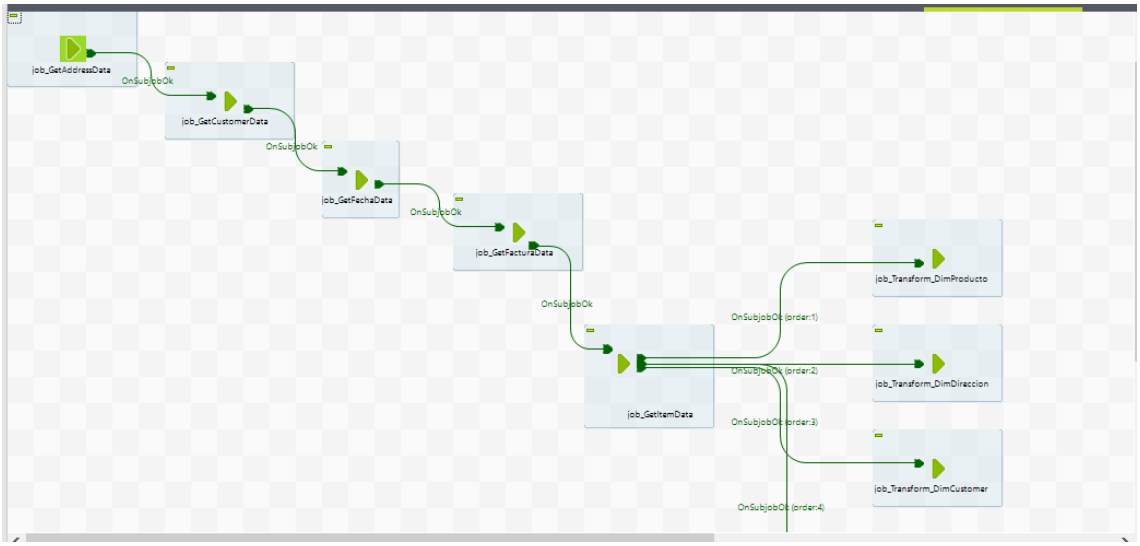

Imagen 37: Job padre en talend

La tabla de hechos y sus respectivas dimensiones, a excepción de la dimensión Fecha, son transformados a partir de los archivos csv almacenados en la carpeta local. Para posterior subirlos a S3 en AWS en la carpeta Staggin.

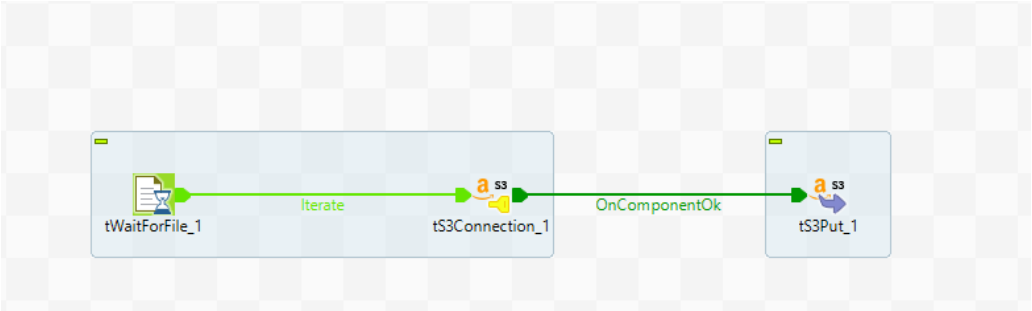

Imagen 38: Carga de archivos hacia el bucket en AWS

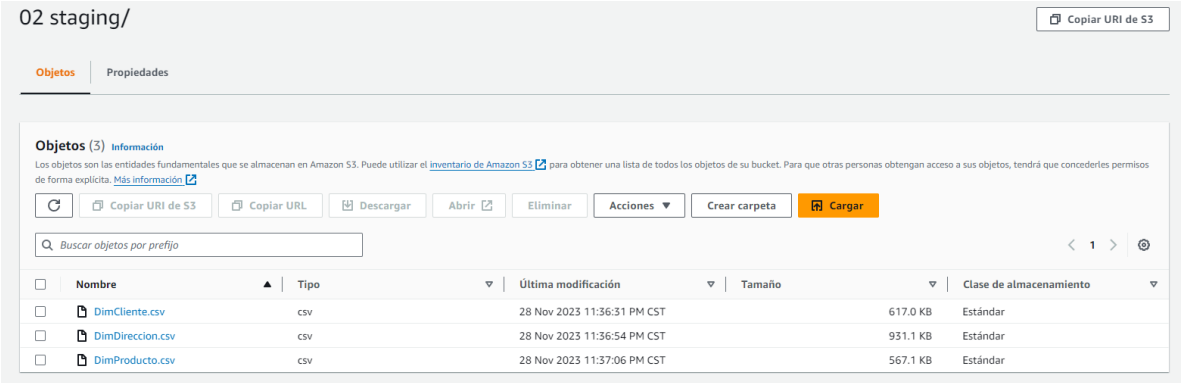

Imagen 39: Carpeta 02 Staging en bucket-grupo2

Carga de datos a Redshift

Después de transformar los datos, se generaron archivos csv en la computadora local y usando la herramienta de talend se creó una carpeta llamada presentación.

