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

CURSO DE ESPECIALIZACIÓN DE INGENIERÍA DE DATOS.

ANÁLISIS DEL PROCESO DE VENTAS PARA LA EMPRESA ¨LICA S.A. De C.V.¨ QUE UTILIZA EL SOFTWARE DE SOFTLAND

PRESENTADO POR:

BARAHONA RAMOS, ADONAY ABSALÓN

DÍAZ MEJÍA, CARLOS DALTON

HERNÁNDEZ SÁNCHEZ, ELMER ALEXANDER

PARA OPTAR AL TÍTULO DE:

INGENIERO DE SISTEMAS INFORMÁTICOS

CUIDAD UNIVERSITARIA, ENERO DE 2023

UNIVERSIDAD DE EL SALVADOR

RECTOR:

MSC. ROGER ARMANDO ARIAS ALVARADO

SECRETARIO GENERAL:

ING. FRANCISCO ANTONIO ALARCON SANDOVAL

FACULTAD DE INGENIERÍA Y ARQUITECTURA

DECANO:

PhD. EDGAR ARMANDO PEÑA FIGUEROA

SECRETARIO:

ING. JULIO ALBERTO PORTILLO

ESCUELA DE INGENIERÍA DE SISTEMAS INFORMÁTICOS

DIRECTOR:

ING. RUDY WILFREDO CHICAS VILLEGAS

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

Trabajo de Graduación previo a la opción al Grado de:

INGENIERO DE SISTEMAS INFORMÁTICOS

Título:

ANÁLISIS DEL PROCESO DE VENTAS PARA LA EMPRESA ¨LICA S.A. De C.V.¨ QUE UTILIZA EL SOFTWARE DE SOFTLAND

Presentado por:

BARAHONA RAMOS, ADONAY ABSALÓN

DÍAZ MEJÍA, CARLOS DALTON

HERNÁNDEZ SÁNCHEZ, ELMER ALEXANDER

Trabajo de Graduación Aprobado por:

Docente asesor:

MSC. RENÉ FABRICIO QUINTANILLA GÓMEZ

CUIDAD UNIVERSITARIA, ENERO DE 2023

Trabajo de Graduación Aprobado por:

Docente asesor:

MSC. RENÉ FABRICIO QUINTANILLA GÓMEZ

ÍNDICE

INTRODUCCIÓN i

CAPÍTULO I: ESPECIFICACIÓN DEL PROYECTO 2

a. Situación actual 2

i. Antecedentes 2

ii. Descripción del problema 6

iii. Planteamiento del problema 7

b. Objetivos 8

i. Objetivo general 8

ii. Objetivos específicos 8

c. Alcances 9

d. Justificación 10

e. Cronograma de actividades 11

f. Presupuesto 13

CAPÍTULO II: ANÁLISIS Y DISEÑO DE LA PROPUESTA DE SOLUCIÓN 15

a. Metodología de trabajo 15

b. Descripción de la propuesta de solución 15

i. Descripción del data set y diccionario de datos 15

ii. Resultados del data profiling 24

iii. Especificación de las necesidades analíticas que el modelo dimensional propuesto solventará 30

iv. Modelo dimensional 31

v. Diagrama Entidad-Relación del Data Warehouse del proceso de ventas 33

vi. Diagrama de estrella del modelo dimensional 34

vii. Mapping de las tablas del modelo dimensional 35

c. Descripción de la tecnología a utilizar 42

d. Diagrama arquitectónico de la solución 44

e. Descripción de cada componente de la solución 45

CAPÍTULO III: ESTRATEGIA DE IMPLEMENTACIÓN DE PROPUESTA DE SOLUCIÓN 47

a. Estrategia de implementación 47

i. Plan de capacitación. 47

ii. Configuración de componentes AWS 47

iii. Configuración de MS SQL Server 58

iv. Configuración de procesos ETL en Talend 59

v. Plan de migración para migración de datos históricos 63
vi. Frecuencia y horarios para la ejecución de los ETLs 65
b. Presupuesto de implementación 65
c. Análisis de resultados 67
i. Bases de datos MS SQL Server 67
ii. Bucket de S3 68
iii. Carpetas de archivos temporales 68
iv. Redshift 69
v. Procesos ETL en Talend 70
vi. Power BI 82
CONCLUSIONES 84
RECOMENDACIONES 85
Bibliografía 86
ANEXOS 87
Anexo 1: Diagrama de estrella del Data Warehouse del proceso de ventas 87

ÍNDICE DE FIGURAS

Ilustración 1: Logo de la empresa 2
Ilustración 2: Marcas 3
Ilustración 3: Categoría de productos que ofrecen 3
Ilustración 4: Diagrama BPM sobre el proceso de ventas y facturación 5
Ilustración 5: Diagrama de planteamiento de problema 7
Ilustración 6: Diagrama de Gantt 12
Ilustración 7: Costo del servicio Amazon S3 durante el desarrollo 13
Ilustración 8: Costo del servicio de Amazon Redshift 13
Ilustración 9: Modelo relacional del sistema transaccional 16
Ilustración 10: Diagrama arquitectónico de la solución 44
Ilustración 11: Listado de servicios utilizados en AWS 47
Ilustración 12: Panel principal del servicio IAM 48
Ilustración 13: Botón agregar usuarios 48
Ilustración 14: Formulario de creación de usuario IAM – Datos generales 48
Ilustración 15: Formulario de creación de usuario IAM - Permisos 49
Ilustración 16: Asignación de claves de acceso IAM 49
Ilustración 17: Creación de Bucket en S3 - Datos generales 50
Ilustración 18: Creación de Bucket en S3 - Propiedad de objetos 50
Ilustración 19: Creación de Bucket en S3 - Acceso público 51
Ilustración 20: Creación de Bucket en S3 - Control de versiones 51
Ilustración 21: Creación de Bucket en S3 - Finalización del proceso 51
Ilustración 22: Panel principal del Bucket 51
Ilustración 23: Proceso de creación de carpetas en S3 52
Ilustración 24: Carpetas principales del bucket 52
Ilustración 25: Subcarpetas correspondientes a las zonas de procesamiento 52
Ilustración 26: Cinta de opciones en la pantalla principal de Redshift 53
Ilustración 27: Proceso de creación del clúster - Datos generales 53
Ilustración 28: Proceso de creación del clúster - Elección del tamaño y numero de nodos 53
Ilustración 29: Opciones de almacenamiento del cluster 54
Ilustración 30: Resumen de configuración del cluster 54
Ilustración 31: Configuración de la base de datos en el cluster 54
Ilustración 32: Listado de clústeres creados 55
Ilustración 33: Menú de acciones del clúster 55
Ilustración 34: Activación de acceso público 55
Ilustración 35: Cinta de opciones del clúster 55
Ilustración 36: Configuración de red y seguridad 56
Ilustración 37: Reglas de entrada del cluster 56
Ilustración 38: Editor de reglas de entrada 56
Ilustración 39: Botón para abrir el editor de consultas 56
Ilustración 40: Abrir nuevo editor de consultas 56
Ilustración 41: Sentencia de creación de la base de datos 57

Ilustración 42: Ejecución del script de creación de esquemas y tablas 57
Ilustración 43: Explorador de objetos de Redshift - Esquemas y tablas 57
Ilustración 44: Ejecución del script de base de datos staging en MSSQL 58
Ilustración 45: Mensaje de ejecución exitosa 58
Ilustración 46: Explorador de objetos de Management Studio 58
Ilustración 47: Carpetas locales para archivos temporales 59
Ilustración 48: Subcarpetas locales correspondientes a las zonas de procesamiento 59
Ilustración 49: Variables de contexto a modificar 60
Ilustración 50: Variables de contexto – Data Lake 60
Ilustración 51: Variables de contexto - Base de datos de Staging 60
Ilustración 52: Variables de contexto - Base de datos transaccional 61
Ilustración 53: Variables de contexto - Data warehouse 61
Ilustración 54: Variables de contexto - Variables de entorno 62
Ilustración 55: Conexiones a bases de datos utilizadas por el ETL 62
Ilustración 56: Detalles de la conexión a MSSQL Server 62
Ilustración 57: Mensaje de conexión exitosa 62
Ilustración 58: Procesos ETL - MasterJob 63
Ilustración 59:Costo del servicio de Amazon Redshift - Implementación 66
Ilustración 60: MSSQL - Tabla de artículos en DB transaccional 67
Ilustración 61: MSSQL - Tabla de artículos en DB de staging durante la limpieza 67
Ilustración 62: MSSQL - Tabla de artículos en DB de staging finalizado el procesamiento 68
Ilustración 63: Datalake en S3 - Estructura de carpetas 68
Ilustración 64: Estructura de carpetas locales 68
Ilustración 65:Redshift - Datos de la tabla DimArticulo 69
Ilustración 66: Redshift - Registros de cargas delta 69
Ilustración 67: Proceso ETL - Actualizar delta 70
Ilustración 68: Proceso ETL - Descargar archivo de S3 70
Ilustración 69: Proceso ETL - Descargar lista de archivos 70
Ilustración 70: Proceso ETL - Subir archivo al datalake 71
Ilustración 71: Proceso ETL - Mover Archivo procesado en S3 71
Ilustración 72: Proceso ETL - Eliminar archivo local 71
Ilustración 73: Procesos ETL – Ejecución del Proceso principal de extracción de la base transaccional 72
Ilustración 74: Procesos ETL - Extracción de registros de la tabla Artículos 72
Ilustración 75: Procesos ETL - Extracción de registros de la tabla FacturaLinea 73
Ilustración 76: Procesos ETL - Consolidación de registros de Factura, FacturaLinea y Articulo 73
Ilustración 77: Procesos ETL - Extracción de registros de la tabla Bodega 74
Ilustración 78: Procesos ETL - Extracción de registros de la tabla Clientes 74
Ilustración 79: Proceso maestro de la zona RAW 75
Ilustración 80: Procesos ETL - Transformación de datos de Vendedores 75
Ilustración 81: Procesos ETL - Transformación de la información sobre ventas 76
Ilustración 82: Procesos ETL - Mapeo y cálculo de métricas sobre el proceso de venta 76

ILUSTRACIÓN 83: PROCESOS ETL - TRANSFORMACIÓN Y LOOKUP DE INFORMACIÓN DE ARTÍCULOS DE VENTA 77
ILUSTRACIÓN 84: PROCESOS ETL - PROCESO MAESTRO DE LA ZONA STAGE 77
ILUSTRACIÓN 85: PROCESOS ETL - PREPARACIÓN DE DATOS DE FACTURALINEA PARA LA ZONA DE PRESENTACIÓN 78
ILUSTRACIÓN 86: PROCESOS ETL - PREPARACION DE DATOS SOBRE PROVEEDORES PARA LA ZONA DE PRESENTACIÓN 78
ILUSTRACIÓN 87: PROCESOS ETL - PROCESO MAESTRO DE LA ZONA PRESENTATION 78
ILUSTRACIÓN 88: CARGA DE DATOS A TABLAS DE DIMENSIONES - ETAPA PRELIMINAR 79
ILUSTRACIÓN 89: CARGA DE DATOS A TABLAS DE DIMENSIONES - IDENTIFICACIÓN Y ELIMINACIÓN DE DUPLICADOS 79
ILUSTRACIÓN 90: CARGA DE DATOS A TABLAS DE DIMENSIONES - CLASIFICACIÓN DE DATOS 79
ILUSTRACIÓN 91: CREACIÓN DEL ARCHIVO PARA LA INSERCIÓN DE DATOS EN LA DIMENSIÓN 80
ILUSTRACIÓN 92:CARGA DE DATOS A TABLAS DE DIMENSIONES - ETAPA POSTERIOR A LA EJECUCIÓN 80
ILUSTRACIÓN 93: ETAPA PRELIMINAR A CARGA DE DATOS EN TABLA DE HECHOS 80
ILUSTRACIÓN 94: ELIMINACIÓN DE DATOS DUPLICADOS Y DESACTUALIZADOS 80
ILUSTRACIÓN 95: LOOKUP DE LLAVES SUBROGADAS PARA LA TABLA DE HECHOS 81
ILUSTRACIÓN 96: CREACIÓN DEL ARCHIVO PARA REALIZAR LA CARGA DE DATOS 81
ILUSTRACIÓN 97: REGISTROS EN LA TABLA DE HECHOS 81
ILUSTRACIÓN 98: TOTAL DE REGISTROS EN LA TABLA DE HECHOS 81
ILUSTRACIÓN 99: INFORME GENERAL DE VENTAS 82
ILUSTRACIÓN 100: INFORME DE INGRESOS VS VOLUMEN DE ORDENES 83
ILUSTRACIÓN 101: INFORME DE RENDIMIENTO DE VENDEDORES 83

ÍNDICE DE TABLAS

TABLA 1: CRONOGRAMA DE ACTIVIDADES 11
TABLA 2: CAMPOS DE LA TABLA ARTÍCULO 16
TABLA 3: ÍNDICES DE LA TABLA ARTÍCULO 17
TABLA 4: CAMPOS DE LA TABLA CLASIFICACIÓN 17
TABLA 5: ÍNDICES DE LA TABLA CLASIFICACIÓN 17
TABLA 6: CAMPOS DE LA TABLA CLIENTE 18
TABLA 7: ÍNDICES DE LA TABLA CLIENTE 18
TABLA 8: CAMPOS DE LA TABLA CATEGORIA_CLIENTE 19
TABLA 9: ÍNDICES DE LA TABLA CATEGORIA_CLIENTE 19
TABLA 10: CAMPOS DE LA TABLA VENDEDOR 19
TABLA 11: ÍNDICES DE LA TABLA VENDEDOR 19
TABLA 12: CAMPOS DE LA TABLA PROVEEDOR 20
TABLA 13: ÍNDICES DE LA TABLA PROVEEDOR 20
TABLA 14: CAMPOS DE LA TABLA COBRADOR 20
TABLA 15: ÍNDICES DE LA TABLA COBRADOR 20
TABLA 16: CAMPOS DE LA TABLA ZONA 21
TABLA 17: ÍNDICES DE LA TABLA ZONA 21
TABLA 18: CAMPOS DE LA TABLA BODEGA 21
TABLA 19: ÍNDICES DE LA TABLA BODEGA 21
TABLA 20: CAMPOS DE LA TABLA FACTURA 22
TABLA 21: ÍNDICES DE LA TABLA FACTURA 22
TABLA 22: CAMPOS DE LA TABLA FACTURA_LINEA 23
TABLA 23: ÍNDICES DE LA TABLA FACTURA_LINEA 23
TABLA 24: DATA PROFILING DE LA TABLA ARTÍCULO 24
TABLA 25: DATA PROFILING DE LA TABLA CLASIFICACIÓN 25
TABLA 26: DATA PROFILING DE LA TABLA CLIENTE 25
TABLA 27: DATA PROFILING DE LA TABLA CATEGORIA_CLIENTE 26
TABLA 28: DATA PROFILING DE LA TABLA VENDEDOR 26
TABLA 29: DATA PROFILING DE LA TABLA PROVEEDOR 26
TABLA 30: DATA PROFILING DE LA TABLA COBRADOR 27
TABLA 31: DATA PROFILING DE LA TABLA ZONA 27
TABLA 32: DATA PROFILING DE LA TABLA BODEGA 27
TABLA 33: DATA PROFILING DE LA TABLA FACTURA 28
TABLA 34: DATA PROFILING DE LA TABLA FACTURA_LINEA 29
TABLA 35:PROCESO DEL NEGOCIO 31
TABLA 36: DEFINICIÓN DE GRANULARIDAD 31
TABLA 37: DEFINICIÓN DE DIMENSIONES 32
TABLA 38: MÉTRICAS 32
TABLA 39: DIMENSIÓN CLIENTE 35
TABLA 40: MAPEO DE LA DIMENSIÓN CLIENTE 35
TABLA 41: DIMENSIÓN BODEGA 36

TABLA 42: MAPEO DE LA DIMENSIÓN BODEGA 36
TABLA 43: DIMENSIÓN COBRADOR 36
TABLA 44: MAPEO DE LA DIMENSIÓN COBRADOR 36
TABLA 45: DIMENSIÓN PROVEEDOR 37
TABLA 46: MAPEO DE LA DIMENSIÓN PROVEEDOR 37
TABLA 47: DIMENSIÓN VENDEDOR 37
TABLA 48: MAPEO DE LA DIMENSIÓN VENDEDOR 37
TABLA 49: DIMENSIÓN FECHA 38
TABLA 50: MAPEO DE LA DIMENSIÓN FECHA 38
TABLA 51: DIMENSIÓN ARTÍCULO 39
TABLA 52: MAPEO DE LA DIMENSIÓN ARTÍCULO 39
TABLA 53: TABLA DE HECHOS VENTAS 40
TABLA 54: MAPEO DE LA TABLA DE HECHOS VENTAS 40

INTRODUCCIÓN

En la actualidad, la gestión de la información se ha vuelto fundamental para el correcto funcionamiento y éxito de las empresas. En este sentido, el uso de sistemas de almacenamiento y procesamiento de datos, como el Data Warehouse, se ha vuelto esencial para brindar a las organizaciones la capacidad de analizar y obtener información valiosa a partir de sus datos históricos.

La empresa LICA S.A de C.V ha identificado la necesidad de mejorar la información que se maneja en relación a su proceso de ventas y facturación, para ello se ha planteado la creación de un Data Warehouse que permita analizar y obtener información valiosa a través de reportes y visualizaciones.

Actualmente, la empresa utiliza el sistema de gestión de recursos empresariales (ERP) Softland para almacenar y registrar datos transaccionales e históricos relacionados con el proceso de ventas y facturación. Sin embargo, estos datos no están organizados de manera óptima para su análisis y toma de decisiones, lo que dificulta el procesamiento de los mismos para generar información de valor de forma oportuna.

Para abordar este problema, se propone implementar un modelo multidimensional de Data Warehouse utilizando herramientas tecnológicas como Microsoft SQL Server, Softland, Talend Open Studio, Amazon Web Services (AWS) tales como: Amazon S3 (Simple Storage Service), y Amazon Redshift, Amazon IAM. Esto permitirá extraer, transformar y cargar datos de un sistema transaccional en estructuras de datos optimizadas para facilitar el análisis y generación de información para apoyar en la toma de decisiones de la empresa.

El proceso de implementación incluirá la creación de zonas de almacenamiento y procesamiento en Amazon S3 para los datos crudos (RAW), los datos procesados y preparados para la carga (STAGE) y los datos procesados y limpios para su análisis (PRESENTATION). Además, se aplicarán tareas de ETL (Extracción, Transformación, Carga) para eliminar duplicados, realizar cálculos y definir el formato de los datos. Los datos se cargarán en Amazon Redshift, una base de datos en la nube, y se utilizarán para crear visualizaciones y facilitar el análisis y la comprensión de los mismos.

Para garantizar la seguridad y el control de acceso a los recursos y servicios utilizados en AWS, se implementará Amazon IAM, un sistema de gestión de seguridad que permite crear roles, permisos y políticas de acceso basados en usuarios y grupos.

Este proyecto busca mejorar la estructuración y procesamiento de datos en la empresa LICA S.A de C.V mediante la implementación de un Data Warehouse utilizando herramientas tecnológicas de vanguardia y un sistema de gestión de seguridad para garantizar la confidencialidad y el control de acceso a los datos. Se espera que esto permita a la empresa tomar decisiones más informadas y mejorar su rendimiento empresarial, esto por medio de reportes analíticos que se realizarán en la herramienta de análisis de datos Power BI dando respuesta a la necesidad de LICA S.A de C.V

CAPÍTULO I: ESPECIFICACIÓN DEL PROYECTO

a. Situación actual

i. Antecedentes

LICA S.A de C.V (Distribuidores exclusivos) es una empresa que se dedica a la comercialización de productos varios a consumidores y comerciantes, realizando ventas al por mayor y menor. Los productos que ofrecen van desde artículos de abarrotes, belleza/cosméticos y limpieza. El mayor cliente que poseen es una cadena de supermercados muy conocida en el país.

La empresa fue fundada el 20 de mayo de 1975 con la finalidad de fabricar y distribuir productos cosméticos a nivel regional. Años más tarde, la dirección decide separar la fabricación de la distribución. Creándose de esta forma dos empresas hermanas, Industrias Cosméticas SA de CV, dedica a la fabricación de cosméticos, y LICA SA de CV, dedicada a la comercialización. Se concentra en sus inicios a la distribución y comercialización de productos a nivel nacional.

A través de los años, LICA ha ido innovando con nuevas líneas en el área de distribución, con el fin de satisfacer a sus clientes en más aspectos de sus vidas. Se esfuerza por brindar a sus clientes el mejor servicio posible, que se traduce en una atención permanente, despachos a tiempo, material de soporte, promociones y ofertas.