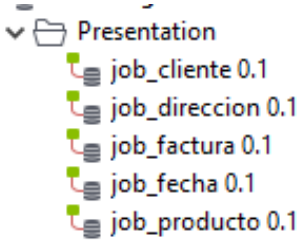

Imagen 40: Carpeta Presentation creada en talend

Carga del archivo csv de la dimensión DimCliente hacia Redshift

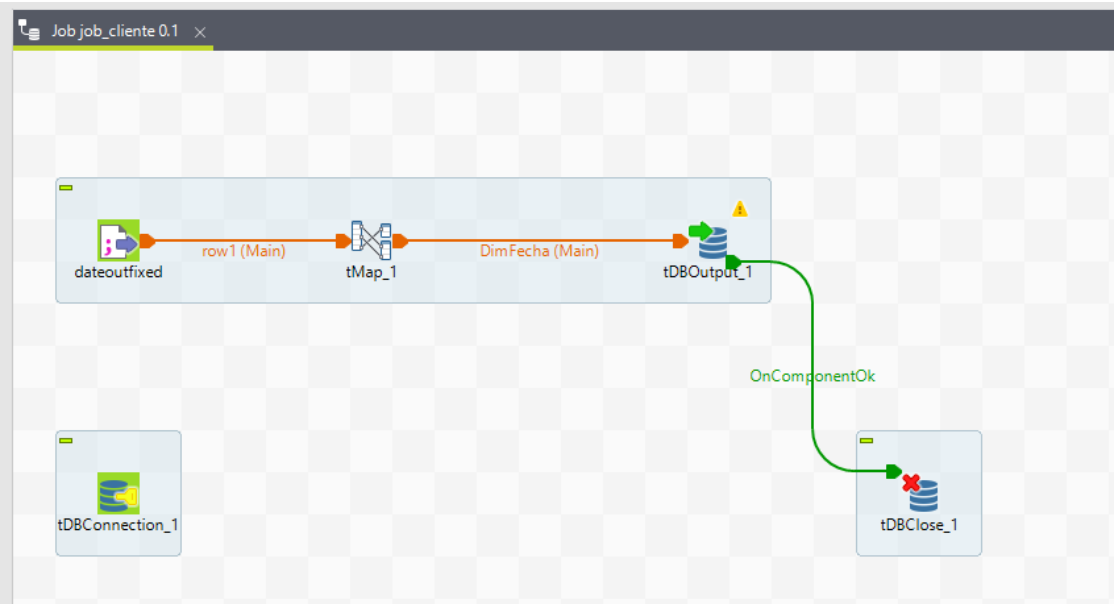

Imagen 41: Carga de dimensión DimCliente hacia Redshift

Carga del archivo csv de la dimensión DimDireccion hacia Redshift

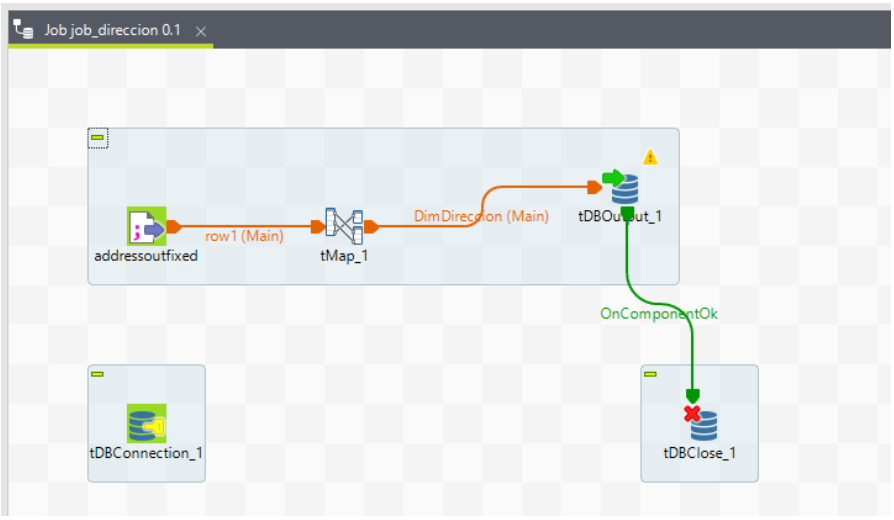

Imagen 42: Carga de dimensión DimDireccion hacia Redshift

Carga del archivo csv de la tabla de hechos FacFactura hacia Redshift