En 2010, se inició un nuevo proyecto de producción en el área de Alimentos, entre sus marcas nace "Delirice" con el objetivo de ofrecerle a sus consumidores, una alternativa de galletas sana, saludable y nutritiva.

Logo de la empresa

Ilustración 1: Logo de la empresa

Marcas de la empresa

Ilustración 2: Marcas

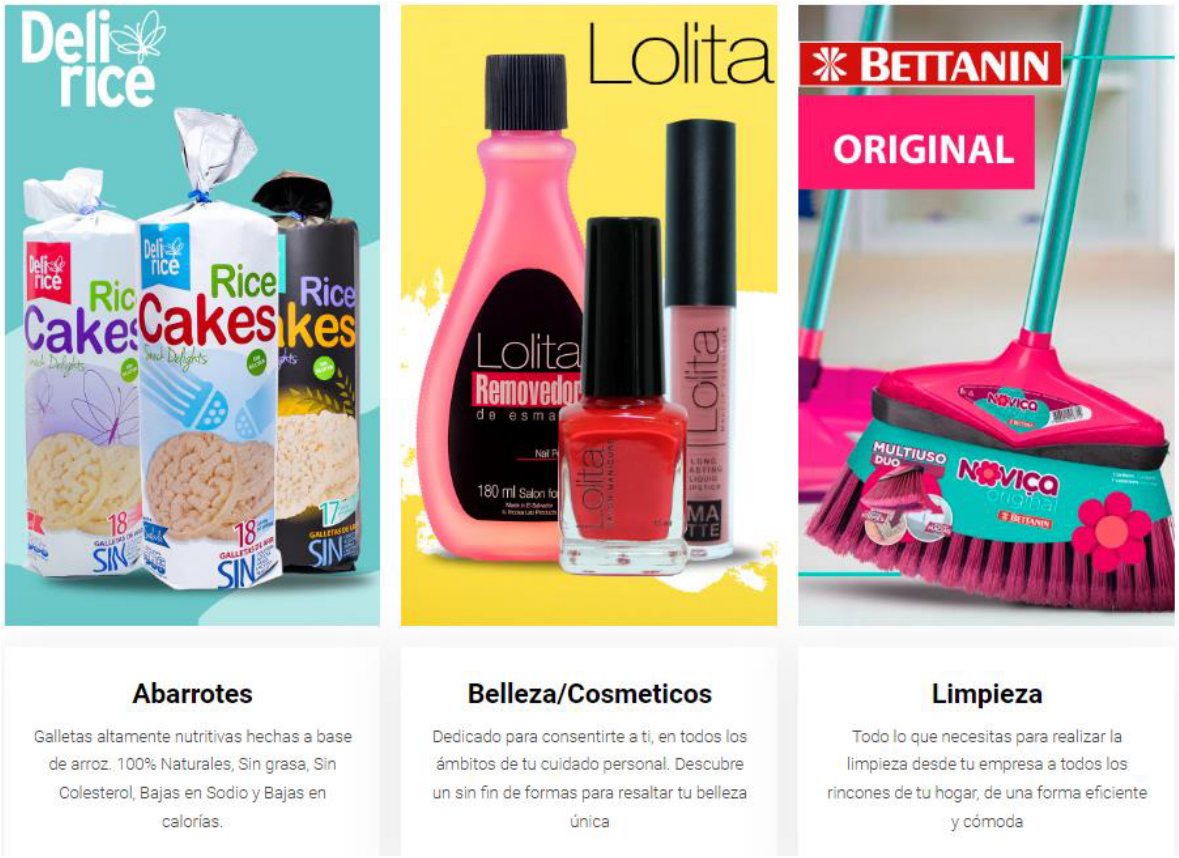


Ilustración 3: Categoría de productos que ofrecen

Descripción del proceso de ventas en el sistema transaccional

Las transacciones de ventas en el sistema inician mediante pedidos, lo cuales pueden ser ingresados por un vendedor o por un validador (persona encargada de verificar el pedido y aprobar dicho pedido). Aunque el sistema permite la gestión de cotizaciones, la institución no hace uso de esa función en el sistema, sino que, proceden directamente a la gestión de los pedidos.

Durante la creación del pedido al asignarle el cliente al pedido, el usuario debe seleccionar el valor consecutivo que se le agregara al documento, pudiendo ser de tipo Factura de consumidor final o Crédito fiscal. Esta selección afectara los valores impresos en la factura. Si fuese crédito fiscal, al subtotal debe restársele un 1% de la venta neta, siguiendo lo establecido en el Art. 163 del Código Tributario; además de detallarse el valor del IVA de la venta. Caso contrario sólo se muestro los totales sin mayor detalle.

Una vez los pedidos han sido creados, estos son revisados por un validador, el cual es responsable de verificar, a través del ERP[1], si se cuenta con el inventario suficiente en la bodega para cada uno de los productos que requiere el cliente. No se realizan pedidos parciales, la institución entrega la cantidad solicitada del producto o no lo entrega. En caso de no contar con la cantidad suficiente para surtir el pedido, el sistema permite la eliminación de las líneas del pedido cuyo inventario no es mayor o igual a lo solicitado. Cabe mencionar, que la institución únicamente hace uso de dos bodegas: General y Averías. Sus artículos sólo son despachados desde la bodega general, es por ello que, otro paso dentro de la validación es cerciorarse que los artículos serán tomados de la bodega correcta, previo a validar el inventario. Adicionalmente, en esta fase inicial, en la cual los pedidos están identificados con el estado de 'Normal', el ERP permite aplicar descuentos sobre cada producto y de forma global a todo el pedido, si así se requiere. Otro paso que el validador debe realizar es asignarle un cobrador al pedido con base a la ruta configurada en el cliente, el cual es copiada al pedido, y una hoja de ruta que maneja la institución internamente donde detalla que ruta que tendrá cada cobrador diariamente. Posterior a la verificación, el validador 'Aprueba' el pedido, cuando esto sucede el ERP cambia el estado del pedido de 'Normal' a 'Aprobado'.

El sistema permite la reserva de inventario para cada pedido en estado 'Normal' o 'Aprobado', esto causa que el inventario del producto en la bodega seleccionada se vea reducido cuando se consulta. Sin embargo, esta función no es utilizada por la institución, lo que causa que el ERP actualice el valor del inventario hasta que el pedido ha llegado a su fase final; es decir, ha sido facturado.

Cuando a un pedido se le ha aplicado descuentos por producto y, además, un descuento global, el sistema totaliza las líneas del pedido (total mercadería), dicho total ya ha sido afectado por el descuento por producto, sobre ese total de mercadería se aplica el descuento global, el cual se maneja mediante porcentaje. El sistema no permite agregar el monto directamente.

Luego de que el pedido ha sido aprobado, pasa al departamento de facturación. Este departamento realiza una revisión muy parecida a la del validador para verificar que no se haya escapado ningún detalle. En caso, de alguna inconsistencia, el departamento de facturación desaprueba el pedido, a lo cual el sistema reacciona actualizando el pedido a estado 'Normal'. Si todo está en orden se utiliza la función de 'Generar factura/Boleta'. Esta acción desencadena un proceso que realiza la creación de los registros en la base de datos en las tablas que almacenan la información del encabezado de la factura y sus líneas y se imprime el documento. Cabe destacar que, una vez impresa la factura no se puede modificar en caso requiera algún ajuste el documento, de ser necesario se efectúa una refacturación; es decir, realizar todo el proceso nuevamente, posterior a la anulación de la factura emitida.

A continuación, un diagrama BPM del proceso que la institución lleva a cabo en el sistema.

[1] Enterprise Resource Planning, por sus siglas en inglés. La planificación de recursos empresariales es un sistema que ayuda a automatizar y administrar los procesos empresariales de distintas áreas: finanzas, fabricación, venta al por menor, cadena de suministro, recursos humanos y operaciones. – *Definición de ERP, Microsoft Dynamics 365*

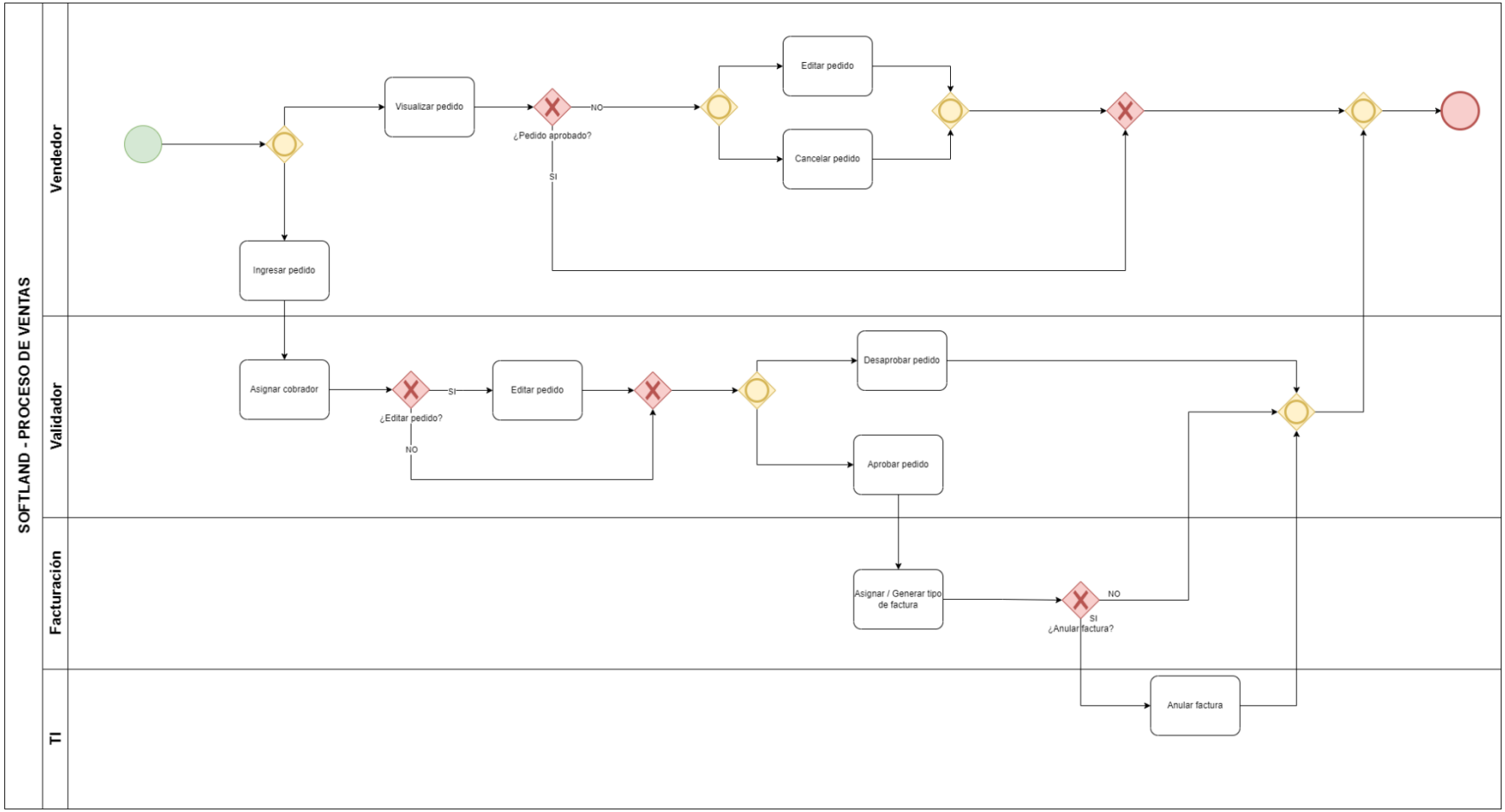


Ilustración 4: Diagrama BPM sobre el proceso de ventas y facturación

ii. Descripción del problema

La empresa presenta necesidades analíticas que, al ser resueltas, estas pueden brindar un apoyo al proceso del análisis sobre la situación de la empresa en el tiempo en función de sus clientes y vendedores. La institución cuenta con datos históricos, generados mediante el ERP de Softland, de aproximadamente 6 años. Durante los procesos de entrevista con los stakeholders, el perfilado de datos, y análisis de los mismos se lograron identificar la necesidad de dos reportes nuevos: Reporte de ventas, Reporte comparativo de ingresos vs volumen de ventas y la automatización del Reporte de rendimiento de vendedores que actualmente generan de forma manual usando la base de datos del sistema transaccional.

El reporte de ventas tiene como objetivo dar a conocer los siguientes indicadores:

- ✓ La utilidad total de las ventas
- ✓ El costo total de las ventas
- ✓ El total de lo vendido
- ✓ La cantidad de ventas hechas
- ✓ El total de artículos vendidos
- ✓ El total de impuesto de las ventas
- ✓ El total de descuento
- ✓ El producto más y menos vendido
- ✓ El tiempo de recompra de cada cliente

Por otro lado, el reporte comparativo de ingresos y crecimiento en volumen de ventas busca brindar información sobre incrementos en los montos vendidos para determinar si los posibles incrementos en los ingresos se deben al aumento en los precios de los productos por factores económicos externos o a un mayor número de productos vendidos.

Finalmente, la automatización del reporte de rendimiento de los vendedores para conocer la situación de las ventas en función sus vendedores y con el objetivo de brindar transparencia en el establecimiento de las metas de ventas de sus colaboradores. Además, ofrecer incentivos y bonos de acuerdo al avance que cada uno presente a lo largo de un periodo establecido.

iii. Planteamiento del problema

Objetivo: Mejorar el acceso a la información relacionada con el proceso de ventas y facturación de la empresa LICA S.A de C.V a través de la creación de reportes analíticos utilizando un análisis dimensional y la implementación de un Data Warehouse.

Entradas: Conformado por el Dataset del proceso de venta y facturación, estos datos serán los insumos de entrada principales para solventar los requerimientos solicitados

Proceso: El proceso está conformado por las tareas de Extracción, transformación y carga de datos realizados por medio de ETL´s, estos procesos se realizan por cada una de las zonas que conforman el Data Warehouse.

Salidas: Están conformadas por visualizaciones que mostrará a modo de reporte los requerimientos solicitados los cuales son: Registro de ventas, Reporte comparativo de ingreso, reporte de rendimiento de vendedores.

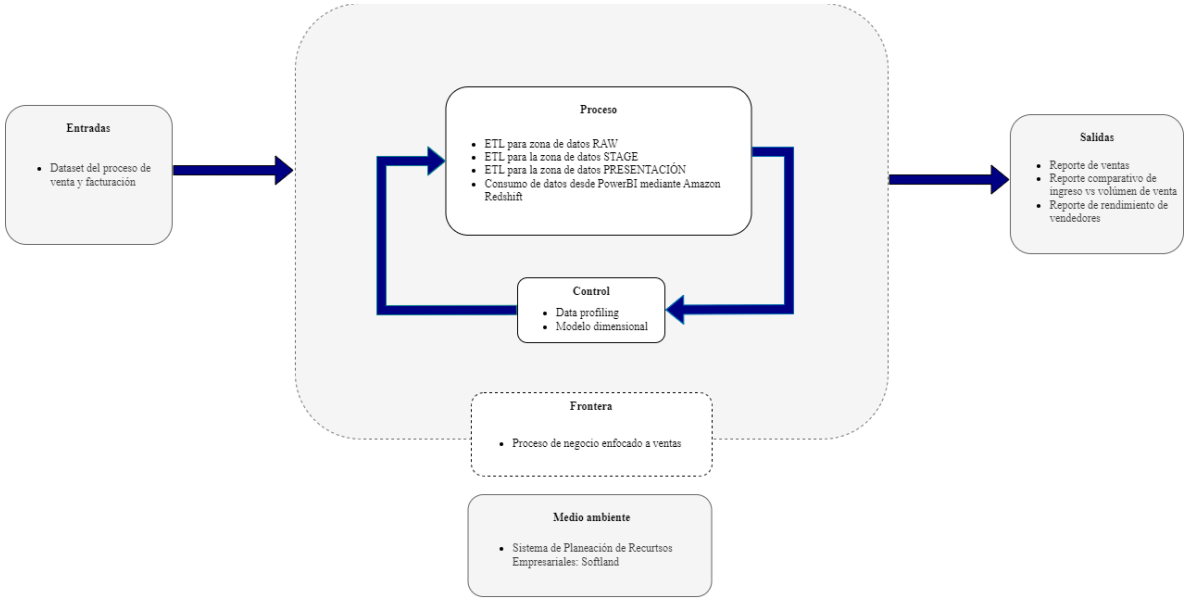


Ilustración 5: Diagrama de planteamiento de problema

b. Objetivos

i. Objetivo general

Diseñar un modelo multidimensional de Data Warehouse que soporte el análisis del proceso de ventas de la empresa “LICA S.A. de C.V.” que utiliza el software ERP Softland

ii. Objetivos específicos

- Conocer el funcionamiento del proceso de ventas de la empresa en el ERP Softland
- Analizar la fuente de datos en la que se almacena los datos transaccionales de las ventas
- Realizar un perfilado de los datos de ventas de la empresa
- Crear un modelo multidimensional que solvente las necesidades analíticas del proceso de ventas.
- Construir el proceso ETL (Extracción, Transformación y Carga) que traslade los datos de la fuente de datos transaccional al Data Warehouse (modelo multidimensional)
- Integrar el modelo dimensional desde Amazon Redshift con la herramienta de Power BI para la construcción de las visualizaciones que den respuesta las necesidades analíticas sobre el proceso de ventas de la empresa.

c. Alcances

- Realizar un esquema en estrella dimensional. Este esquema servirá como base para almacenar los datos que se utilizarán para proporcionar soluciones a las necesidades analíticas de la empresa.

- Presentar un Dashboard consolidado con los requerimientos solicitados. Para ello, se utilizará la herramienta Power BI para proporcionar un conjunto de reportes que contengan la información necesaria para la toma de decisiones de la empresa. De esta manera, se podrá acceder a los datos de manera visual y fácilmente comprensible.

- Implementar un proceso de ETL (extracción, transformación, carga), para llevar a cabo el proceso se construirán ETL's con tareas específicas, las cuales se encargarán de obtener, transformar y cargar los datos entre las distintas áreas del Data Lake. La utilización de estas tareas tiene como objetivo facilitar la manipulación y la integración de los datos en la solución propuesta.

- Crear un repositorio en GitHub para gestionar el proceso de construcción y la documentación técnica asociada al proyecto, en este repositorio se guardarán los JOBS realizados durante el ciclo de construcción de la solución. De esta manera, se podrá acceder a ellos de manera organizada y controlada.

- Presentar la documentación de la solución en formato digital para que los usuarios finales y técnicos puedan comprender el proceso de construcción. Esto permitirá acceder a la información de manera rápida y facilitará la comprensión de cómo se ha desarrollado la solución.

d. Justificación

Las organizaciones cada día se plantean estrategias basadas en priorizar al cliente, existen más demanda en la calidad de los servicios, hay más exigencia en la respuesta inmediata que se le debe de dar al cliente y sobre todo hay necesidad de que el almacenamiento de los datos sea robusto y eficaz; es decir, se necesitan de bases de datos que realicen mejor el procesamiento de datos y sean capaces de dar respuesta inmediata a peticiones que se requieren.

Como consecuencia, las instituciones han optado por una solución para mejorar la información de las bases de datos, siendo esta una forma adecuada de fomentar el éxito, mejorar el desempeño empresarial, reducir costos y lograr cambios.

El Data Warehouse llega como una solución a las organizaciones para una mejor estructuración y procesamiento de datos. Un Data Warehouse (almacén de datos) desempeña in rol esencial en las organizaciones. Como su nombre lo indica, es un depósito de datos inmenso que aloja toda la información que las organizaciones generan, de tal forma que resulte fácil y sencillo el ingreso de los usuarios a este contenido para los análisis requeridos. El diseño y la metodología del almacén de datos debe realizarse adaptada a las necesidades de cada organización.

El data warehouse debe estar perfectamente organizado en secciones y unidades para que la accesibilidad y usabilidad de la información sean garantizadas, atendiendo a cada área o departamento y a la utilidad final que se requiera.

Las organizaciones ya no ponen en duda que los datos son un activo muy importante para el funcionamiento y supervivencia de su negocio, también reconocen la importancia de que estos datos aporten un valor agregado y una guía certera para la toma de decisiones.

Por lo antes mencionado, el presente proyecto explora herramientas para optimizar la elaboración de informes relacionados al proceso de ventas para apoyar a la organización en la toma de decisiones y se plantea brindar una solución de data warehouse mediante un modelo multidimensional que permita extraer, transformar y cargar datos, de un sistema transaccional, en estructuras de datos optimizadas para dar respuesta a interrogantes acerca del rendimiento del negocio.