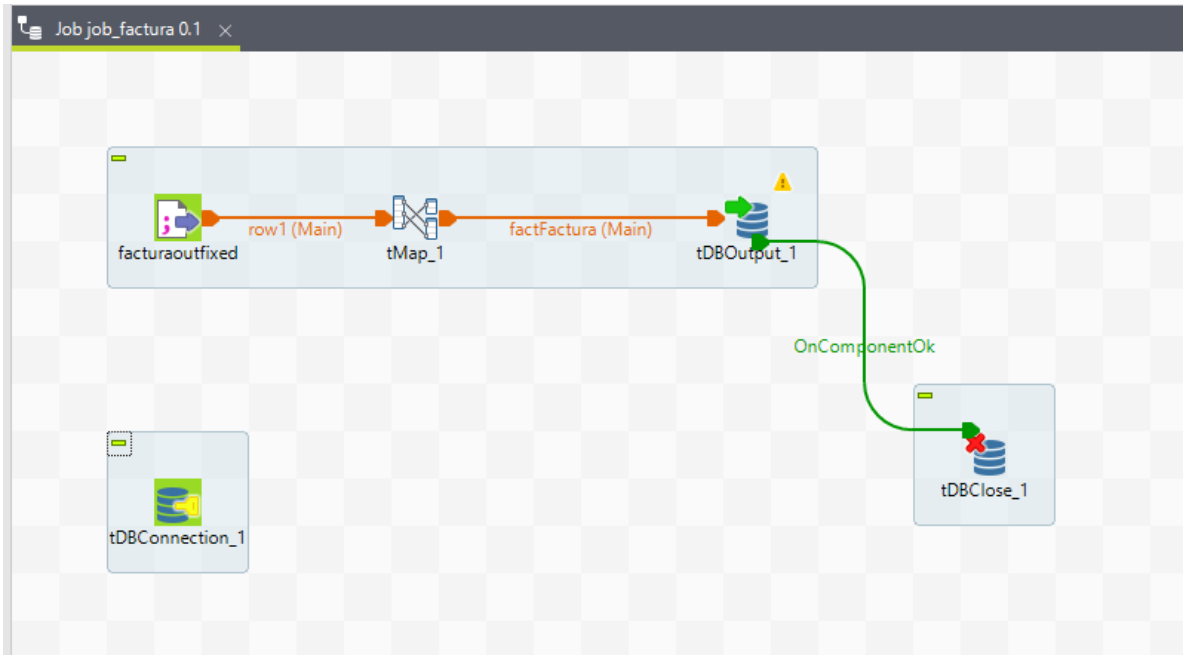

Imagen 43: Carga de tabla de hechos FacFactura hacia Redshift

Carga del archivo csv de la dimensión DimFecha hacia Redshift

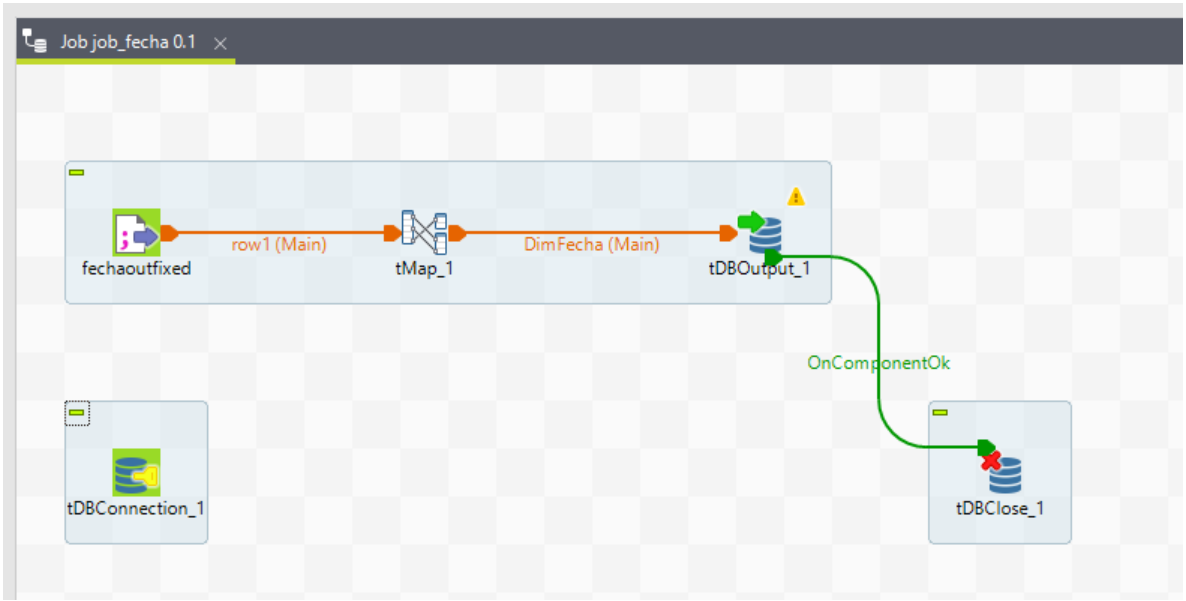

Imagen 44: Carga de dimensión DimFecha hacia Redshift

Carga del archivo csv de la dimensión DimProducto hacia Redshift

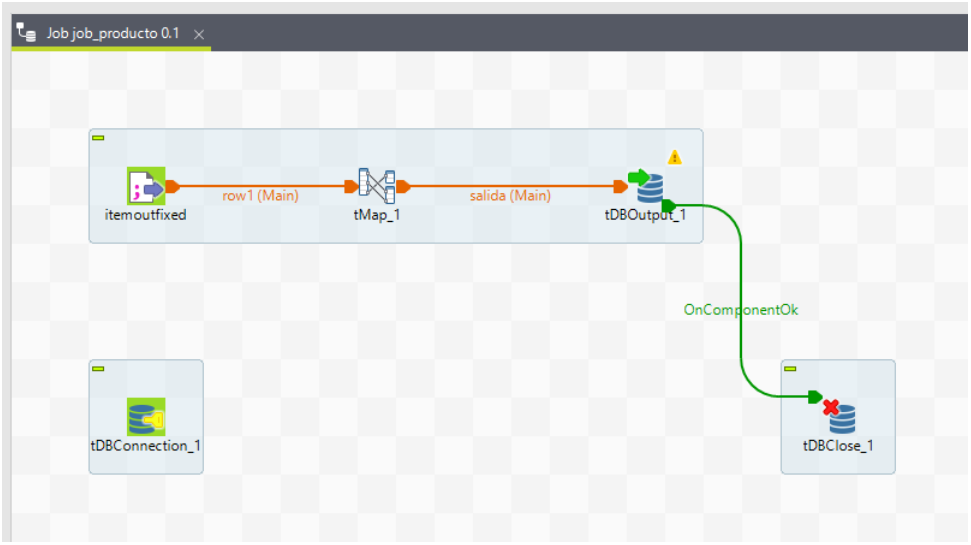

Imagen 45: Carga de dimensión DimProducto hacia Redshift