La propuesta de la solución detallada en este informe es proveer el empleo de herramientas informáticas, conservación y explotación de datos histórica del sistema transaccional, omitir registros innecesarios en las bases de datos y generar aquellos indispensables para generar información valiosa para la toma de decisiones y así facilitar a los administradores de la organización el acceso a sus datos para que les permita afrontar las exigencias competitivas.

e. Cronograma de actividades

Nombre de tarea	Duración	Comienzo	Fin
Primer contacto con la empresa	1 día	sáb 7/5/22	sáb 7/5/22
Reunión equipo UES para planificar propuesta de solución	1 día	dom 8/5/22	dom 8/5/22
Construcción de propuesta de proyecto	4 días	mar 17/5/22	vie 20/5/22
Presentación de propuesta a empresa LICA	1 día	sáb 21/5/22	sáb 21/5/22
Refinamiento de propuesta de solución	3 días	lun 23/5/22	mié 25/5/22
Realizar análisis del data set de la base de datos transaccional	5 días	lun 6/6/22	vie 10/6/22
Detallar resultados del data profiling	3 días	lun 13/6/22	mié 15/6/22
Realizar especificación de las necesidades analíticas	5 días	jue 16/6/22	mié 22/6/22
Realizar análisis del modelo dimensional	6 días	vie 24/6/22	vie 1/7/22
Realizar mapping de modelo dimensional y dataset	3 días	dom 3/7/22	mar 5/7/22
Curva de aprendizaje sobre AWS, herramienta ETL, implementación de Data Warehouse	20 días	lun 1/8/22	vie 26/8/22
Preparación y configuración de entorno de trabajo	4 días	lun 29/8/22	jue 1/9/22
Crear y restaurar base de datos transaccional en SQL Server	2 días	jue 1/9/22	vie 2/9/22
Configurar Amazon IAM	1 día	lun 5/9/22	lun 5/9/22
Configurar Amazon S3	2 días	lun 5/9/22	mar 6/9/22
Configurar Amazon Redshift	2 días	lun 5/9/22	mar 6/9/22
Construcción de JOBS en TALEND OPEN STUDIO	50 días	vie 9/9/22	jue 17/11/22
Realización de pruebas al modelo ETL construido	5 días	jue 17/11/22	mié 23/11/22
Solventar bugs detectados en los ETL's	2 días	jue 24/11/22	vie 25/11/22
Construcción de visualizaciones en POWER BI	7 días	sáb 26/11/22	lun 5/12/22
Validación y revisión de resultados de las visualizaciones en Power BI	3 días	mar 6/12/22	jue 8/12/22
Solventar bugs detectados en las visualizaciones	2 días	jue 8/12/22	vie 9/12/22
Documentación de la solución en GitHub	3 días	jue 8/12/22	sáb 10/12/22

Tabla 1: Cronograma de actividades

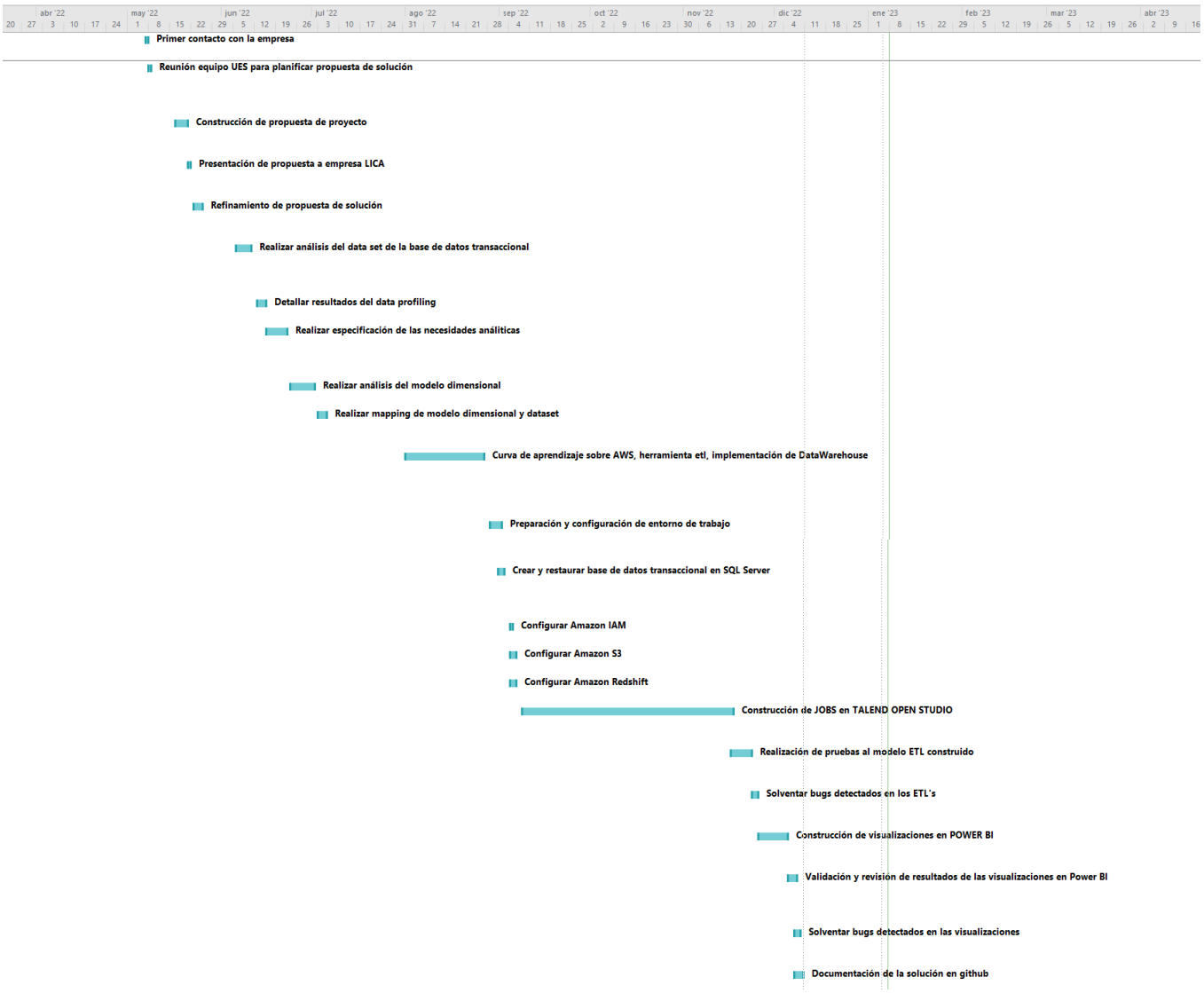


Ilustración 6: Diagrama de Gantt

f. Presupuesto

El presupuesto detallado en este apartado representa el costo de los servicios de AWS y licenciamiento de las herramientas utilizadas durante el periodo de desarrollo. El presupuesto necesario para su implementación y puesta en producción será detallado posteriormente (Capitulo III, literal b.)

Costeo de Amazon S3 (Simple Storage Service)

El costo del servicio de S3 cubre el espacio de almacenamiento utilizado durante el desarrollo, el número de peticiones realizadas de tipo PUT, COPY, POST y LIST, y de tipo GET y LIST. El estimado fue calculado en la calculadora del sitio web de Amazon (Calculadora de AWS). A continuación, se detalla el cálculo:

Ilustración 7: Costo del servicio Amazon S3 durante el desarrollo

Subtotal: $0.27

Costeo de Amazon Redshift (Almacenamiento del modelo de estrella)

El costo del servicio de almacenamiento y procesamiento de datos de Redshift cubre el número de nodos (1), tipo de instancia (dc2.large) y el número de horas de uso de la instancia durante el desarrollo (8 hrs. x 70 días = 560 hrs.). A continuación, se detalla el cálculo:

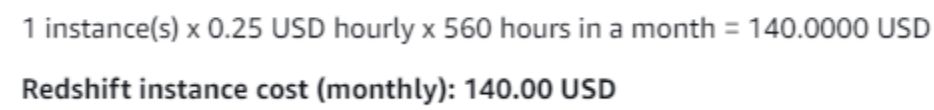


Ilustración 8: Costo del servicio de Amazon Redshift

Subtotal: $140.00

Costeo de Amazon IAM (Zona de Gobernanza)

El servicios Amazon IAM (Identity and Access Management, por sus siglas en inglés) para le manejo de accesos a los recursos de AWS es gratuito. En consecuencia, no hubo un costo.

Subtotal: $0.00

Costeo de Microsoft SQL Server Versión 2019 – Edición Express

La edición utilizada del motor de base de datos para almacenar registros temporales (desactualizados o duplicados en la zona de presentación, previos a ser escritos en Amazon Redshift). En consecuencia, no hubo un costo.

Subtotal: $0.00

Costeo de Talend Open Studio (Construcción de ETL)

La herramienta utilizada para la construcción de procesos ETL´s para la integración de datos es de open-source y sus descarga del sitio oficial es gratuita. Las funciones con las que cuenta esta versión de no-paga fue suficiente para la construcción de la solución. En consecuencia, no hubo un costo.

Subtotal: $0.00

Costeo de Microsoft Power BI

Para el consumo de los datos estructurados almacenado en Amazon Redshift y construcción de repostes-visualizaciones que dieran respuesta a las necesidades analíticas, se hizo uso del licenciamiento gratuita en conjunto con los 30 de prueba de la versión Pro. En consecuencia, no hubo un costo.

Subtotal: $0.00

Total: $140.32

Cabe mencionar que el costo total de la construcción de la solución, desde la perspectiva de servicios tecnológicos y software licenciado, fue de **$0.00** dado que para se utilizaron cuentas de estudiante para tener acceso a la capa gratuita de los servicios de AWS y la versión Pro en periodo de prueba de Microsoft Power BI.

CAPÍTULO II: ANÁLISIS Y DISEÑO DE LA PROPUESTA DE SOLUCIÓN

a. Metodología de trabajo

Para la construcción de Data Warehouse se implementó la metodología de Ralph Kimball, esta metodología es por mucho la arquitectura más utilizada al momento de la construcción de Data Warehouse, conformada por los siguientes componentes:

- **Source transactions:** son las aplicaciones y sistemas que almacenan y procesan los datos de origen operativos de la empresa, como los sistemas de ventas, contabilidad e inventario. Estos sistemas suelen ser transaccionales y se centran en el registro y procesamiento de datos en tiempo real.
  Para el presente proyecto la empresa LICA S.A de C.V cuenta con datos históricos, generados mediante el ERP de Softland, de aproximadamente 6 año.

- **Back room:** es la capa de procesamiento y almacenamiento de datos del sistema de BI. Esto incluye el Data Warehouse y las herramientas de integración de datos (ETL, por sus siglas en inglés) que se utilizan para extraer, transformar y cargar los datos de los sistemas Source transactions al Data Warehouse. También se pueden incluir en esta capa sistemas de almacenamiento y procesamiento de Big Data, como Talend, Hadoop o Spark.
  Para el análisis realizado se utilizó la herramienta Talend Open Studio.

- **Front room:** es la capa de presentación y análisis de datos del sistema de BI. Esto incluye las herramientas de análisis y reportaría que permiten a los usuarios finales acceder y analizar los datos del Data Warehouse. Estas herramientas suelen incluir dashboards, informes y análisis en línea.
  En este proyecto la representación de la información se realiza utilizando la herramienta de Power BI.

b. Descripción de la propuesta de solución

i. Descripción del data set y diccionario de datos

Luego de conocer el proceso de ventas de la organización y la forma en la que interactúan con el sistema ERP de Softland (Refiérase a los antecedentes) se ha identificado un conjunto de tablas del sistema transaccional, como punto de partida, para la solución de análisis del proceso de ventas que serán utilizadas, como fuente de datos, para el Data Warehouse. Estas tablas serán conocidas como el "Data Set", todas las tablas pertenecen al mismo esquema en el motor de base de datos de Microsoft SQL Server. Esquema 'LICASA'. A continuación, se presenta un diagrama del modelo relacional:

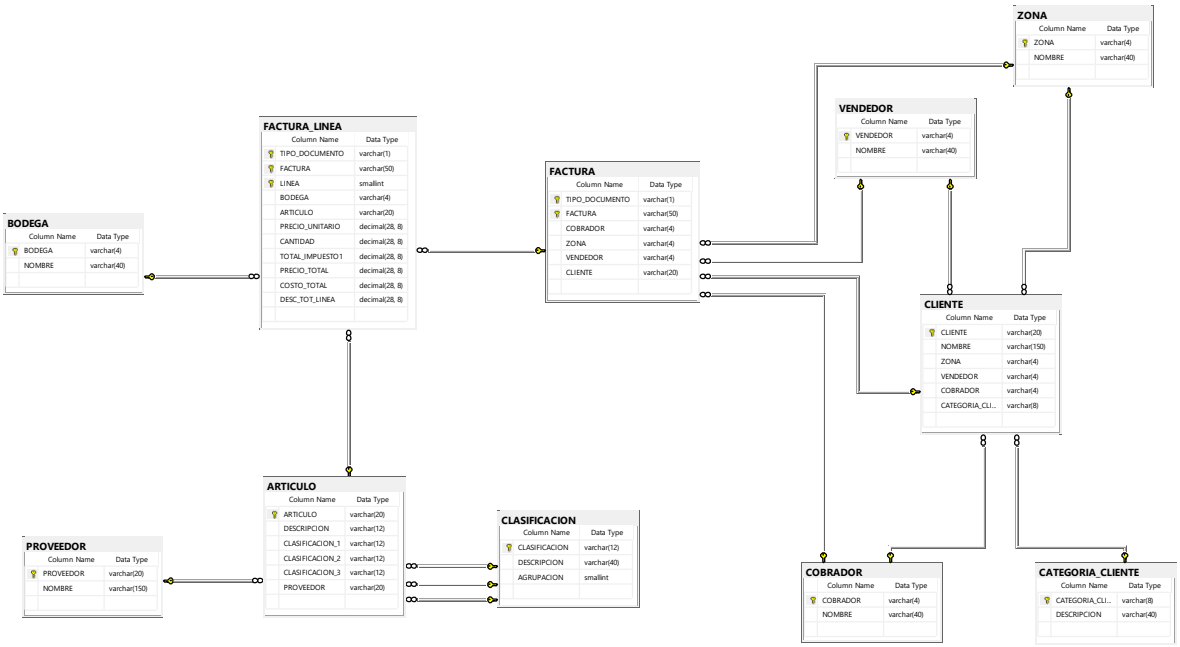


Ilustración 9: Modelo relacional del sistema transaccional

1. **Articulo:** Almacena los productos que la empresa utiliza para su proceso de ventas y para uso interno de sus operaciones.

Columna	Tipo de dato	Descripción	Valores únicos	Es requerido	Acepta valores nulos	Restricción	Notas
ARTICULO	varchar(20)	Código del articulo	Si	Si	No	Llave primaria, Único	
DESCRIPCION	varchar(254)	Nombre del articulo	No	Si	No		
CLASIFICACION_1	varchar(12)	Clasificación general del artículo (Grupo). Referencia a la tabla clasificación	No	No	Si	Llave foránea	
CLASIFICACION_2	varchar(12)	Clasificación específica del artículo (subgrupo). Referencia a la tabla clasificación	No	No	Si	Llave foránea	
CLASIFICACION_3	varchar(12)	Marca del artículo. Referencia a la tabla clasificación	No	No	Si	Llave foránea	
PROVEEDOR	varchar(20)	Proveedor del articulo	No	No	Si	Llave foránea	

Tabla 2: Campos de la tabla artículo

Índices

Nombre de la llave	Tipo	Único	Nulo	Columna(s)	Comentario
ARTICULOPK	Llave primaria	Si	No	ARTICULO	
ARTICULOIFCLASIF1	Llave foránea	No	Si	CLASIFICACION_1	
ARTICULOIFCLASIF2	Llave foránea	No	Si	CLASIFICACION_2	
ARTICULOIFCLASIF3	Llave foránea	No	Si	CLASIFICACION_3	
ARTICPROV	Llave foránea	No	Si	PROVEEDOR	

Tabla 3: Índices de la tabla artículo

Referencias:
LICASA.CLASIFICACION (CLASIFICACION_1 => CLASIFICACION)
LICASA.CLASIFICACION (CLASIFICACION_2 => CLASIFICACION)
LICASA.CLASIFICACION (CLASIFICACION_3 => CLASIFICACION)
LICASA.PROVEEDOR (PROVEEDOR)

Referenciada por:

LICASA.FACTURA_LINEA *(ARTICULO)*

2. **Clasificación:** Guarda los valores del catálogo que se utiliza para clasificar los artículos de la empresa.

Columna	Tipo de dato	Descripción	Valores únicos	Es requerido	Acepta valores nulos	Restricción	Notas
CLASIFICACION	varchar(12)	Código de la clasificación	Si	Si	No	Llave primaria, Único	
DESCRIPCION	varchar(40)	Nombre de la clasificación	No	Si	No		
AGRUPACION	smallint	Jerarquía a la que pertenece la clasificación.	No	Si	No		Tipo de dato Enumerado que permite jerarquizar los tipos de clasificación. Los valores permitidos son: 1 : Grupo 2 : Subgrupo 3 : Marca 4 : Palos 5 : PROV COD

Tabla 4: Campos de la tabla clasificación

Índices

Nombre de la llave	Tipo	Único	Nulo	Columna(s)	Comentario
CLASIFICACIONPK	Llave primaria	Si	No	CLASIFICACION	-
CLASIFIEDESCRIP	No agrupado	No	No	DESCRIPCION	
CLASIFIEAGRUPA	No agrupado	No	No	AGRUPACION	

Tabla 5: Índices de la tabla clasificación

Referenciada por:

LICASA.ARTICULO (CLASIFICACION_1 -> CLASIFICACION)
LICASA.ARTICULO (CLASIFICACION_2 -> CLASIFICACION)
LICASA.ARTICULO (CLASIFICACION_3 -> CLASIFICACION)

3. **Cliente:** Almacena a todas las personas naturales y jurídicas que forman parte de la cartera de clientes de la empresa.

Columna	Tipo de dato	Descripción	Valores únicos	Es requerido	Acepta valores nulos	Restricción	Notas
CLIENTE	varchar(20)	Código del cliente	Si	Si	No	Llave primaria, Único	
NOMBRE	varchar(150)	Nombre del cliente	No	Si	No		
ZONA	varchar(4)	Zona a la que pertenece el cliente	No	Si	No	Llave foránea	
VENDEDOR	varchar(4)	Código del vendedor asignado al cliente	No	No	Si	Llave foránea	
COBRADOR	varchar(4)	Código del cobrador que entrega los artículos al cliente	No	Si	No	Llave foránea	
CATEGORIA_ CLIENTE	varchar(8)	Código de la categoría a la que pertenece el cliente	No	Si	No	Llave foránea	

Tabla 6: Campos de la tabla cliente

Índices

Nombre de la llave	Tipo	Único	Nulo	Columna(s)	Comentario
CLIENTEPK	Llave primaria	Si	No	CLIENTE	-
CLICLIENTE	Agrupado	No	No	CLIENTE	
CLIENTEIENOMB	No agrupado		No	NOMBRE	
CLIENTEIECOBR	Llave foránea	No	No	COBRADOR	
CLIENTEIEVEND	Llave foránea	No	Si	VENDEDOR	
CLIENTEIEZONA	Llave foránea	No	No	ZONA	
CLIENTEIECATCLI	Llave foránea	No	No	CATEGORIA_CLIENTE	

Tabla 7: Índices de la tabla cliente

Referencias

LICASA.CATEGORIA_CLIENTE (CATEGORIA_CLIENTE)
LICASA.COBRADOR (COBRADOR)
LICASA.VENDEDOR (VENDEDOR)
LICASA.ZONA (ZONA)

Referenciada por

LICASA.FACTURA (CLIENTE)
LICASA.FACTURA (CLIENTE_CORPORAC -> CLIENTE)

4. **Categoria_Cliente:** Almacena el catálogo que se utiliza para categorizar a los clientes.