Carga de datos en Redshift en la base llamada cluster-grupo2, en las dimensiones DimCliente, DimDireccion, DimFecha, DimProducto, y la tabla de hechos FacFactura

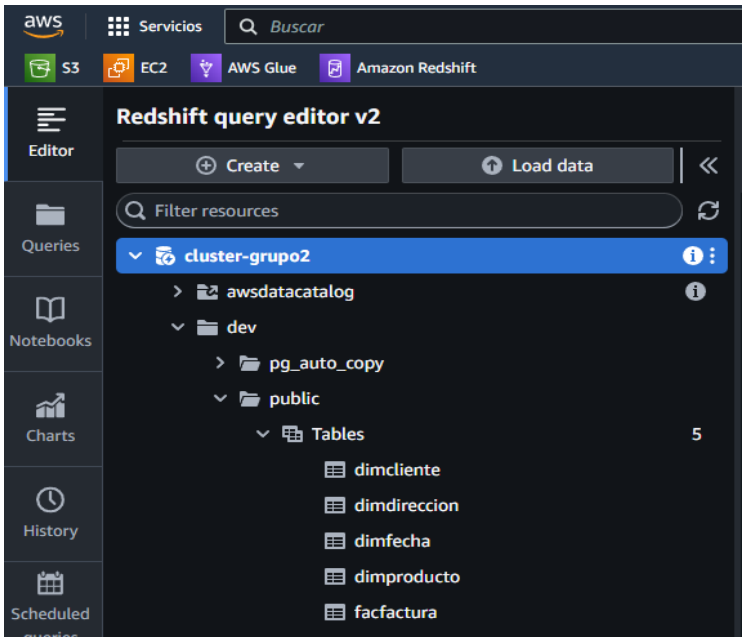

Imagen 46: Dimensión y tabla de hechos en Redshift

Presentación en Power BI

Con la herramienta de Power BI se obtiene los datos de la base cluster-grupo2 que está en redshift y se realizan los siguientes dashboard que cumplen las métricas establecidas antes.