Columna	Tipo de dato	Descripción	Valores únicos	Es requerido	Acepta valores nulos	Restricción	Notas
CATEGORIA_ CLIENTE	varchar(8)	Código de la categoría	Si	Si	No	Llave primaria, Único	
DESCRIPCION	varchar(40)	Nombre de la categoría	No	Si	No		

Tabla 8: Campos de la tabla categoria_cliente

Índices

Nombre de la llave	Tipo	Único	Nulo	Columna(s)	Comentario
CATEGORIA_CLIENTPK	Llave primaria	Si	No	CATEGORIA_CLIENTE	-
CAT_CLIENTEAKDESC	No agrupado	No	No	DESCRIPCION	

Tabla 9: Índices de la tabla categoria_cliente

Referenciada por

LICASA.CLIENTE (CATEGORIA_CLIENTE)

5. **Vendedor:** Guarda todos los vendedores con los que han trabajado en la empresa.

Columna	Tipo de dato	Descripción	Valores únicos	Es requerido	Acepta valores nulos	Restricción	Notas
VENDEDOR	varchar(4)	Código del vendedor	Si	Si	No	Llave primaria, Único	
NOMBRE	varchar(40)	Nombre del vendedor	No	Si	No		

Tabla 10: Campos de la tabla vendedor

Índices

Nombre de la llave	Tipo	Único	Nulo	Columna(s)	Comentario
VENDEDORPK	Llave primaria	Si	No	VENDEDOR	-
VENDEDORIENOMB	No agrupado	No	No	NOMBRE	

Tabla 11: Índices de la tabla vendedor

Referenciada por

LICASA.CLIENTE (VENDEDOR)
LICASA.FACTURA (VENDEDOR)

6. **Proveedor:** Almacena todos los proveedores con los que ha trabajado la empresa.

Columna	Tipo de dato	Descripción	Valores únicos	Es requerido	Acepta valores nulos	Restricción	Notas
PROVEEDOR	varchar(20)	Código del proveedor	Si	Si	No	Llave primaria, Único	
NOMBRE	varchar(150)	Nombre del proveedor	No	Si	No		

Tabla 12: Campos de la tabla proveedor

Índices

Nombre de la llave	Tipo	Único	Nulo	Columna(s)	Comentario
PROVEEDORPK	Llave primaria	Si	No	PROVEEDOR	-
PROVEEDORIENOMB	No agrupado		No	NOMBRE	

Tabla 13: Índices de la tabla proveedor

Referenciada por

LICASA.ARTICULO (PROVEEDOR)

7. **Cobrador:** Guarda los datos de los cobradores, internamente conocidos como motoristas, que han trabajado en la empresa.

Columna	Tipo de dato	Descripción	Valores únicos	Es requerido	Acepta valores nulos	Restricción	Notas
COBRADOR	varchar(4)	Código del cobrador	Si	Si	No	Llave primaria, Único	
NOMBRE	varchar(40)	Nombre del cobrador	No	Si	No		

Tabla 14: Campos de la tabla cobrador

Índices

Nombre de la llave	Tipo	Único	Nulo	Columna(s)	Comentario
COBRADORPK	Llave primaria	Si	No	COBRADOR	-
COBRADORAKNOMBRE	No agrupado	No	No	NOMBRE	

Tabla 15: Índices de la tabla cobrador

Referenciada por

LICASA.CLIENTE (COBRADOR)
LICASA.FACTURA (COBRADOR)

8. **Zona:** Almacena el catálogo que permite a la empresa segregar a sus clientes de forma geográfica y/o por la empleada (display) encargada de promocionar sus productos en sus establecimientos.

Columna	Tipo de dato	Descripción	Valores únicos	Es requerido	Acepta valores nulos	Restricción
ZONA	varchar(4)	Código de zona	Si	Si	No	Llave primaria, Único
NOMBRE	varchar(40)	Nombre de la zona. Incluye el nombre de los 14 departamentos del país y nombres de los encargados de las diferentes zonas de distribución (display).	No	Si	No	

Tabla 16: Campos de la tabla zona

Índices

Nombre de la llave	Tipo	Único	Nulo	Columna(s)	Comentario
ZONAPK	Llave primaria	Si	No	ZONA	-
ZONAIENOMB	No agrupado	No	No	NOMBRE	

Tabla 17: Índices de la tabla zona

Referenciada por

LICASA.CLIENTE (ZONA)
LICASA.FACTURA (ZONA)

9. **Bodega:** Guarda los datos de las bodegas utilizadas por la empresa como almacén de sus productos.

Columna	Tipo de dato	Descripción	Valores únicos	Es requerido	Acepta valores nulos	Restricción	Notas
BODEGA	varchar(4)	Código de la bodega	Si	Si	No	Llave primaria, Único	
NOMBRE	varchar(40)	Nombre de la bodega	No	Si	No		

Tabla 18: Campos de la tabla bodega

Índices

Nombre de la llave	Tipo	Único	Nulo	Columna(s)	Comentario
BODEGAPK	Llave primaria	Si	No	BODEGA	-
BODEGAIENOMBRE	No agrupado	No	No	NOMBRE	

Tabla 19: Índices de la tabla bodega

Referenciada por

LICASA.FACTURA_LINEA (BODEGA)

10. **Factura:** Guarda los datos relacionados al encabezado de una transacción de venta, también conocida como factura.

Columna	Tipo de dato	Descripción	Valores únicos	Es requerido	Acepta valores nulos	Restricción	Notas
TIPO_ DOCUMENTO	varchar(1)	Representa el tipo de factura que se está registrando.	No	Si	No	Llave primaria	Los valores son permitidos: D = Devolución F = Factura
FACTURA	varchar(50)	Código de la factura	No	Si	No	Llave primaria	
COBRADOR	varchar(4)	Código del cobrador	No	Si	No	Llave foránea	
ZONA	varchar(4)	Código de la zona	No	Si	No	Llave foránea	
VENDEDOR	varchar(4)	Código del vendedor	No	Si	No	Llave foránea	
CLIENTE	varchar(20)	Código del cliente	No	Si	No	Llave foránea	
MONTO_DESCUENTO1	decimal(28,8)	Descuento 1	No	No	No		
MONTO_DESCUENTO1	decimal(28,8)	Descuento 2	No	No	No		

Tabla 20: Campos de la tabla factura

Índices

Nombre de la llave	Tipo	Único	Nulo	Columna(s)	Comentario
FACTURAPK	Llave primaria	Si	No	TIPO_DOCUMENTO FACTURA	Llave primaria compuesta
FACTURAIECOBR	Llave foránea	No	No	COBRADOR	
FACTURAIEZONA	Llave foránea	No	No	ZONA	
FACTURAIEVEND	Llave foránea	No	No	VENDEDOR	
FKFACTURCLIE	Llave foránea	No	No	CLIENTE	

Tabla 21: Índices de la tabla factura

Referencias

LICASA.COBRADOR (COBRADOR)
LICASA.ZONA (ZONA)
LICASA.CLIENTE (CLIENTE)
LICASA.VENDEDOR (VENDEDOR)

11. **Factura_Linea:** Almacena los datos detallados de una transacción de venta, el artículo, cantidades y otros detalles pertinentes a la transacción.

Columna	Tipo de dato	Descripción	Valor único	Es requerido	Acepta valores nulos	Restricción
TIPO_DOCUMENTO	varchar(1)	Representa el tipo de factura que se está registrando.	No	Si	No	Llave primaria, Llave foránea
FACTURA	varchar(50)	Código de la factura a la que pertenece la línea	No	Si	No	Llave primaria, Llave foránea
LINEA	smallint	Numero correlativo de la línea de la factura	No	Si	No	Llave primaria
BODEGA	varchar(4)	Código de la bodega	No	Si	No	Llave foránea
ARTICULO	varchar(20)	Código del articulo detallado en la línea de la factura	No	Si	No	Llave foránea
PRECIO_UNITARIO	Decimal (28,8)	Precio unitario del articulo	No	Si	No	
CANTIDAD	Decimal (28,8)	Cantidad de artículos detallados en la línea	No	Si	No	
TOTAL_IMPUESTO1	Decimal (28,8)	Impuesto total aplicado a los artículos de la línea	No	Si	No	
PRECIO_TOTAL	Decimal (28,8)	Precio total de los artículos de la línea	No	Si	No	
COSTO_TOTAL	Decimal (28,8)	Costo total de los artículos de la línea	No	Si	No	
DESC_TOT_LINEA	Decimal (28,8)	Descuento aplicado a los artículos de la línea	No	Si	No	

Tabla 22: Campos de la tabla factura_linea

Índices

Nombre de la llave	Tipo	Único	Nulo	Columna(s)	Comentario
LINEA_FACTURAPK	Llave primaria	Si	No	FACTURA TIPO_DOCUMENTO LINEA	Llave primaria compuesta
FACLINFACTPDCLN	Agrupado	No	No	FACTURA TIPO_DOCUMENTO LINEA	
FACLINIEBODEGA	Llave foránea	No	No	BODEGA	
FACLINIEARTI	Llave foránea	No	No	ARTICULO	

Tabla 23: Índices de la tabla factura_linea

Referencias

LICASA.ARTICULO (ARTICULO)
LICASA.BODEGA (BODEGA)
LICASA.FACTURA (FACTURA, TIPO_DOCUMENTO)

ii. Resultados del data profiling

En lo que respecta al análisis del data set y el perfilamiento los valores almacenados en cada tabla que lo conforma, se realizaron los siguientes hallazgos:

1. Articulo

Articulo	
La tabla cuenta con aproximadamente 1061 registros en la actualidad. Los campos de interés de esta tabla son: • ARTICULO (PK) • DESCRIPCION • CLASIFICACION_1 (FK) • CLASIFICACION_2 (FK) • CLASIFICACION_3 (FK)	
Campo	**Observación**
ARTICULO (PK)	• No existen valores nulos o en blanco.
DESCRIPCION	• No existen valores nulos o en blanco. • No todos los caracteres se encuentran mayúscula. Deben convertirse todos al momento del proceso ETL
CLASIFICACION_1 (FK)	• Valores totales: 1061 • Valores nulos: 68 • Valores en blanco: 0 • Representa una referencia al campo "CLASIFICACION" de la tabla Clasificación
CLASIFICACION_2 (FK)	• Valores totales: 1061 • Valores nulos: 91 • Valores en blanco: 0 • Representa una referencia al campo "CLASIFICACION" de la tabla Clasificación
CLASIFICACION_3 (FK)	• Valores totales: 1061 • Valores nulos: 438 • Valores en blanco: 0 • Representa una referencia al campo "CLASIFICACION" de la tabla Clasificación

Tabla 24: Data profiling de la tabla artículo

2. **Clasificacion**

Clasificacion	
La tabla cuenta con aproximadamente 106 registros en la actualidad. Los campos de interés de esta tabla son: • CLASIFICACION (PK) • DESCRIPCION • AGRUPACION	
Campo	**Observación**
CLASIFICACION	• No existen valores nulos o en blanco
DESCRIPCION	• No existen valores nulos o en blanco • No todos los caracteres se encuentran mayúscula. Deben convertirse todos al momento del proceso ETL
AGRUPACION	• Tipo de dato Enumerado que permite jerarquizar los tipos de clasificación. • Los valores permitidos son: o 1 : Grupo o 2 : Subgrupo o 3 : Marca o 4 : Palos o 5 : PROV COD • Únicamente se hará uso del 1- 3

Tabla 25: Data profiling de la tabla clasificación

3. **Cliente**

Cliente	
La tabla cuenta con aproximadamente 2059 registros en la actualidad. Los campos de interés de esta tabla son: • CLIENTE (PK) • NOMBRE • CATEGORIA_CLIENTE • ZONA	
Campo	**Observación**
CLIENTE (PK)	• No existen valores nulos o en blanco
NOMBRE	• No existen valores nulos o en blanco • No todos los caracteres se encuentran mayúscula. Deben convertirse todos al momento del proceso ETL
CATEGORIA_CLIENTE	• No existen valores nulos o en blanco y tampoco se permiten.
ZONA	• No existen valores nulos o en blanco y tampoco se permiten.

Tabla 26: Data profiling de la tabla cliente

4. **Categoria_Cliente**

Categoria_Cliente	
La tabla cuenta con aproximadamente 34 registros en la actualidad. Los campos de interés de esta tabla son: • DESCRIPCION	
Campo	**Observación**
DESCRIPCION	• No existen valores nulos o en blanco • No todos los caracteres se encuentran mayúscula. Deben convertirse todos al momento del proceso ETL • Posee 14 registros donde la CATEGORIA_CLIENTE no posee letras únicamente números

Tabla 27: Data profiling de la tabla categoria_cliente

5. **Vendedor**

Vendedor	
La tabla cuenta con aproximadamente 25 registros en la actualidad. Los campos de interés de esta tabla son: • VENDEDOR (PK) • NOMBRE	
Campo	**Observación**
VENDEDOR (PK)	• No existen valores nulos o en blanco
NOMBRE	• No existen valores nulos o en blanco • No todos los caracteres se encuentran mayúscula. Deben convertirse todos al momento del proceso ETL

Tabla 28: Data profiling de la tabla vendedor

6. **Proveedor**

Proveedor	
La tabla cuenta con aproximadamente 566 registros en la actualidad. Los campos de interés de esta tabla son: • PROVEEDOR (PK) • NOMBRE	
Campo	**Observación**
PROVEEDOR (PK)	• No existen valores nulos o en blanco.
NOMBRE	• No existen valores nulos o en blanco. • No todos los caracteres se encuentran mayúscula. Deben convertirse todos al momento del proceso ETL

Tabla 29: Data profiling de la tabla proveedor

7. **Cobrador**

Cobrador	
La tabla cuenta con aproximadamente 43 registros en la actualidad. Los campos de interés de esta tabla son: • COBRADOR (PK) • NOMBRE	
Campo	**Observación**
COBRADOR (PK)	• No existen valores nulos o en blanco.
NOMBRE	• No existen valores nulos o en blanco. • No todos los caracteres se encuentran mayúscula. Deben convertirse todos al momento del proceso ETL

Tabla 30: Data profiling de la tabla cobrador

8. **Zona**

Zona	
La tabla cuenta con aproximadamente 36 registros en la actualidad. Los campos de interés de esta tabla son: • NOMBRE	
Campo	**Observación**
NOMBRE	• No existen valores nulos o en blanco • No todos los caracteres se encuentran mayúscula. Deben convertirse todos al momento del proceso ETL • Posee 14 registros donde la ZONA no posee letras únicamente números. Estos representan los 14 departamentos

Tabla 31: Data profiling de la tabla zona

9. **Bodega**

Bodega	
La tabla cuenta con aproximadamente 8 registros en la actualidad. Se utilizarán únicamente los registros relacionados al proceso de venta. Los campos de interés de esta tabla son: • BODEGA (PK) • NOMBRE	
Campo	**Observación**
BODEGA (PK)	• No existen valores nulos o en blanco. • Posee 1 registro donde el valor del campo no posee números, solo letras.
NOMBRE	• No existen valores nulos o en blanco, ni se permiten.

Tabla 32: Data profiling de la tabla bodega

10. Factura

Factura	
La tabla cuenta a la actualidad con aproximadamente 100,198 registros La tabla está compuesta por una llave foránea compuesta los cuales son los atributos de interés para nuestro análisis: • FACTURA • TIPO_DOCUMENTO • MONTO_DESCUENTO1 • MONTO_DESCUENTO2 Existen registros que no poseen línea, pero si esta la factura	
Campo	**Observación**
FACTURA	• Todos los campos se encuentran completos • No existen valores nulos o en blanco
TIPO_DOCUMENTO	• Todos los campos se encuentran completos • No existen valores nulos o en blanco • Realizar la transformación del dato al realizar la carga, en la base transaccional se maneja “D” y “F”, donde: o D = Devolución o F = Factura
MONTO_DESCUENTO1	• No cuenta con valores nulos • Su valores, hasta la fecha, van desde 0 hasta 342.43
MONTO_DESCUENTO2	• No cuenta con valores nulos. • No presenta ningún valore diferente de 0. • El campo parece no ser utilizado

Tabla 33: Data profiling de la tabla factura

11. Factura_Linea

Factura_Linea	
La tabla cuenta a la actualidad con aproximadamente 870,598 registros La tabla está compuesta por una llave foránea compuesta los cuales son los atributos de interés para nuestro análisis: • LINEA • FACTURA • TIPO_DOCUMENTO	
Campo	**Observación**
PRECIO_UNITARIO	• Todos los campos se encuentran completos • No existen valores nulos o en blanco
CANTIDAD	• Todos los campos se encuentran completos • No existen valores nulos o en blanco
PRECIO_TOTAL	• Todos los campos se encuentran completos • No existen valores nulos o en blanco • No se están considerando las cantidades devueltas, solamente el descuento
TOTAL_IMPUESTO1	• Todos los campos se encuentran completos • No existen valores nulos o en blanco
COSTO_TOTAL	• Todos los campos se encuentran completos • No existen valores nulos o en blanco
DESC_TOT_LINEA	• Todos los campos se encuentran completos • No existen valores nulos o en blanco

TIPO_DOCUMENTO	• Todos los campos se encuentran completos
	• No existen valores nulos o en blanco
	• Realizar la transformación del dato al realizar la carga, en la base transaccional se maneja “D” y “F”, donde:
	  ○ D = Devolución
	  ○ F = Factura

Tabla 34: Data profiling de la tabla factura_linea

iii. Especificación de las necesidades analíticas que el modelo dimensional propuesto solventará

Mediante entrevistas con diferentes usuarios con conocimiento del negocio, funcionamiento de la herramienta ERP Softland e interesados en la solución del modelo de Data Warehouse, logramos identificar algunas necesidades analíticas con las que se pretende brindar un apoyo al proceso de análisis sobre la situación de la empresa en el tiempo desde la perspectiva de sus clientes y vendedores. Los usuarios manifestaron el interés de contar con tres reportes, los cuales se detallan a continuación:

Reporte de ventas: Informe donde se desea conocer valores montos totales sobre sus transacciones. En el desean conocer los siguientes valores:

- ✓ La utilidad total de las ventas
- ✓ El costo total de las ventas
- ✓ El total de lo vendido
- ✓ La cantidad de ventas hechas
- ✓ El total de artículos vendidos
- ✓ El total de impuesto de las ventas
- ✓ El total de descuento
- ✓ El producto más y menos vendido
- ✓ El tiempo de recompra de cada cliente

Reporte comparativo de ingresos y crecimiento entre periodos: Este informe busca brindar información acerca de incrementos en los montos vendidos para determinar si esos posibles incrementos en los ingresos generados se deben al aumento de precios por factores económicos externos, como la inflación, o a un crecimiento el número de artículos vendidos en los diferentes periodos.

Lo que se busca mostrar son los montos vendidos y la cantidad de artículos vendidos por periodos iguales a un año para que los usuarios interesados puedan observar los valores y comparar si hay un aumento en los ingresos y si se debe a una mayor cantidad de artículos vendidos. Dado que la empresa vende todos sus productos de forma individual, la unidad de medida para cada uno es la misma: Unidad.

Reporte de rendimiento de vendedores: El informe de rendimiento de los vendedores busca conocer la situación de las ventas que cada empleado genera para poder conocer el avance en el tiempo hacia la meta planteada por la empresa año con año y también para brindar incentivos a los empleados según su rendimiento. Lo que se desea conocer son el margen de utilidad que cada empleado aporta con relación a la meta establecida para cada uno.

La meta del periodo se desea calcular como el promedio de lo vendido en los últimos 6 meses por cada vendedor más el 20% de ese valor en relación al rango de fechas que se desea evaluar para, posteriormente, utilizar dicho valor y calcular el margen de utilidad que representa cada venta hecha en el periodo en cuestión.

La fórmula a utilizar seria: $Margen = \frac{TotalVendido}{MetaDelPeriodo} \times 100$

iv. Modelo dimensional

Paso 1 - Seleccionar proceso del negocio
Proceso de ventas

Tabla 35:Proceso del negocio

Paso 2 - Definir el nivel de granularidad:	
Granularidad deseada	• Quiere ver las ventas en función de articulo, Bodega, cliente, vendedor, Cobrador, Proveedor, por día/semana mes/semana año/mes/trimestre/ semestre/año
Granularidad posible	• Se puede ver las ventas en función de Articulo, Bodega, Cliente, Vendedor, Cobrador, Proveedor, por día/semana mes/semana año/mes/trimestre/semestre/año

Tabla 36: Definición de granularidad

Paso 3 - Identificar las dimensiones	
Dimensión	**Atributos**
Artículo	• ArticuloID (BK) • ArticuloKey (SK) • ArticuloDescripcion • Clasificacion1Grupo • Clasificacion2SubGrupo • Clasificacion3Marca
Cliente	• ClienteID (BK) • ClienteKey (SK) • ClienteNombre • ClienteCategoria • Zona
Bodega	• BodegaID (BK) • BodegaKey (SK) • BodegaNombre
Vendedor	• VendedorID (BK) • VendedorKey (SK) • VendedorNombre • MetaDelPeriodo
Fecha	• FechaID (BK) • FechaKey (SK) • Dia • SemanaMes • SemanaAnio • Mes • Trimestre • Semestre • Anio • FinDeMes • DiaNombre • MesNombre

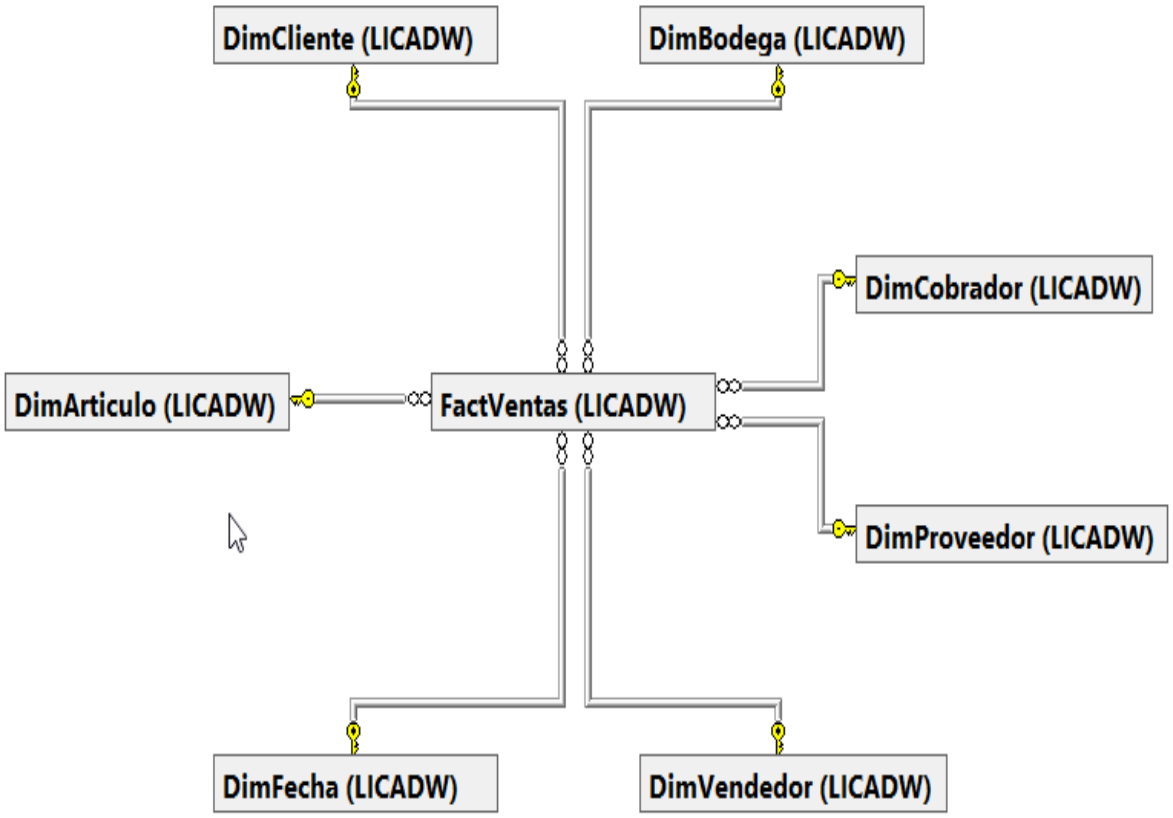
Cobrador	• CobradorID (BK) • CobradorKey (SK) • CobradorNombre
Proveedor	• ProveedorID (BK) • ProveedorKey (SK) • ProveedorNombre

Tabla 37: Definición de dimensiones

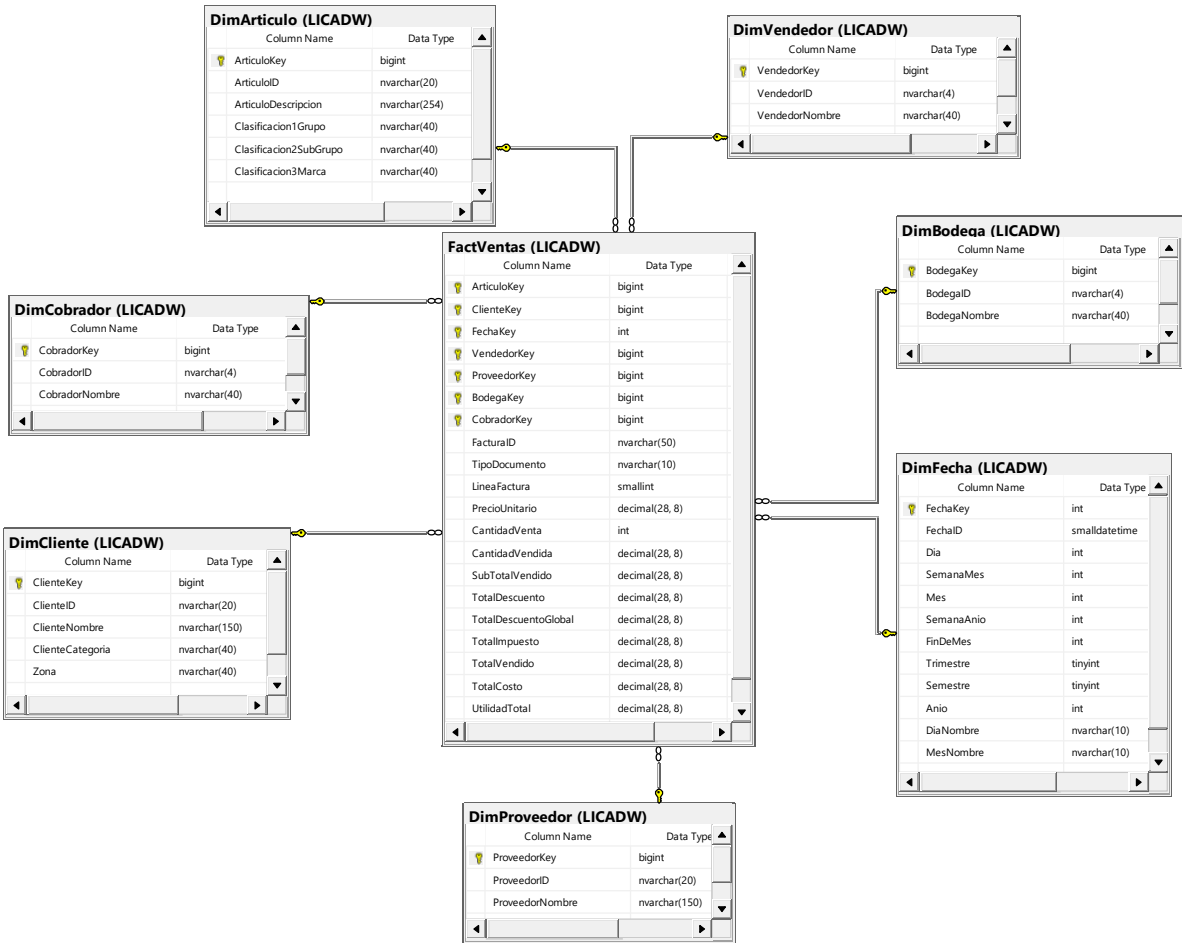
Paso 4 - Identificar métricas
• Quiero saber la utilidad total. (Total vendido - Costo total) • Quiero saber el costo total (Valor puntual de la base transaccional) • Quiero saber la cantidad de ventas • Quiero saber cuál es el articulo más vendido. (Contabilizar la cantidad de artículos y un Max) • Quiero saber cuál es el articulo menos vendido. (Contabilizar la cantidad de artículos y un Min) • Quiero saber el total de lo vendido. (Sumarizar total de lo vendido) • Quiero saber el total de impuesto. (Sumarizar impuesto total) • Quiero saber el total de descuento. (Sumarizar el descuento por producto y global) • Quiero saber el tiempo de recompra por cliente. (Calculo propio. Ejemplo. Promedio que pasa entre cada transacción. Se calcula on runtime por su dinamismo, por consiguiente, no tendrá presencia física en el DW) • Quiero cantidad de artículos vendidos. (Contabilizar la cantidad de artículos • Quiero calcular el monto de venta meta por vendedor. (Este valor no estará presente en el DW dado su dinamismo, ya que depende de parámetros de fechas. Será calculado en tiempo de ejecución.) • Quiero conocer el margen de utilidad que representa la transacción (Este valor no estará presente en el DW dado su dinamismo, ya que depende del monto de venta meta del vendedor. Será calculado en tiempo de ejecución.)

Tabla 38: Métricas

v. Diagrama Entidad-Relación del Data Warehouse del proceso de ventas



vi. Diagrama de estrella del modelo dimensional



Importante

- Para visualizar el diagrama en mejor resolución puede remitirse al Anexo 1: Diagrama de estrella del data warehouse del proceso de ventas.

- Para verificación de código fuente del proyecto y un mayor detalle del proceso de ETL desarrollado en Talend Open Studio, hacer clic en el siguiente enlace Repositorio en GitHub

vii. Mapping de las tablas del modelo dimensional

1. DimCliente

Tipo de tabla	Dimensión
Nombre	DimCliente
Nombre visual	DimCliente
Descripción	La dimensión DimCliente, incluye todos los clientes de la empresa
Usado en el esquema	LICADW

Tabla 39: Dimensión cliente

Objetivo								Fuente							
Columna	**Nombre visual**	**Descripción**	**Grupo de atributo**	**Tipo de dato**	**Tamaño**	**Precisión**	**Llave primaria**	**NULL?**	**Tipo de SCD**	**Sistema**	**Esquema**	**Tabla**	**Campo**	**Tipo de datos**	**Reglas ETL**
ClienteKey	Cliente Key	Surrogate primary key	Identificador	bigint			PK	N	0						
ClienteID	Identificador Cliente	Business Key	Identificador	nvarchar	20			N	1	SOFTLAND	LICASA	Cliente	Cliente	varchar (20)	
ClienteNombre	Nombre Cliente	Nombre correspondiente del cliente	Nombre	nvarchar	150			N	1	SOFTLAND	LICASA	Cliente	Nombre	varchar (150)	Convertir a mayúscula
ClienteCategoria	Categoría Cliente	Categoría a la que pertenece el cliente	Nombre	nvarchar	40			N	1	SOFTLAND	LICASA	Categoria_Cliente	Descripcion	varchar (40)	Convertir a mayúscula
Zona	Zona	Nombre del departamento o display correspondiente	Nombre	nvarchar	40			N	1	SOFTLAND	LICASA	Zona	Nombre	varchar (40)	Convertir a mayúscula

Tabla 40: Mapeo de la dimensión cliente

2. DimBodega

Tipo de tabla	Dimensión
Nombre	DimBodega
Nombre visual	DimBodega
Descripción	La dimensión DimBodega, incluye todas las bodegas de la empresa
Usado en el esquema	LICADW

Tabla 41: Dimensión bodega

Objetivo										Fuente					
Columna	**Nombre visual**	**Descripción**	**Grupo de atributo**	**Tipo de dato**	**Tamaño**	**Precisión**	**Llave primaria**	**NULL?**	**Tipo de SCD**	**Sistema**	**Esquema**	**Tabla**	**Campo**	**Tipo de datos**	**Reglas ETL**
BodegaKey	Bodega Key	Surrogate primary key	Identificador	bigint			PK	N	0						
BodegaID	Identificador Bodega	Business Key	Identificador	nvarchar	4			N	1	SOFTLAND	LICASA	Bodega	Bodega	varchar (4)	
BodegaNombre	Nombre Bodega	Nombre correspondiente de bodega	Nombre	nvarchar	40			N	1	SOFTLAND	LICASA	Bodega	Nombre	varchar (40)	Convertir a mayúscula

Tabla 42: Mapeo de la dimensión bodega

3. DimCobrador

Tipo de tabla	Dimensión
Nombre	DimCobrador
Nombre visual	DimCobrador
Descripción	La dimensión DimCrobrador, incluye todos los cobradores de la empresa
Usado en el esquema	LICADW

Tabla 43: Dimensión cobrador

Objetivo										Fuente					
Columna	**Nombre visual**	**Descripción**	**Grupo de atributo**	**Tipo de dato**	**Tamaño**	**Precisión**	**Llave primaria**	**NULL?**	**Tipo de SCD**	**Sistema**	**Esquema**	**Tabla**	**Campo**	**Tipo de datos**	**Reglas ETL**
CobradorKey	Cobrador Key	Surrogate primary key	Identificador	bigint			PK	N	0						
CobradorID	Identificador Cobrador	Business Key	Identificador	nvarchar	4			N	1	SOFTLAND	LICASA	Cobrador	cobrador	varchar(4)	
CobradorNombre	Nombre Cobrador	Nombre correspondiente del cobrador	Nombre	nvarchar	40			N	1	SOFTLAND	LICASA	Cobrador	nombre	varchar(40)	Convertir a mayúscula

Tabla 44: Mapeo de la dimensión cobrador

4. DimProveedor

Tipo de tabla	Dimensión
Nombre	DimProveedor
Nombre visual	DimProveedor
Descripción	La dimensión DimProveedor, incluye todos los proveedores de la empresa
Usado en el esquema	LICADW

Tabla 45: Dimensión proveedor

Objetivo										Fuente					
Columna	**Nombre visual**	**Descripción**	**Grupo de atributo**	**Tipo de dato**	**Tamaño**	**Precisión**	**Llave primaria**	**NULL?**	**Tipo de SCD**	**Sistema**	**Esquema**	**Tabla**	**Campo**	**Tipo de datos**	**Reglas ETL**
ProveedorKey	Proveedor Key	Surrogate primary key	Identificador	int			PK	N	0						
ProveedorID	Identificador Proveedor	Business Key	Identificador	nvarchar	20			N	1	SOFTLAND	LICASA	Proveedor	Proveedor	varchar (20)	
ProveedorNombre	Nombre Proveedor	Nombre correspondiente del proveedor	Nombre	nvarchar	150			N	1	SOFTLAND	LICASA	Proveedor	Proveedor	varchar (150)	Convertir a mayúscula

Tabla 46: Mapeo de la dimensión proveedor

5. DimVendedor

Tipo de tabla	Dimensión
Nombre	DimVendedor
Nombre visual	DimVendedor
Descripción	La dimensión DimProveedor, incluye todos los proveedores de la empresa
Usado en el esquema	LICADW

Tabla 47: Dimensión vendedor

Objetivo										Fuente					
Columna	**Nombre visual**	**Descripción**	**Grupo de atributo**	**Tipo de dato**	**Tamaño**	**Precisión**	**Llave primaria**	**NULL?**	**Tipo de SCD**	**Sistema**	**Esquema**	**Tabla**	**Campo**	**Tipo de datos**	**Reglas ETL**
VendedorKey	Vendedor Key	Surrogate primary key	Identificador	bigint			PK	N	0						
VendedorID	Identificador Vendedor	Business Key	Identificador	nvarchar	4			N	1	SOFTLAND	LICASA	Vendedor	Vendedor	varchar (4)	
VendedorNombre	Nombre Vendedor	Nombre del proveedor	Nombre	nvarchar	40			N	1	SOFTLAND	LICASA	Vendedor	Nombre	varchar (150)	Convertir a mayúscula

Tabla 48: Mapeo de la dimensión vendedor

6. DimFecha

Tipo de tabla	Dimensión
Nombre	DimFecha
Nombre visual	DimFecha
Descripción	La DimFecha, contiene el detalle de todas las fechas desde el 2016 al 2025
Usado en el esquema	LICADW

Tabla 49: Dimensión fecha

Objetivo										Fuente						
Columna	**Nombre visual**	**Descripción**	**Grupo de atributo**	**Tipo de dato**	**Tamaño**	**Precisión**	**Llave primaria**	**NULL?**	**Tipo de SCD**	**Sistema**	**Esquema**	**Tabla**	**Campo**	**Tipo de datos**	**Reglas ETL**	**Comentarios**
FechaKey	Fecha Key	Surrogate primary key	Identificador	bigint			PK	N	0							
FechaID	Identificador Fecha	Business Key	Identificador	smalldatetime				N	0							
Dia	Dia	Indica el día correspondiente	Nombre	int				N	0							
SemanaMes	Semana Mes	Indica la semana en cuestión del mes	Nombre	int				N	0							
Mes	Mes	Indica el mes correspondiente	Nombre	int				N	0							
SemanaAnio	Semana Anio	Indica la semana en cuestión del año	Nombre	int				N	0							
FinDeMes	Fin de mes	Indica el día de fin de mes	Nombre	int				N	0							
Trimestre	Trimestre	Indica el trimestre del año	Nombre	tinyInt				N	0							
Semestre	Semestre	Indica el semestre del año	Nombre	tinyInt				N	0							
Anio	Anio	Indica el año	Nombre	int				N	0							
DiaNombre	Nombre dia	Indica el nombre correspondiente del día de la semana	Nombre	nvarchar	10			N	0							
MesNombre	Nombre mes	Indica el nombre correspondiente del mes	Nombre	nvarchar	10			N	0							

Tabla 50: Mapeo de la dimensión fecha

7. DimArticulo

Tipo de tabla	Dimensión
Nombre	DimArticulo
Nombre visual	DimArticulo
Descripción	La dimensión DimArticulo, incluye todos los articulos y la clasificación 1 - Grupo, 2 - SubGrupo y 3 - Marca
Usado en el esquema	LICADW

Tabla 51: Dimensión artículo

Objetivo									
Columna	**Nombre visual**	**Descripción**	**Grupo de atributo**	**Tipo de dato**	**Tamaño**	**Precisión**	**Llave primaria**	**NULL?**	**Tipo de SCD**
ArticuloKey	Artículo Key	Surrogate primary key	Identificador	bigint			PK	N	0
ArticuloID	Identificador Articulo	Business Key	Identificador	nvarchar	20			N	1
ArticuloDescripcion	Descripción Articulo	Nombre correspondiente del articulo	Nombre	nvarchar	254			N	1
Clasificacion1Grupo	Clasificación 1 Grupo	Clasificación - Grupo a la que pertenece el articulo	Nombre	nvarchar	40			N	1
Clasificacion2SubGrupo	Clasificación 2 Subgrupo	Clasificación - Subgrupo a la que pertenece el articulo	Nombre	nvarchar	40			N	1
Clasificacion3Marca	Clasificación 3 Marca	Clasificación - Marca del articulo	Nombre	nvarchar	40			N	1

Tabla 52: Mapeo de la dimensión artículo

Fuente						
Sistema	**Esquema**	**Tabla**	**Campo**	**Tipo de datos**	**Reglas ETL**	**Comentarios**
SOFTLAND	LICASA	Articulo	Articulo	varchar (20)		
SOFTLAND	LICASA	Articulo	Descripcion	varchar (254)	Convertir a mayúscula	
SOFTLAND	LICASA	Clasificacion	Descripcion	varchar (40)	Convertir a mayúscula	El valor debe buscarse utilizando la llave foránea y la agrupación respectiva (1)
SOFTLAND	LICASA	Clasificacion	Descripcion	varchar (40)	Convertir a mayúscula	El valor debe buscarse utilizando la llave foránea y la agrupación respectiva (2)
SOFTLAND	LICASA	Clasificacion	Descripcion	varchar (40)	Convertir a mayúscula	El valor debe buscarse utilizando la llave foránea y la agrupación respectiva (3)

8. FactVentas

Tipo de tabla	Fact – tabla de hechos
Nombre	FactVentas
Nombre visual	FactVentas
Descripción	FactVentas captura transacciones de venta al nivel de línea de factura