Métrica 1

Imagen 47: Métrica 1 en Power BI

Métrica 2

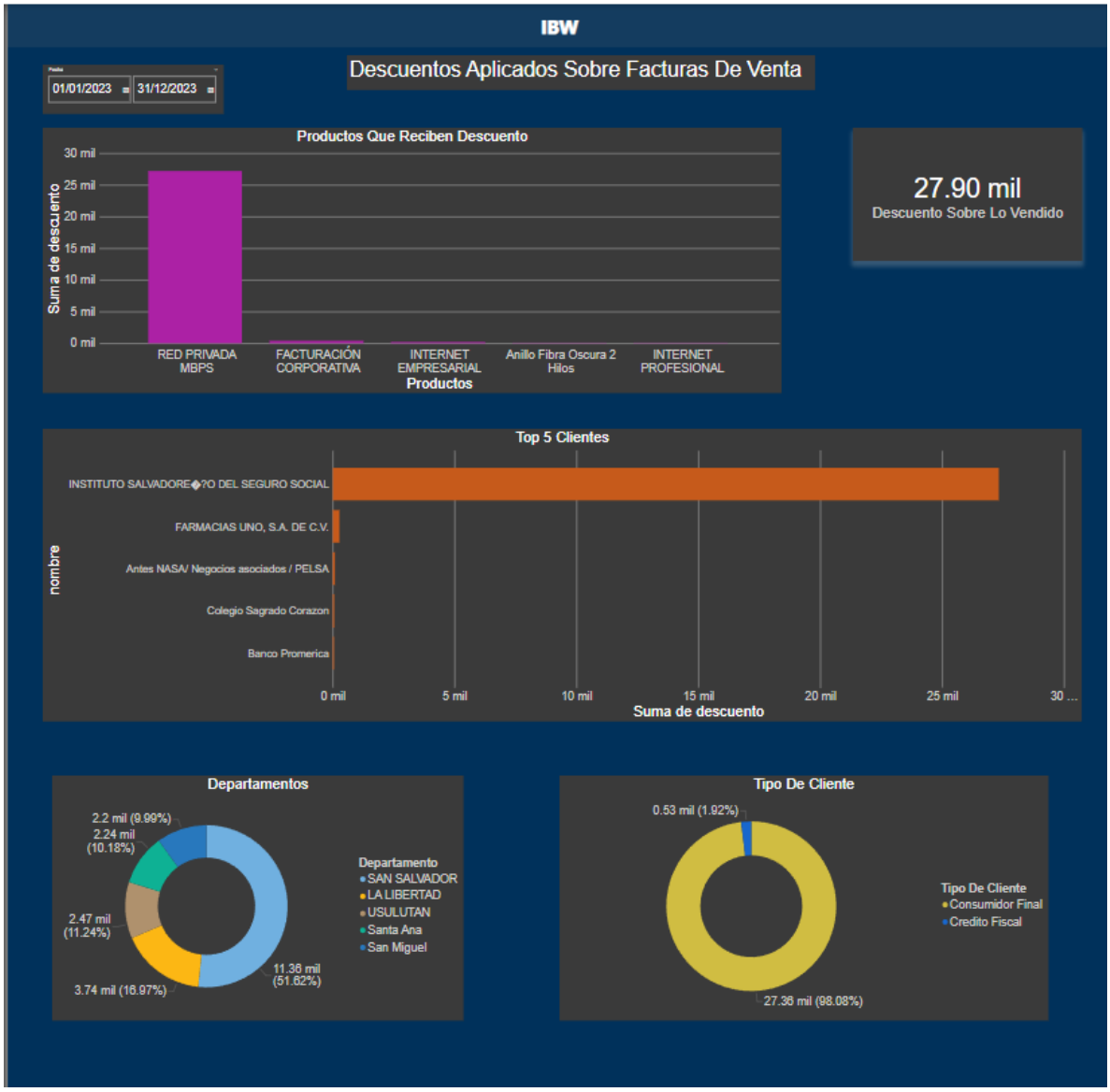

Imagen 48: Métrica 2 en Power BI

Métrica 3

Imagen 49: Métrica 3 en Power BI

Conclusiones y recomendaciones

Conclusiones

- La aplicación práctica de los conocimientos obtenidos durante el curso de especialización en Ingeniería de Datos fue esencial para materializar una solución integral y eficiente. Este proceso comenzó con la creación del modelo dimensional, donde se emplearon las técnicas y mejores prácticas aprendidas para diseñar una estructura que reflejara de manera fiel la complejidad y las interconexiones de los datos en el ámbito empresarial de IBW.

- Se realizó una profunda investigación de la lógica del negocio con el objetivo primordial de obtener un conocimiento profundo sobre el funcionamiento intrínseco de los procesos en el área de ventas de IBW. Como resultado de esta investigación en profundidad, se identificaron y definieron claramente las dimensiones que serán fundamentales en la solución propuesta.

- La capacidad de generar informes en Power BI no solo radica en la presentación visual de datos, sino también en la capacidad de realizar análisis interactivos y personalizados. Estos informes no solo sirven como instantáneas estáticas, sino como herramientas dinámicas que permiten a los usuarios explorar datos, realizar desgloses detallados y mejora en la toma decisiones.

Recomendaciones

- Para la solución brindada Amazon Web Services es una opción altamente beneficiosa gracias a sus destacadas características en seguridad y disponibilidad. No obstante, es importante señalar que, para empresas de menor envergadura, especialmente las pequeñas y medianas, estos servicios pueden conllevar un costo sustancialmente elevado por lo que se sugiere explorar alternativas que se adecuen mejor al presupuesto de dichas empresas, considerando incluso la posibilidad de utilizar recursos informáticos propios.