Tabla 53: Tabla de hechos ventas

Objetivo										
Columna	**Nombre visual**	**Descripción**	**Tipo de dato**	**Tamaño**	**Precisión**	**Tipo de llave**	**Llave foránea a**	**NULL?**	**Valor por defecto**	**Valores de ejemplo**
ArticuloKey	ArticuloKey	Llave a dimensión DimArticulo	bigint			FK	DimArticulo.DimArticuloKey	N		1, 2, 3
ClienteKey	ClienteKey	Llave a dimensión DimCliente	bigint			FK	DimCliente.DimClienteKey	N		1, 2, 3
FechaKey	FechaKey	Llave a dimensión DimFecha	bigint			FK	DimFecha.FechaKey	N		20220629, 20220601
ProveedorKey	ProveedorKey	Llave a dimensión DimProveedor	bigint			FK	DimProveedor.ProveedorKey	N		1, 2, 3
BodegaKey	BodegaKey	Llave a dimensión DimBodega	bigint			FK	DimBodega.BodegaKey	N		1, 2, 3
VendedorKey	VendedorKey	Llave a dimensión DimVendedor	bigint			FK	DimVendedor.VendedorKey	N		1, 2, 3
CobradorKey	CobradorKey	Llave a dimensión CobradorKey	bigint			FK	DimCobrador.CobradorKey	N		1, 2, 3
FacturaID	Identificador Factura	Código de la factura a la que pertenece la línea	nvarchar	20		PK		N		17DS000F09403, 1DS000C0863
TipoDocumento	Tipo de documento	Representa el tipo de factura que se está registrando.	nvarchar	10		PK		N		Devoluciónón, Factura
LineaFactura	Linea	Numero correlativo de la línea de la factura	smallint			PK		N		1, 2, 3
PrecioUnitario	Precio Unitario	Precio unitario del articulo	decimal		28, 8			N		
CantidadVendida	Cantidad Vendida	Cantidad de artículos detallados en la línea	int					N		
TotalVendido	Total Vendido	Precio total de los artículos de la línea	decimal		28, 8			N		
TotalImpuesto	Total Impuesto	Impuesto total aplicado a los artículos de la línea	decimal		28, 8			N		
TotalCosto	Total Costo	Costo total de los artículos de la línea	decimal		28, 8			N		
TotalDescuentoProducto	Total Descuento Producto	Descuento aplicado a los artículos de la línea	decimal		28, 8			N		
TotalDescuentoGlobal	Total Descuento Global	Descuento aplicado sobre la venta per se	decimal		28,8			N		
UtilidadTotal	Utilidad Total	Resultado de restar el costo total del precio total	decimal		28, 8			N		
CantidadVenta	Cantidad Venta	Campo utilizado para contabilizar el total de ventas	int					N		24, 10, 14, 3
SubTotalVendido	Subtotal Vendido	Subtotal vendido antes de aplicar descuentos	decimal		28, 8					

Tabla 54: Mapeo de la tabla de hechos ventas

Fuente						
Sistema	**Esquema**	**Tabla**	**Campo**	**Tipo de dato**	**Reglas de extracción / transformación**	**Comentarios**
Derivado					Key lookup de Articulo.ARTICULO	
Derivado					Key lookup de Cliente.CLIENTE	
Derivado					Valor "date key" desde "Ch10 dateDim.xlsx"	
Derivado					Key lookup de Proveedor.PROVEEDOR	
Derivado					Key lookup de Bodega.BODEGA	
Derivado					Key lookup de Vendedor.VENDEDOR	
Derivado					Key lookup de Cobrador.COBRADOR	
SOFTLAND	LICASA	Factura_Linea	FACTURA	varchar(50)		
SOFTLAND	LICASA	Factura_Linea	TIPO_DOCUMENTO			D=Devolucion y F=Factura
SOFTLAND	LICASA	Factura_Linea	LINEA	smallint		
SOFTLAND	LICASA	Factura_Linea	PRECIO_UNITARIO	decimal(28, 8)		
Derivado					CANTIDAD - CATIDAD_DEVUELT	
SOFTLAND	LICASA	Factura_Linea	PRECIO_TOTAL	decimal(28, 8)		
SOFTLAND	LICASA	Factura_Linea	TOTAL_IMPUESTO1	decimal(28, 8)		
SOFTLAND	LICASA	Factura_Linea	COSTO_TOTAL	decimal(28, 8)		
SOFTLAND	LICASA	Factura_Linea	DESC_TOT_LINEA	decimal(28, 8)		
Derivado					((Monto_Descuento1 + Monto_Descueno2)/SUMA(Cantidad))*Cantidad	Este valor será un prorrateo del monto de descuento global
Derivado					PRECIO_TOTAL - COSTO_TOTAL	
Derivado					CantidadVenta = 1	Se contabilizarán todas las facturas ignorando duplicados
Derivado					PRECIO_TOTAL + DESC_TOT_LINEA	

c. Descripción de la tecnología a utilizar

Las herramientas tecnológicas que se utilizó para la construcción de la solución son las siguientes:

- **Microsoft SQL Server**

Microsoft SQL Server es un sistema de gestión de base de datos relacional, desarrollado por la empresa Microsoft. El lenguaje de desarrollo utilizado es Transact-SQL, una implementación del estándar ANSI del lenguaje SQL, utilizado para manipular y recuperar datos, crear tablas y definir relaciones entre ellas.

- **Softland**

Softland es un completo Sistema de Gestión para Pymes que permite realizar una contabilidad, gestión comercial y pago de remuneraciones de manera eficiente y controlada. Se desarrolla mediante una línea de productos de gestión para Pymes, orientado especialmente a cubrir las necesidades totales de la empresa del siglo XXI.

- **Talend Open Studio**

Talend Open Studio (TOS) es una suite que aporta un conjunto muy complejo, variado y completo de herramientas para llevar a cabo la integración de datos que se ofrece en una versión de código libre (open source). Precisamente por ello, esta es una de las herramientas de integración ETL (extract, transform, load) más utilizadas dentro del mundo Big Data; es más, es la cuarta en la lista después de Informática Powercenter, IBM InfoSphere Data stage y Oracle Data Integrator (ODI).

- **Amazon Web Services (AWS)**

Amazon Web Services (AWS) es la plataforma en la nube más adoptada y completa en el mundo, que ofrece más de 200 servicios integrales de centros de datos a nivel global. Millones de clientes, incluso las empresas emergentes que crecen más rápido, las compañías más grandes y los organismos gubernamentales líderes, están usando AWS para reducir los costos, aumentar su agilidad e innovar de forma más rápida.

- **Amazon Redshift**

AWS Redshift es un almacén de datos rápido y completamente administrado que permite analizar todos los datos empleando de forma sencilla y rentable SQL estándar y las herramientas de inteligencia empresarial (BI) existentes. Esta herramienta ofrece la oportunidad de ejecutar consultas analíticas complejas en petabytes de datos estructurados, utilizando una sofisticada optimización de consultas, almacenamiento en columnas en discos locales de alto desempeño y ejecución masiva de consultas paralelas. La mayoría de los resultados se producen en segundos.

- **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.

- **Amazon IAM**

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

- **Power BI**

Power BI es un conjunto de herramientas de análisis empresarial que pone el conocimiento al alcance de toda la organización. Power BI, como solución integrada en Office 365, permite la conexión a cientos de orígenes de datos, la preparación de datos simplificada, generación de análisis ad hoc. Además, esta herramienta de Business Intelligence permite dar vida a tus datos con los paneles e informes dinámicos

d. Diagrama arquitectónico de la solución

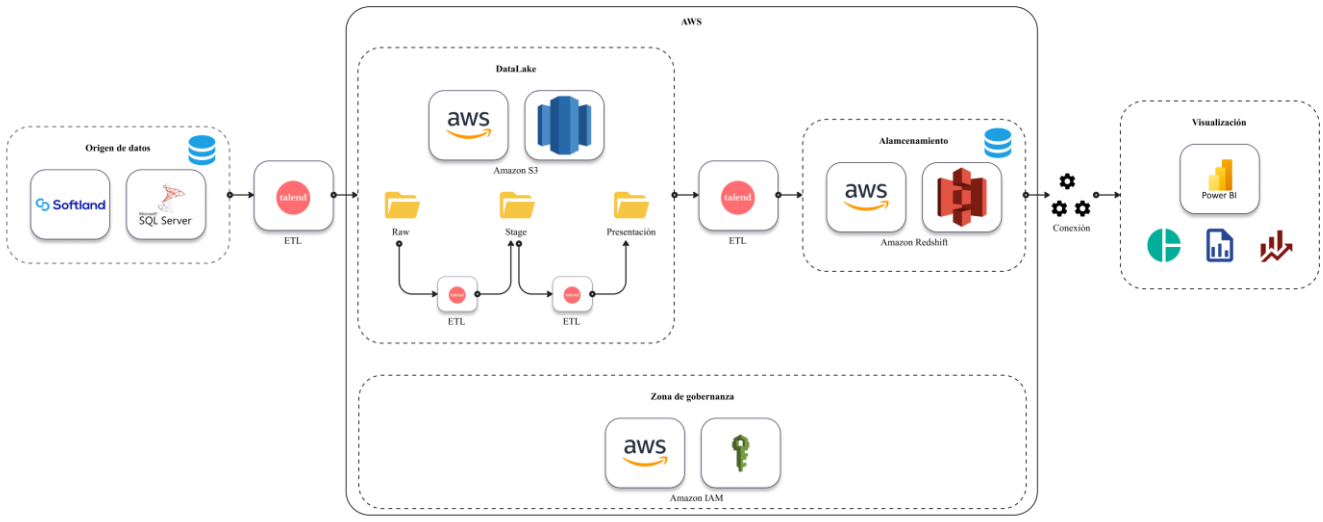


Ilustración 10: Diagrama arquitectónico de la solución

e. Descripción de cada componente de la solución

Origen de datos: Los orígenes de datos para un Data Warehouse pueden obtener de diversas fuentes como bases de datos, datos transaccionales, archivos, sistemas operativos, entre otras fuentes de datos, para este caso la empresa LICA S.A de C.V utilizan los siguientes componentes

- **SOFTALND**: La empresa LICA S.A de C.V utiliza este sistema de planificación de recursos empresariales (ERP, por sus siglas en inglés) como herramienta para registrar y almacenar datos transaccionales e históricos relacionados con el proceso de ventas y facturación.

- **SQL SERVER**: SOFTLAND utiliza SQL SERVER como motor de base de datos. En esta base de datos se almacenan los datos históricos del proceso de ventas y facturación de la empresa y se aplican las tareas iniciales de procesamiento ETL (Extracción, Transformación, Carga) para poder utilizarlos en la solución del Data Warehouse.

AWS: Dada que es una plataforma de Cloud Computing, AWS ofrece una amplia gama de servicios de computación, almacenamiento, bases de datos y herramientas de análisis de datos, AWS nos brinda diversos beneficios los cuales son: escalabilidad, Integración con herramientas de análisis de datos, bajos costos, flexibilidad y fiabilidad, para el desarrollo del proyecto se utilizaron los siguientes componentes:

- **AMAZON S3:** La herramienta brindada por AWS nos permite almacenar y procesar grandes cantidades de datos.
  En Amazon S3 se crearon las zonas para almacenar y procesar, estas zonas son las siguientes: RAW, STAGE y PRESENTATION

- **RAW:** En la zona RAW se llevó a cabo el almacenamiento de los datos crudos es decir los datos origen obtenidos del insumo que nos brinda el software TALEND, esto se realizó utilizando tareas de ETL.

- **STAGE:** La zona SATEGE se utilizó para el almacenamiento y procesamiento de los datos preparándolos para llevar a cabo la carga hacia la zona de PRESENTACIÓN, en la zona STAGE es donde se aplicaron diversas tareas sobre los datos origen como eliminar duplicados, realizar cálculos, definir formato de los datos.

- **PRESENTATIÓN:** En la zona de presentación es donde almacenamos los datos procesados y limpios, los cuales se utilizarán para el análisis y toma de decisiones.

  Estos datos se cargan en AWS REDSHIFT a través de tareas de ETL y se utilizan para crear visualizaciones de los datos y facilitar su análisis y comprensión.

- **AMAZON REDSHIFT:** Es una base de datos en cloud donde se llevó a cabo la implementación del modelo dimensional permitiéndonos obtener gracias a su estructura columnar un muy alto rendimiento y facilitando la integración con la herramienta de visualización POWER BI.

- **AMAZON IAM:** Para tener un mejor control en el acceso a los recursos y servicios que se utilizaron de AWS se implementó esta gestión en la seguridad, permitiéndonos crear a partir de un usuario raíz diversos roles, permisos y políticas de acceso que luego se asignaron a los usuarios correspondientes, protegiendo el Data Warehouse de accesos no autorizados y cumplir con los requisitos de seguridad.

ETL: Para llevar a cabo la carga de los datos a un Data Warehouse y realizar la preparación de los mismos, se requieren diversas tareas de ETL permitiendo llevar a cabo tareas de limpieza y validación de datos, mejorando así la calidad y la integración de los mismos. Al implementar estas tareas nos ayuda ahorrando tiempo y esfuerzo manual mejorando así la eficiencia de los procesos.

- **Talend:** La herramienta de integración y transformación de datos nos facilita realizar estas tareas de una manera interactiva ya que nos brinda una interfaz visual e intuitiva la cual disminuye la complejidad al realizar las tareas de extracción, transformación y carga de los datos.
  Para el desarrollo del presente proyecto se hizo uso al momento de la gestión de los datos crudos de la base de datos histórica llevándolos a la primera zona en S3 llamada RAW de nuestra arquitectura, de igual manera se realiza para el tratamiento de los datos de la zona Raw a STAGE, de STAGE a PRESENTATION y finalmente de PRESENTATION a REDSHIFT que es donde los datos están listos para poblar el modelo dimensional planteado y servir como insumo para llevar a cabo la representación de los datos en PowerBI.

Visualización: Una vez los datos hayan sido gestionados por medio de las herramientas y técnicas antes mencionadas, es necesario representarlos gráficamente para facilitar su análisis y comprensión.

- **PowerBI:** Es la herramienta utilizada en el presente proyecto para poder llevar a cabo la visualización de los datos, los cuales son representados por medio de un informe o Dashboard, estas visualizaciones están conformadas por diversos componentes como gráficos, tablas y datos puntuales.
  Con estos reportes se solventa la necesidad que la empresa planteó en los requerimientos correspondientes, permitiendo así tener una fácil comprensión de los resultados, detectar diversos patrones y tendencias, identificar oportunidades y problemas, sirviendo de apoyo en la toma de decisiones.

CAPÍTULO III: ESTRATEGIA DE IMPLEMENTACIÓN DE PROPUESTA DE SOLUCIÓN

a. Estrategia de implementación

La estrategia utilizada en la implementación será producto nuevo y por sustitución.

Producto nuevo con respecto a los procesos ETL y el Data Warehouse construidos para el aplicativo, ya que actualmente la empresa no cuenta con una solución de inteligencia de datos. Y por sustitución debido a que se busca automatizar la generación del reporte general de ventas, los ingresos versus el volumen de ventas y el rendimiento de los vendedores.

Para llevar a cabo la implementación se debe realizar una serie de pasos preliminares

i. Plan de capacitación.

En su mayoría de los usuarios del aplicativo tienen conocimiento técnico sobre las tecnologías utilizadas al desarrollar la solución, por lo tanto, se capacitará a los responsables de mantenimiento de las tareas ETL, mediante la realización pruebas piloto en un ambiente de desarrollo, con el objetivo de que se familiaricen con el flujo de información y los componentes utilizados.

Posteriormente se realizarán presentaciones para los usuarios gerenciales, enfocadas en el manejo de la aplicación de visualización de datos y la interacción con los reportes.

ii. Configuración de componentes AWS

A continuación, se muestra paso a paso las acciones necesarias para la puesta en marcha de la propuesta de solución.

Se hará uso de tres servicios de AWS: IAM, S3 y Redshift

Ilustración 11: Listado de servicios utilizados en AWS

Identity and Access Management (IAM)

Comenzando por el servicio IAM, se deberá crear un usuario con capacidades irrestrictas para acceder al servicio de S3 desde el ETL.

Ubicándose en el panel principal de IAM, se deberá seleccionar la opción “Usuarios”.

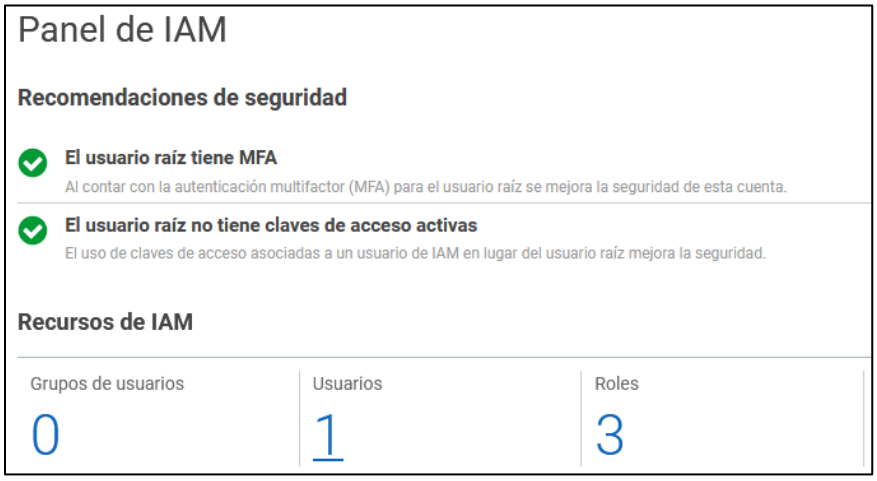


Ilustración 12: Panel principal del servicio IAM

El panel de creación de usuarios acción se debe dar clic en la opción “Agregar usuarios”.

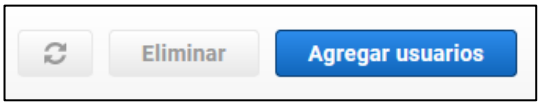


Ilustración 13: Botón agregar usuarios

AIM solicitará los detalles del usuario como el nombre y el tipo de credenciales de AWS, para el caso particular de la solución se debe elegir las credenciales del tipo “Clave de acceso mediante programación” ya que no se accede directamente a AWS sino a través de los procesos ETL.

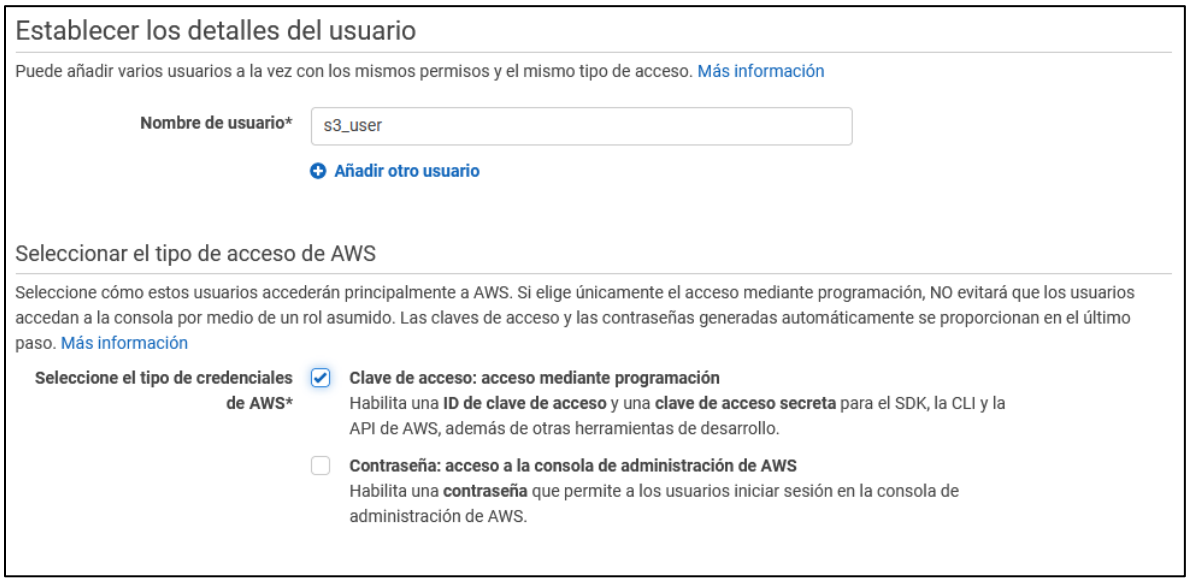


Ilustración 14: Formulario de creación de usuario IAM – Datos generales

Luego de elegir los detalles del usuario, se deberá seleccionar el permiso necesario para el usuario de S3, en este caso se hará uso de una política llamada AmazonS3FullAccess

Para ello debe seleccionar la opción denominada “Asociar directamente las políticas existentes”, luego buscar el nombre de la política antes mencionada y seleccionarla.

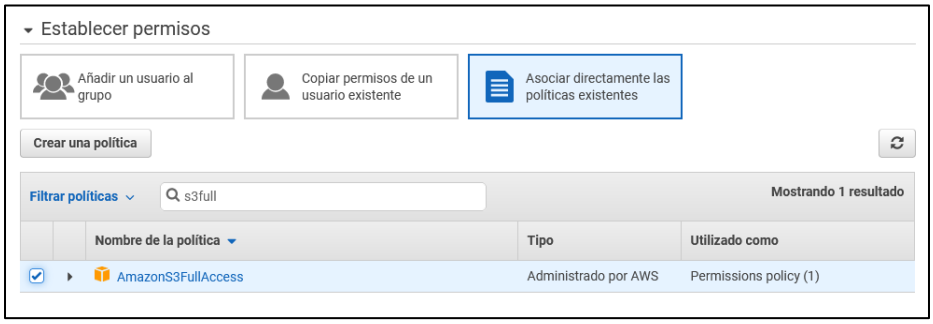


Ilustración 15: Formulario de creación de usuario IAM - Permisos

En el último paso de creación de usuario se muestra el ID y clave de acceso secreta. Se debe tomar nota de ambas ya que se utilizarán más adelante para configurar los procesos ETL.

Opcionalmente se puede descargar la información como un archivo CSV.

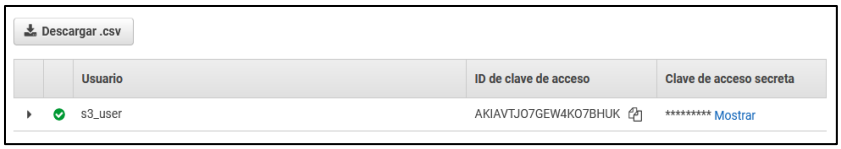


Ilustración 16: Asignación de claves de acceso IAM

Simple Storage Service (S3)

En S3 se debe crear un contenedor (Bucket) que servirá como Data Lake que almacenará todos los archivos creados como resultado de la ejecución de los procesos ETL en cada una de sus etapas.

Para crear el bucket de S3 se selecciona la opción "Crear bucket" en la página principal de S3.

Primero se debe especificar el nombre (el cual debe ser único) y la región donde será creado.

Importante: El bucket de S3 y el cluster de Redshift deben estar en la misma región. Para motivos demostrativos se ha seleccionado la región "EE.UU. Este (Norte de Virginia) us-east-1"

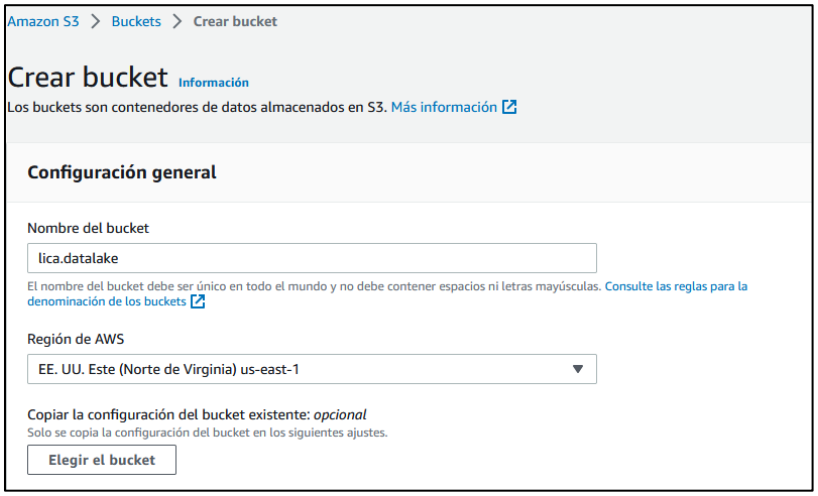


Ilustración 17: Creación de Bucket en S3 - Datos generales

Como recomendación general se deben deshabilitar las ACL.

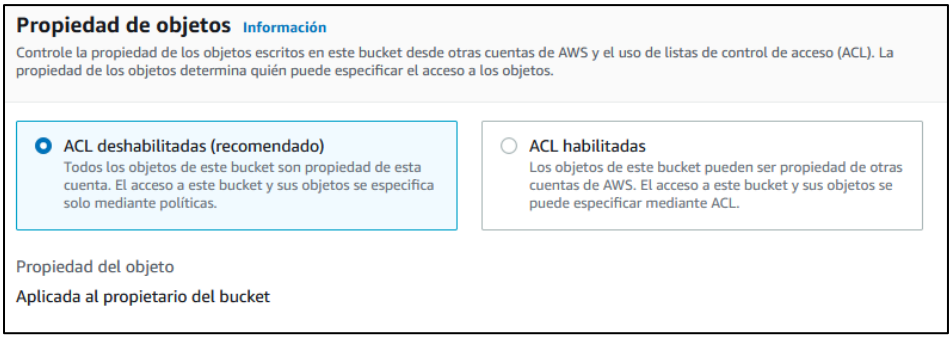


Ilustración 18: Creación de Bucket en S3 - Propiedad de objetos

También se debe bloquear el acceso público como lo muestra la siguiente imagen.

Ilustración 19: Creación de Bucket en S3 - Acceso público

Para este caso en particular no se necesita control de versiones para los objetos del bucket debido a que se consideró realizarlo dentro de los procesos ETL.

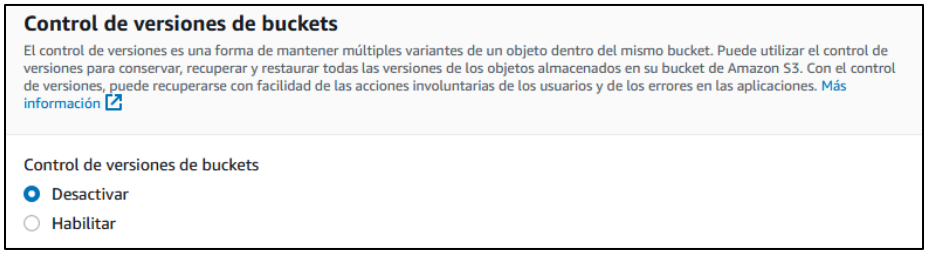


Ilustración 20: Creación de Bucket en S3 - Control de versiones

Opcionalmente, se puede configurar el cifrado de datos o la configuración de la replicación.

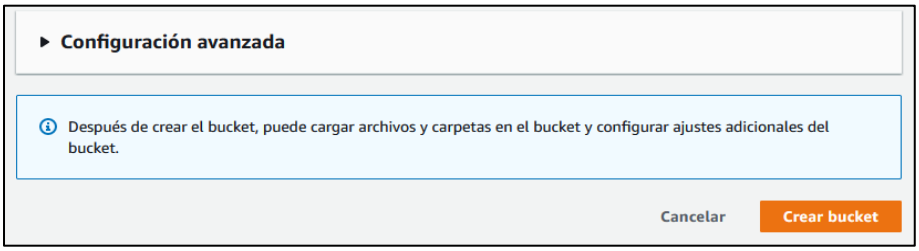


Ilustración 21: Creación de Bucket en S3 - Finalización del proceso

Una vez creado el Data Lake, se procede a crear la estructura de carpetas. Ubicándose en el panel principal de S3, se selecciona el bucket creado.

Ilustración 22: Panel principal del Bucket

Luego dando clic en la opción “Crear carpeta”, aparecerán las siguientes opciones

Nombre de la carpeta y cifrado del lado del servidor (Se dejará desactivado el cifrado para todas las carpetas).

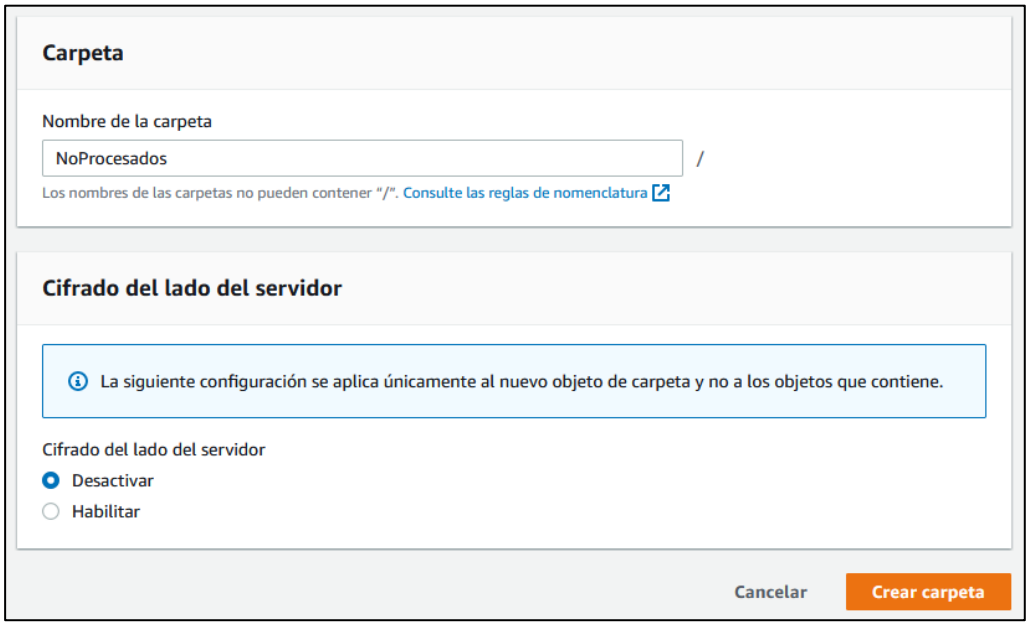


Ilustración 23: Proceso de creación de carpetas en S3

En total se crearán 2 carpetas principales “NoProcesados” y “Archivados”, las cuales contendrán tres subcarpetas “raw”, “stage” y “presentation”. Esta estructura de carpetas corresponde a las diferentes zonas de procesamiento de datos que posee la solución.

Se debe repetir el proceso de creación hasta tener una estructura como la siguiente

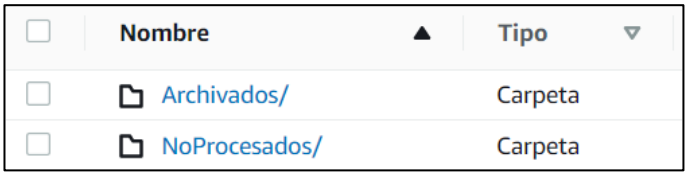


Ilustración 24: Carpetas principales del bucket

Y sus respectivas subcarpetas.

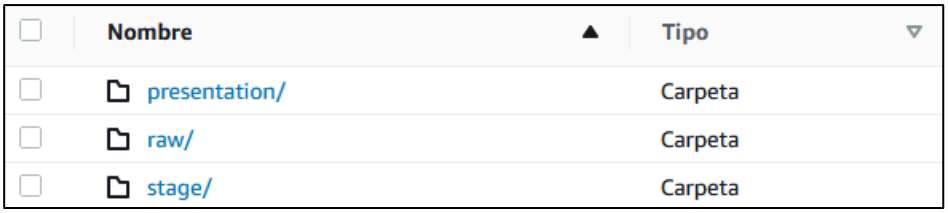


Ilustración 25: Subcarpetas correspondientes a las zonas de procesamiento

Redshift

Se creará un cluster de Redshift que servirá para alojar la base de datos del Data Warehouse para luego ser consumido por Power BI.

Para crear un nuevo clúster se debe dar clic en el botón “Crear clúster” en el panel principal de Redshift.

Ilustración 26: Cinta de opciones en la pantalla principal de Redshift

Una vez ubicado en pantalla de creación, se mostrarán las siguientes opciones.

Se debe elegir el nombre y el propósito. Si la solución está en un entorno de pruebas y es la primera vez que se creará un cluster con la cuenta de AWS, se tiene la posibilidad de crear un cluster gratuito. Para este caso en particular se elegirá la opción “Producción”.

En el apartado elegir tamaño del clúster se dejará la opción por defecto “Yo elegiré”.

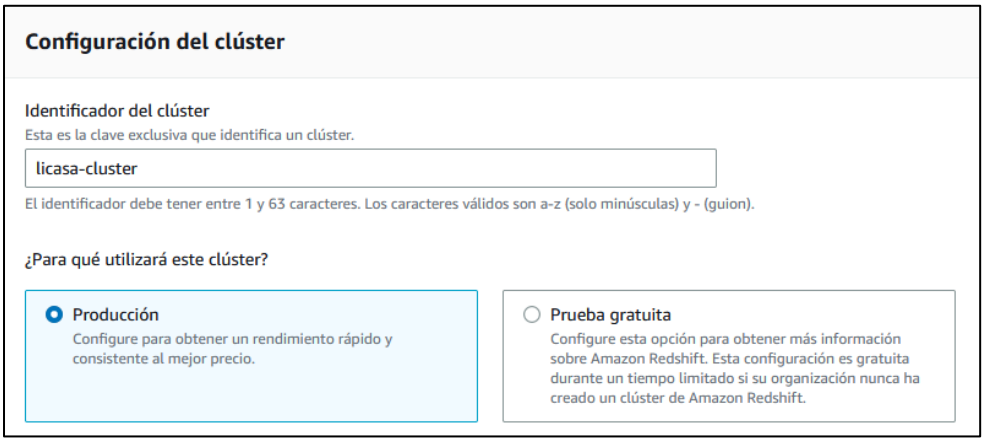


Ilustración 27: Proceso de creación del clúster - Datos generales

Se determinó que la configuración de un único nodo del tipo “dc2.large” es la adecuada para el volumen de información que produce la empresa.

Ilustración 28: Proceso de creación del clúster - Elección del tamaño y numero de nodos

El nodo cuenta con las siguientes especificaciones:

Almacenamiento: 160 GB por nodo.
Capacidad de procesamiento: 2 CPU virtuales.

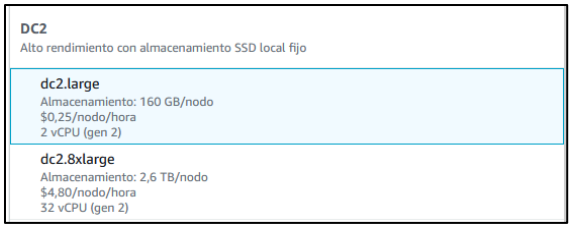


Ilustración 29: Opciones de almacenamiento del cluster

En la última sección del formulario de creación se muestra el resumen del clúster incluyendo su costo mensual.

Ilustración 30: Resumen de configuración del cluster

El ultimo es apartado la configuración de la base datos se debe introducir el nombre del usuario administrador y la contraseña del mismo.

Opcionalmente se puede utilizar una contraseña generada por Amazon.

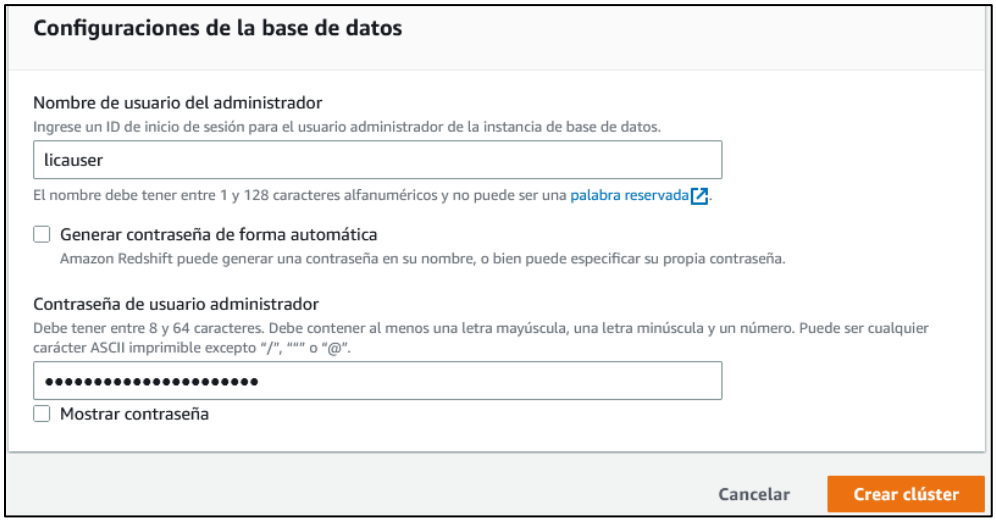


Ilustración 31: Configuración de la base de datos en el cluster

Importante: Tomar nota de las credenciales introducidas, ya que se utilizarán más adelante durante la configuración de Talend.

Una vez creado el clúster se deberá cambiar la configuración de accesibilidad pública, para ello se debe ingresar al panel de configuración del clúster.

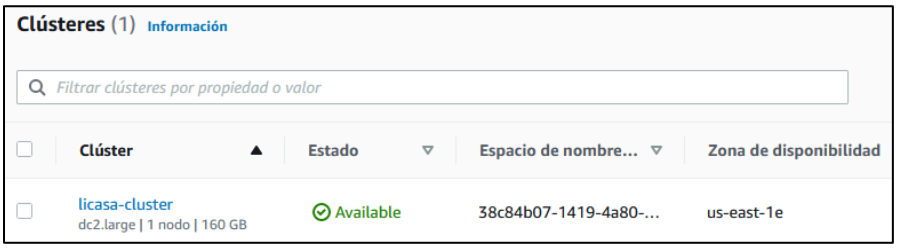


Ilustración 32: Listado de clústeres creados

Luego buscar la opción "Modificar la configuración de accesibilidad pública" del menú desplegable "Acciones".

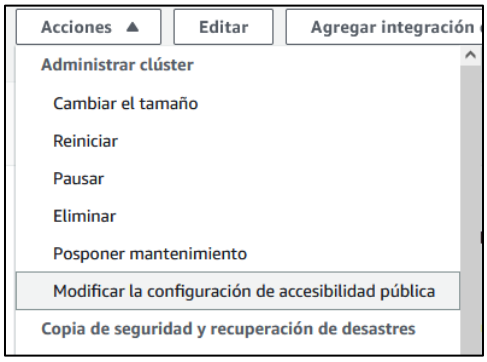


Ilustración 33: Menú de acciones del clúster

Aparecerá un cuadro de dialogo como el siguiente. Donde se debe seleccionar la casilla "Activar accesibilidad publica"

Cabe aclarar que esto no significa que cualquier persona puede realizar una conexión al clúster, sino que se puede acceder desde fuera de los componentes de AWS, siempre que se cumplan una serie de requisitos en los grupos de seguridad.

Ilustración 34: Activación de acceso público

El siguiente paso es modificar el grupo de seguridad para que permita a los procesos ETL conectarse a la base de datos para realizar el proceso de inserción / actualización de datos en el Data Warehouse.

Esta configuración se encuentra en la sección "Propiedades" del menú de opciones del clúster.

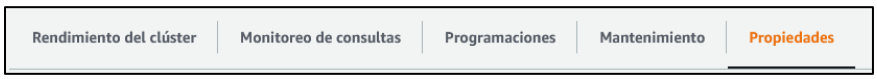


Ilustración 35: Cinta de opciones del clúster

Específicamente en el apartado “Configuración de red y seguridad”.

Ilustración 36: Configuración de red y seguridad

Al dar clic en el enlace se mostrará una página con los detalles del grupo de seguridad, luego se debe buscar la opción “Editar reglas de entrada”.

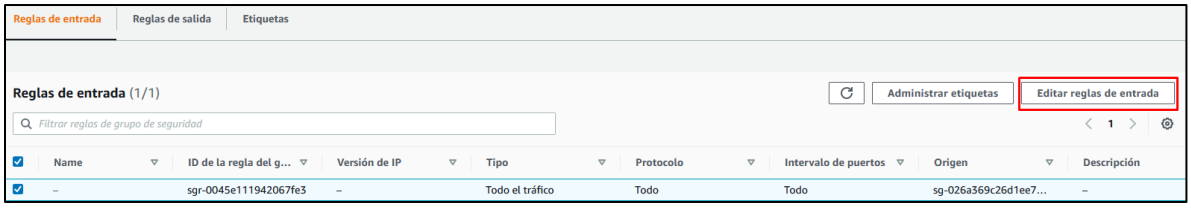


Ilustración 37: Reglas de entrada del cluster

Cuando se muestre el cuadro de edición de reglas de entradas se creará una nueva de tipo “Redshift” y con origen “Personalizada”.

Luego se debe introducir la dirección IP y mascará de subred del equipo informático donde será ejecutada la solución, para este caso particular es 103.158.32.123/32.

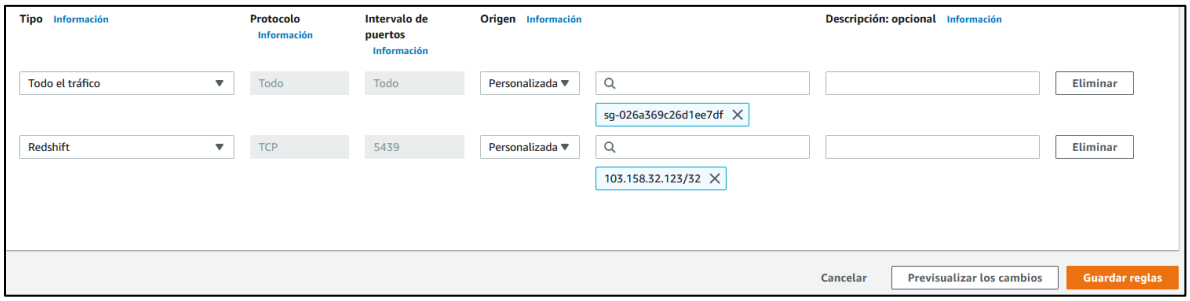


Ilustración 38: Editor de reglas de entrada

Después de guardar la nueva regla de entrada, se procede a la creación de la base de datos desde el editor de consultas, esquemas y tablas para el Data warehouse en Redshift.