- Todas las herramientas empleadas en la construcción de la solución demuestran su eficacia, pero en algunas situaciones se torna imperativo ya que no se puede hacer uso de todos los recursos, en ciertas tareas pueden requerir el despliegue de herramientas de pago para alcanzar niveles óptimos de rendimiento y funcionalidad. En consecuencia, la consideración de herramientas de pago se vuelve esencial para aquellas instancias en las que se busca maximizar la eficiencia y obtener resultados más avanzados.

- Se debe definir un intervalo temporal específico para el análisis de los datos obtenidos de la solución. No obstante, la decisión de cuándo realizarlos recae en la discrecionalidad del líder o coordinador designado para la gestión, dado que el tiempo dedicado a estos estudios y análisis dependerá de la naturaleza y enfoque de las actividades desempeñadas por la organización.

Bibliografía

- ✓ ***Kimball, R., & Ross, M. (2013). The data warehouse toolkit (3rd ed.). Wiley.***
- ✓ ***Talend Studio Documentation***
  https://help.talend.com/r/en-US/8.0/open-studio-user-guide/what-is-talend-studio
- ✓ ***Power BI Documentation***
  https://learn.microsoft.com/en-us/power-bi/
- ✓ ***Guia de procesos ETL***
  https://www.iebschool.com/blog/que-son-los-procesos-etl-big-data/
- ✓ ***Descripción general de Buckets en Amazon Web Services***
  https://docs.aws.amazon.com/es_es/AmazonS3/latest/userguide/UsingBucket.html
- ✓ ***¿Qué es Amazon S3? Definición y características***
  https://docs.aws.amazon.com/es_es/AmazonS3/latest/userguide/Welcome.html
- ✓ ***Clúster y sus capacidades en Amazon Web Services***
  https://docs.aws.amazon.com/es_es/AmazonECS/latest/developerguide/clusters.html
- ✓ ***AWS Documentation***
  https://docs.aws.amazon.com/AmazonS3/latest/gsg/CreatingABucket.html
- ✓ ***AWS Identity and Access Management (IAM) Documentation***
  https://docs.aws.amazon.com/iam/index.html
- ✓ ***AWS S3 Bucket Policies***
  https://docs.aws.amazon.com/AmazonS3/latest/dev/access-policy-language-overview.html

Glosario

A

Almacén de Datos (Data Waterhouse): Repositorio centralizado de datos que se utiliza para el análisis y la generación de informes.

AWS: Es una plataforma de servicios en la nube ofrecida por Amazon. Proporciona una amplia variedad de servicios de computación,

B

Big Data: Conjunto de técnicas y tecnologías diseñadas para manejar conjuntos de datos extremadamente grandes y complejos.

Bucket: Contenedor de almacenamiento que almacena datos, archivos o información de Amazon Web Services.

C

CSV: Es un formato de archivo simple utilizado para almacenar y representar datos tabulares en forma de texto plano

Clúster: En Amazon Redshift, un "clúster" se refiere a un grupo de nodos que trabajan juntos para proporcionar una base de datos en la nube altamente escalable y eficiente para análisis de datos.

D

Dashboard: En Power BI, un dashboard se refiere a una representación visual interactiva y consolidada de datos e información clave.

Dataset: Se refiere a un conjunto de datos que comparten una o varias características comunes o que están relacionados de alguna manera

Data Warehouse: Almacén de datos, es una colección de datos orientada a un determinado ámbito, integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza

Dimensiones: Las dimensiones son elementos clave en el diseño de modelos dimensionales que ayudan a organizar y estructurar los datos para facilitar el análisis y la consulta eficientes.

E

ETL: Es un proceso concreto que permite la correspondiente extracción, transformación y carga de los datos.

Esquema estrella: Es un diseño específico utilizado en el modelado dimensional de bases de datos, que se enfoca en organizar la información de manera que sea eficiente para realizar consultas analíticas

ERPNext: Es un software gratuito y de código abierto integrado de planificación de recursos empresariales desarrollado por la empresa de software india Frappe Technologies. y basado en el sistema de base de datos MariaDB que utiliza Frappe, un marco de trabajo del lado del servidor basado en Python.

F

Fact table: Es un componente fundamental en el modelado dimensional de bases de datos, utilizado en almacenes de datos y sistemas de inteligencia empresarial. La tabla de hechos almacena las métricas o medidas cuantitativas clave que se están analizando en un entorno de negocios.

G

GitHub: es una plataforma en línea para el desarrollo colaborativo de software y la gestión de versiones. Ofrece servicios de alojamiento de repositorios de código, seguimiento de problemas, control de versiones e integración continua.

Granularidad: El término "granularidad" se refiere al nivel de detalle o a la escala en la que se recopilan, almacenan o procesan datos.

I

Ingeniera de datos: Se refiere al conjunto de procesos, técnicas y metodologías utilizadas para diseñar, construir, gestionar y optimizar la arquitectura de datos, así como para facilitar el flujo eficiente de información

IAM: Es un servicio AWS que permite gestionar de manera segura el acceso a los recursos y servicios de la plataforma y controlar quién tiene acceso a qué recursos en su entorno.

M

Métricas: Las métricas proporcionan indicadores clave que permiten a los profesionales de ingeniería de datos evaluar el éxito de sus operaciones, identificar áreas de mejora y garantizar que los sistemas de datos cumplan con los objetivos comerciales y técnicos.

Modelado Dimensional: Es una técnica utilizada en ingeniería de datos para diseñar bases de datos que sean eficientes para realizar consultas analíticas y facilitar la presentación de informes en sistemas de inteligencia empresarial

P

Power BI: Es una suite de herramientas de análisis y visualización de datos desarrollada por Microsoft. Permite a los usuarios conectarse a diversas fuentes de datos, transformar y modelar esos datos, y crear informes interactivos.

R

Raw: Zona de almacenamiento de archivos crudos, sin procesar, en un bucket de Amazon S3.

Redshift: es un servicio de almacenamiento de datos en la nube ofrecido por AWS. Se trata de un sistema de gestión de bases de datos relacionales diseñado específicamente para el análisis y la generación de informes a gran escala.

S

Staging: Zona del bucket destinada al resguardo de los datos que han sido transformados mediante los procesos ETL.

T

Talend Studio: Es una suite de herramientas de código abierto que ofrece capacidades de integración de datos, transformación, migración y carga, así como también otras funciones de integración y procesamiento de datos.

Anexos

Anexo 1: Script para la estructura estrella en RedShift.

```
-- Crear las tablas de dimensiones para AWS Redshift

CREATE TABLE DimProducto (
    ProductoKey INT PRIMARY KEY IDENTITY(1,1) not null sortkey distkey,
    ProductoID VARCHAR(50),
    Nombre VARCHAR(100),
    Categoria VARCHAR(50),
    Subcategoria VARCHAR(50),
    Descripcion VARCHAR(100)
) diststyle key;


CREATE TABLE DimCliente (
    ClienteKey INT PRIMARY KEY IDENTITY(1,1) not null sortkey distkey,
    ClienteID VARCHAR(50),
    Nombre VARCHAR(100),
    Nit VARCHAR(50),
    Giro VARCHAR(150),
    TipoDeCliente VARCHAR(50),
    FechaRegistro DATE
) diststyle key;


CREATE TABLE Dimfecha (
    TiempoKey INT PRIMARY KEY IDENTITY(1,1) not null sortkey distkey,
    datekey integer ENCODE az64,
    date date ENCODE az64,
    daynumberofweek integer ENCODE az64,
    daynumberofmonth integer ENCODE az64,
    daynumberofyear integer ENCODE az64,
    weeknumberofyear integer ENCODE az64,
    monthnumberofyear integer ENCODE az64,
    calendarquarterofyear integer ENCODE az64,
    calendarsemesterofyear integer ENCODE az64,
    calendaryear integer ENCODE az64,
    calendaryearweek character varying(256) ENCODE lzo,
    calendaryearmonth character varying(256) ENCODE lzo,
    calendaryearquarter character varying(256) ENCODE lzo,
    calendaryearsemester character varying(256) ENCODE lzo,
```

```
    englishdaynameofweek character varying(256) ENCODE lzo,
    spanishdaynameofweek character varying(256) ENCODE lzo,
    englishmonthname character varying(256) ENCODE lzo,
    spanishmonthname character varying(256) ENCODE lzo
) diststyle key;


CREATE TABLE DimDireccion (
    DireccionKey INT PRIMARY KEY IDENTITY(1,1) not null sortkey distkey,
    DireccionID VARCHAR(50),
    Zona VARCHAR(100),
    Ciudad VARCHAR(50),
    Distrito VARCHAR(50),
    Departamento VARCHAR(50)
) diststyle key;


-- Crear la tabla de hecho
CREATE TABLE FacFactura (
    FacturaKey INT PRIMARY KEY IDENTITY(1,1) not null sortkey distkey,
    FacturaID VARCHAR(50),
    ProductoKey INT,
    ClienteKey INT,
    TiempoKey INT,
    DireccionKey INT,
    Cantidad INT,
    Precio DECIMAL(10,2),
    Iva Decimal (10,2),
    Importe Decimal (10,2),
    Descuento DECIMAL (10,2)
) diststyle key;


ALTER TABLE facfactura ADD constraint producto FOREIGN KEY (ProductoKey)
REFERENCES dimproducto(ProductoKey);
ALTER TABLE facfactura ADD constraint cliente FOREIGN KEY (ClienteKey)
REFERENCES DimCliente(ClienteKey);
ALTER TABLE facfactura ADD constraint tiempo FOREIGN KEY (TiempoKey)
REFERENCES DimFecha(TiempoKey);
ALTER TABLE facfactura ADD constraint direccion FOREIGN KEY (DireccionKey)
REFERENCES DimDireccion(DireccionKey)
```

Anexo 2: Presupuesto de una computadora de escritorio gama media

	Componente	Precio (USD)
2	Procesador (CPU): Intel Core i5-11400	$200
3	Placa base: MSI B460M PRO-VDH WiFi	$100
4	Memoria RAM: Corsair Vengeance LPX 16GB (2 x 8GB) DDR4-3200	$70
5	Almacenamiento SSD: Kingston A2000 500GB NVMe	$70
6	Tarjeta gráfica (GPU): Integrada con el procesador	$0
7	Fuente de alimentación (PSU): EVGA 500 W1 80+ WHITE 500W	$40
8	Gabinete: NZXT H510	$70
9	Sistema operativo: Windows 10 Home	$120
10	Monitor: Acer R240HY 23.8 pulgadas	$120
11	Teclado y ratón: Logitech MK270 Combo	$25
12	Altavoces o auriculares: Logitech S120 Speaker System	$20
13	**Costo Total:**	$835

Imagen 50: Presupuesto de hardware

Anexo 3: Cálculos de presupuesto de implementación

Para calcular los días laborales y el costo mensual, primero necesitamos determinar cuántas horas se trabajan en un mes y luego multiplicar ese número por la tarifa por hora.

1. Horas trabajadas diariamente: 4 horas diarias
2. Días laborales a la semana: 6 días (de lunes a sábado)
3. Horas trabajadas a la semana: 4 horas/día $\times$ 6 días/semana = 24 horas/semana
4. Horas trabajadas al mes: 24 horas/semana $\times$ 4 semanas/mes = 96 horas/mes
5. Costo mensual: 96 horas/mes $\times$× 15 $/hora = 1,440 $/mes

Este costo mensual es por persona y tomando en cuenta que son 3 desarrolladores y se trabajara un total de 9 meses, el costo aumenta significativamente, no obstante, estos gastos no se cargaran.

Los servicios bacisco se calculan de la siguiente forma.

Para la Energía eléctrica se tomó un gasto promedio de $15 mensuales, esto se debe de multiplicar por 9 meses como también por la cantidad de recursos humanos que serían 3, haciendo un total de: $405.00

El Servicios de internet residencial se tomará como base el de 100 mbs, con un costo de $60 para un total de 9 meses: $540.00

Anexo 4: Script para crear base de datos en SQL Server

```
-- Dimensión Producto
CREATE TABLE DimProducto (
    ProductoKey INT PRIMARY KEY,
    ProductoID VARCHAR(50),
    Nombre VARCHAR(100),
    Categoria VARCHAR(50),
    Subcategoria VARCHAR(50),
    Descripcion VARCHAR(100),
    -- Otros atributos del producto
);

-- Dimensión Cliente
CREATE TABLE DimCliente (
    ClienteKey INT PRIMARY KEY,
    ClienteID VARCHAR(50),
    Nombre VARCHAR(100),
    Nit VARCHAR(50),
    Giro VARCHAR(150),
    TipoDeCliente VARCHAR(50),
    FechaRegistro DATE,
    -- Otros atributos del cliente
);

-- Dimensión Tiempo
CREATE TABLE DimTiempo (
    TiempoKey INT PRIMARY KEY,
    datekey integer,
    date date,
```

```
    daynumberofweek integer,
    daynumberofmonth integer,
    daynumberofyear integer,
    weeknumberofyear integer,
    monthnumberofyear integer,
    calendarquarterofyear integer,
    calendarsemesterofyear integer,
    calendaryear integer,
    calendaryearweek character varying(256),
    calendaryearmonth character varying(256),
    calendaryearquarter character varying(256),
    calendaryearsemester character varying(256),
    englishdaynameofweek character varying(256),
    spanishdaynameofweek character varying(256),
    englishmonthname character varying(256),
    spanishmonthname character varying(256),
    -- Otros atributos relacionados con la fecha
);


-- Dimensión Dirección
CREATE TABLE DimDireccion (
    DireccionKey INT PRIMARY KEY,
    DireccionID VARCHAR(50),
    Zona VARCHAR(100),
    Ciudad VARCHAR(50),
    Distrito VARCHAR(50),
    Departamento VARCHAR(50),
    -- Otros atributos de la dirección
);
```

```
-- Crear la tabla de hecho
CREATE TABLE FacFactura (
    FacturaKey INT PRIMARY KEY,
    FacturaID VARCHAR(50),
    ProductoKey INT,
    ClienteKey INT,
    TiempoKey INT,
    DireccionKey INT,
    Cantidad INT,
    Precio DECIMAL(10,2),
Iva Decimal (10,2),
Importe Decimal (10,2),
Descuento DECIMAL (10,2),
  -- Otros atributos de la factura
    FOREIGN KEY (ProductoKey) REFERENCES DimProducto(ProductoKey),
    FOREIGN KEY (ClienteKey) REFERENCES DimCliente(ClienteKey),
    FOREIGN KEY (TiempoKey) REFERENCES DimTiempo(TiempoKey),
    FOREIGN KEY (DireccionKey) REFERENCES DimDireccion(DireccionKey)
);
```