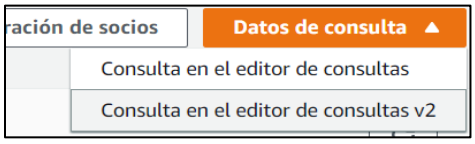


Ilustración 39: Botón para abrir el editor de consultas

Para acceder al editor de consultas se selecciona el botón “Datos de consulta” en el panel de configuración del cluster y luego dar clic en “Consulta en el editor de consultas v2”.

Al entrar se debe abrir una nueva pestaña de consultas para poder ejecutar el script del Data Warehouse.

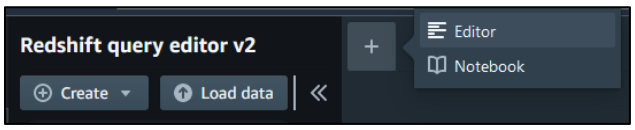


Ilustración 40: Abrir nuevo editor de consultas

Importante: En el siguiente paso se ejecutará el script de Redshift. Para visualizar el script que se debe ejecutar, referirse al enlace:
https://raw.githubusercontent.com/CEID-GRUPO5/LICASA_DW/main/Script%20Redshift.sql

Para ejecutar el script se da clic en el botón "Run".
Primero se ejecutará únicamente la sentencia de creación de la base de datos misma.

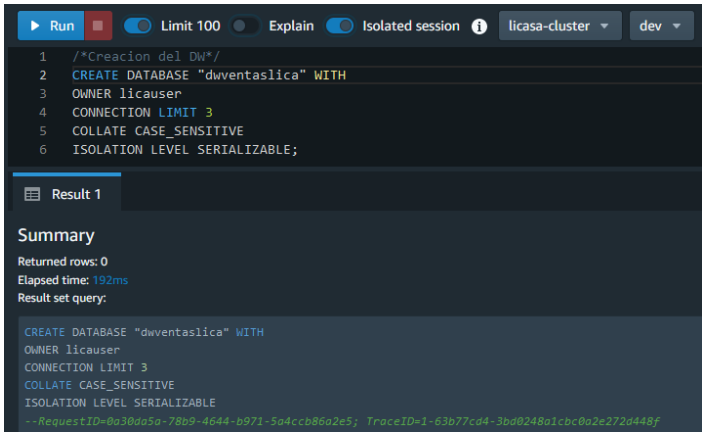


Ilustración 41: Sentencia de creación de la base de datos

Al terminar la ejecución de creación se procede a ejecutar el resto del script en otra pestaña, teniendo en cuenta que se ha de cambiar la base de datos donde será ejecutado como se muestra señalado en la siguiente imagen.

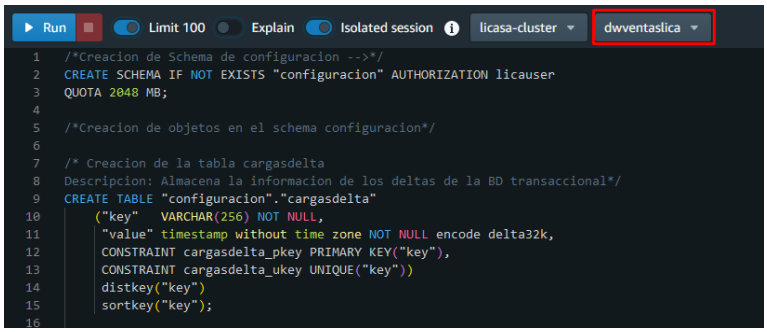


Ilustración 42: Ejecución del script de creación de esquemas y tablas

Al finalizar la ejecución del resto del script se debe revisar que cada una de las sentencias se haya ejecutado exitosamente, viendo el resultado en el explorador de objetos.

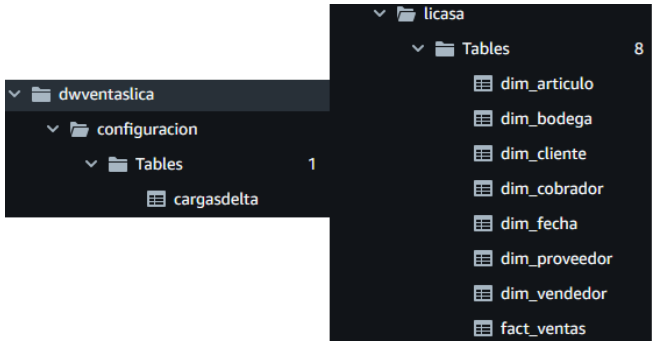


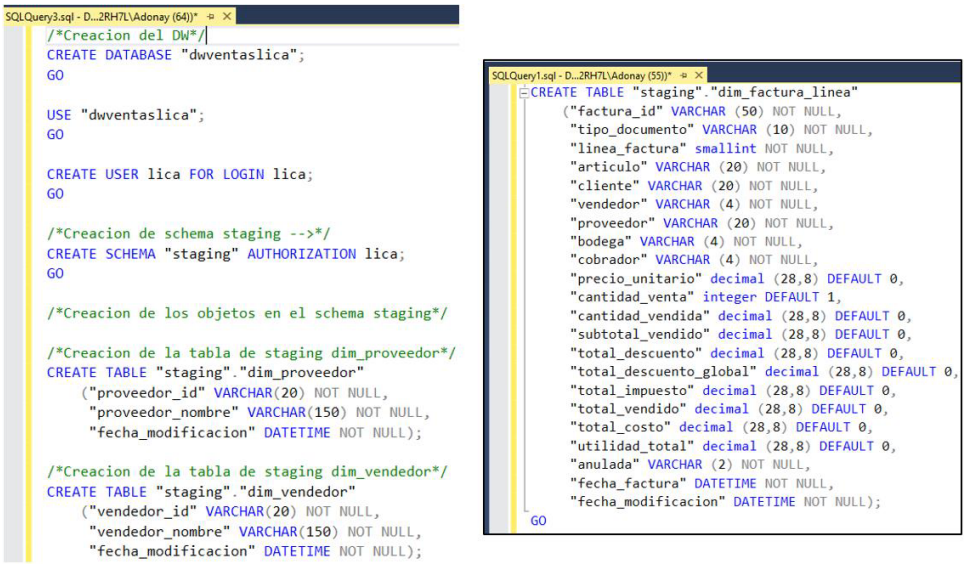
Ilustración 43: Explorador de objetos de Redshift - Esquemas y tablas

iii. Configuración de MS SQL Server

SQL Server es gestor de base de datos principal del sistema transaccional de la organización para la cual se ha creado la solución propuesta. Por tal motivo, se ha decidido utilizar la misma tecnología para realizar la limpieza preliminar de los datos, esto es, quitar duplicados y actualizar los datos que presentan modificaciones.

Utilizando el software SQL Server Management Studio se debe ejecutar el script de creación de la base de datos, esquema "staging" y sus respectivas tablas.

Para ver el script a ejecutar se puede visitar el siguiente enlace: https://raw.githubusercontent.com/CEID-GRUPO5/LICASA_DW/main/Script%20SQL%20Server.sql



```
/*Creacion del DW*/
CREATE DATABASE "dwventaslica";
GO

USE "dwventaslica";
GO

CREATE USER lica FOR LOGIN lica;
GO

/*Creacion de schema staging -->*/
CREATE SCHEMA "staging" AUTHORIZATION lica;
GO

/*Creacion de los objetos en el schema staging*/

/*Creacion de la tabla de staging dim_proveedor*/
CREATE TABLE "staging"."dim_proveedor"
    ("proveedor_id" VARCHAR(20) NOT NULL,
     "proveedor_nombre" VARCHAR(150) NOT NULL,
     "fecha_modificacion" DATETIME NOT NULL);

/*Creacion de la tabla de staging dim_vendedor*/
CREATE TABLE "staging"."dim_vendedor"
    ("vendedor_id" VARCHAR(20) NOT NULL,
     "vendedor_nombre" VARCHAR(150) NOT NULL,
     "fecha_modificacion" DATETIME NOT NULL);
```

```
CREATE TABLE "staging"."dim_factura_linea"
    ("factura_id" VARCHAR (50) NOT NULL,
     "tipo_documento" VARCHAR (10) NOT NULL,
     "linea_factura" smallint NOT NULL,
     "articulo" VARCHAR (20) NOT NULL,
     "cliente" VARCHAR (20) NOT NULL,
     "vendedor" VARCHAR (4) NOT NULL,
     "proveedor" VARCHAR (20) NOT NULL,
     "bodega" VARCHAR (4) NOT NULL,
     "cobrador" VARCHAR (4) NOT NULL,
     "precio_unitario" decimal (28,8) DEFAULT 0,
     "cantidad_venta" integer DEFAULT 1,
     "cantidad_vendida" decimal (28,8) DEFAULT 0,
     "subtotal_vendido" decimal (28,8) DEFAULT 0,
     "total_descuento" decimal (28,8) DEFAULT 0,
     "total_descuento_global" decimal (28,8) DEFAULT 0,
     "total_impuesto" decimal (28,8) DEFAULT 0,
     "total_vendido" decimal (28,8) DEFAULT 0,
     "total_costo" decimal (28,8) DEFAULT 0,
     "utilidad_total" decimal (28,8) DEFAULT 0,
     "anulada" VARCHAR (2) NOT NULL,
     "fecha_factura" DATETIME NOT NULL,
     "fecha_modificacion" DATETIME NOT NULL);
GO
```

Ilustración 44: Ejecución del script de base de datos staging en MSSQL

Se debe comprobar que se ejecute todo el script. Al finalizar el proceso se mostrará el siguiente mensaje.

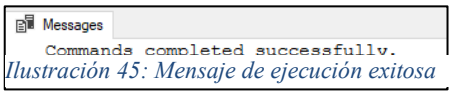


Ilustración 45: Mensaje de ejecución exitosa

Utilizando el explorador de archivos se verifica que todas las tablas fueron creadas exitosamente en el esquema denominado "staging".

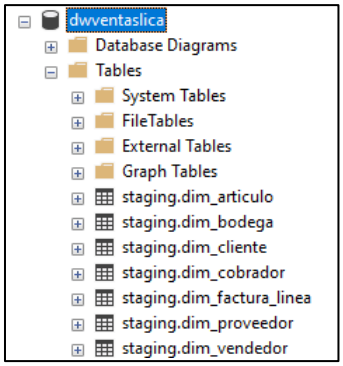


Ilustración 46: Explorador de objetos de Management Studio

iv. Configuración de procesos ETL en Talend

Creación de carpetas locales

Utilizadas para almacenar temporalmente los archivos descargados desde Amazon S3 para su procesamiento mediante los ETL.

Se debe crear un directorio de archivos temporales denominado “temp” en una ubicación fija y fácil de acceder. Para el caso particular, se ha seleccionado la ruta “C:\Users\USUARIO”

Una vez creada la carpeta, se debe emular la misma estructura creada en S3, con dos directorios principales “Archivados” y “NoProcesados”.

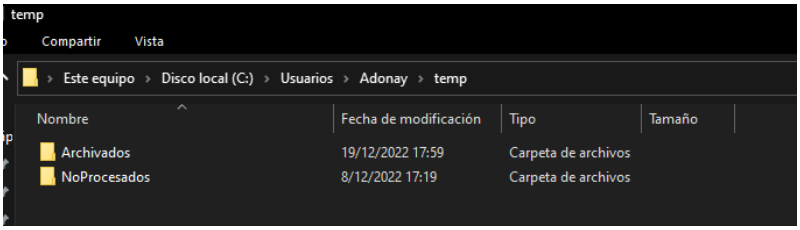


Ilustración 47: Carpetas locales para archivos temporales

Y sus respectivas subcarpetas “raw”, “stage” y “presentation”.

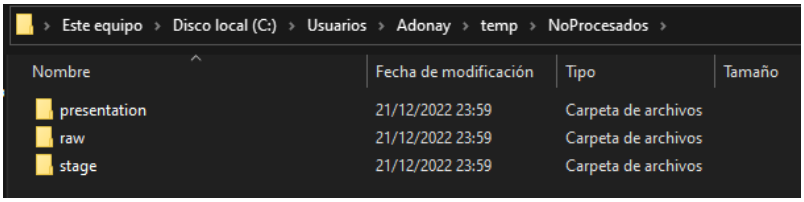


Ilustración 48: Subcarpetas locales correspondientes a las zonas de procesamiento

Importante: Anotar la ruta del directorio principal para ser utilizada en la asignación de variables más adelante.

Asignación de variables de contexto

Las variables de contexto que necesitan modificarse se encuentran en la sección *Contexts*, se asignaran valores a todos los grupos de contexto en la carpeta *Connections* y al grupo denominado *Environment*.

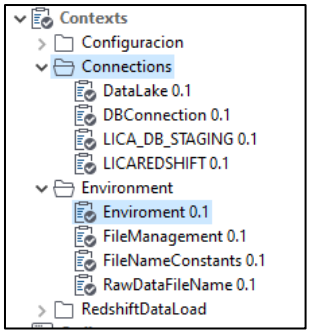


Ilustración 49: Variables de contexto a modificar

Data Lake

Variables que contienen las credenciales para conectarse al bucket de S3.Se modificarán las variables:

AccessKey por el ID de acceso
SecretKey por la clave de acceso secreta.
Ambos valores proporcionados por Amazon al momento de crear el usuario en la sección **Identity and Access Management (IAM).**

Name	Type	Comment	Default Value
BuketName	String	Nombre del Bucket en S3	lica.datalake
AccessKey	String	ID de acceso de S3	XXXXXXXXXXXXXXXXX
SecretKey	String	Clave secreta para acceder e interactuar con S3	XXXXXXXXXXXXXXXXXXX

Ilustración 50: Variables de contexto – Data Lake

LICA_DB_STAGING - base de datos de staging

Variables que contienen las credenciales y parámetros para conectarse a la base de datos de staging para realizar el proceso de limpieza de duplicados conservando los datos más recientes.Únicamente se modificarán las variables:

Password: ingresar clave del usuario asignado para acceder a la base de datos staging
Port: Ingresar el puerto utilizado para escuchar peticiones de conexión en el servidor
Server: Ingresar el nombre del servidor

Name	Type	Comment	Default Value
LICA_DB_STAGING_AdditionalParams	String		"noDatetimeStringSync=true;encrypt=true;trustServerCertificate=true;ssl=require"
LICA_DB_STAGING_Database	String	Nombre de la base de datos de staging	dwventaslica
LICA_DB_STAGING_Login	String	Nombre del usuario de SQL Server	lica
LICA_DB_STAGING_Password	Password	Clave del usuario de SQL Server	**********
LICA_DB_STAGING_Port	String	Puerto TCP de SQL Server	1433
LICA_DB_STAGING_Schema	String	Nombre del esquema de staging	staging
LICA_DB_STAGING_Server	String	Nombre del servidor	DESKTOP-XXXXXX

Ilustración 51: Variables de contexto - Base de datos de Staging

DBConnection
Variables que contienen las credenciales y parámetros para conectarse a la base de datos transaccional para realizar el proceso de extracción de datos.

Se modificarán todas las variables según los valores correspondientes al servidor de base de datos de la empresa, a excepción de "LICA_DB_AdditionalParams" de la siguiente manera.

Database: nombre de la base de datos transaccional
Login: nombre del usuario asignado para acceder a la base de datos
Password: clave del usuario asignado para acceder a la base de datos
Port: el puerto utilizado para escuchar peticiones de conexión en el servidor
Schema: nombre del esquema principal de la base de datos
Server: Ingresar el nombre del servidor

Name	Type	Comment	Default Value
LICA_DB_AdditionalParams	String		"noDatetimeStringSync=true;encrypt=true;trustServerCertificate=true;ssl=require"
LICA_DB_Database	String	Nombre de la base de datos	LICA
LICA_DB_Login	String	Nombre del usuario de SQL Server	lica
LICA_DB_Password	Password	Clave del usuario de SQL Server	****
LICA_DB_Port	String	Puerto TCP de SQL Server	1433
LICA_DB_Schema	String	Nombre del esquema	LICASA
LICA_DB_Server	String	Nombre del servidor	DESKTOP-XXXXXX

Ilustración 52: Variables de contexto - Base de datos transaccional

LICAREDSHIFT - Conexión a data warehouse en Redshift
Variables que contienen las credenciales y parámetros para conectarse al Data Warehouse para realizar el proceso de carga (inserción o actualización) de los datos producto del ETL.

Se modificarán las variables:

Login por el nombre de usuario administrador del clúster.
Password por la clave de usuario administrador.
Utilizar los valores introducidos al momento de crear el clúster en la sección **Redshift**

Name	Type	Comment	Default Value
LICAREDSHIFT_AdditionalParams	String		
LICAREDSHIFT_Database	String	Nombre de la base de datos de DW	dwventaslica
LICAREDSHIFT_Login	String	Nombre de usuario de Redshift	licauser
LICAREDSHIFT_Password	Password	Clave de acceso del usuario	**********
LICAREDSHIFT_Port	String	Puerto de enlace	5439
LICAREDSHIFT_Schema	String	Nombre del esquema de DW	licasa
LICAREDSHIFT_Schema_Configuracion	String	Nombre del esquema de configuración	configuracion
LICAREDSHIFT_Server	String	Ruta de enlace	licasa-cluster.cbzbtfsamhwu.us-east-1.redshift.amazonaws.com

Ilustración 53: Variables de contexto - Data warehouse

Environment
Variables que contienen las rutas al directorio de archivos temporales y el nombre del directorio mismo.

Modificar la variable **RootPath** con la ruta donde se ubica el directorio de archivos temporales creado en la sección **Creación de carpetas locales**.

Name	Type	Comment	Default Value
RootPath	String	Ruta local donde se unica el folder para el almacenamiento der archivos temporales	C:\Users\USUARIO
LocalTempPath	String	Folder para almacenar archivos temporales	\temp

Ilustración 54: Variables de contexto - Variables de entorno

Comprobar conexiones
En esta sección cubre los pasos necesarios para comprobar las conexiones configuradas en las variables de contexto.

Cada una de las conexiones a bases de datos que utilizan los procesos ETL se encuentran en la sección *Db Connections* en el menú *Metadata.*

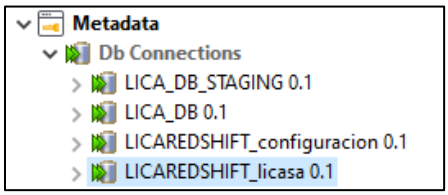


Ilustración 55: Conexiones a bases de datos utilizadas por el ETL

Al dar doble clic sobre ellas se abre un cuadro de dialogo como el siguiente.

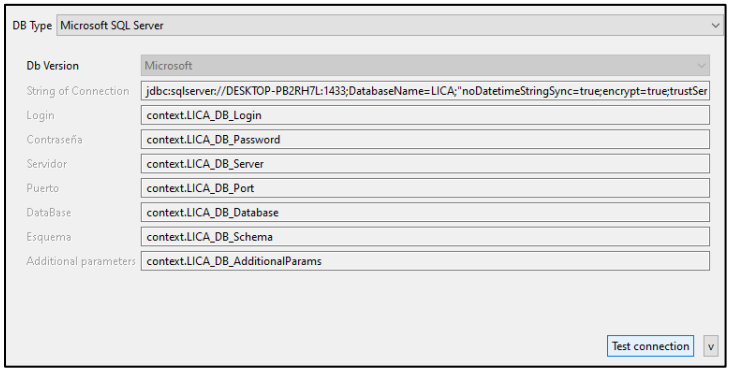


Ilustración 56: Detalles de la conexión a MSSQL Server

Donde se puede comprobar la conexión al dar clic al botón “Test connection”
Si la conexión es exitosa aparecerá el siguiente mensaje.

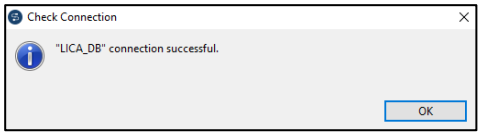


Ilustración 57: Mensaje de conexión exitosa

Se debe realizar el mismo proceso para todas las conexiones.

Interacción con los procesos ETL
Finalmente, para iniciar la ejecución de los procesos ETL se dispone de dos opciones:

- Ejecutar el trabajo denominado “MasterJob” que incluye los trabajos principales de cada etapa del ETL.
- Ejecutar los trabajos principales (Master) de manera individual en cada etapa.

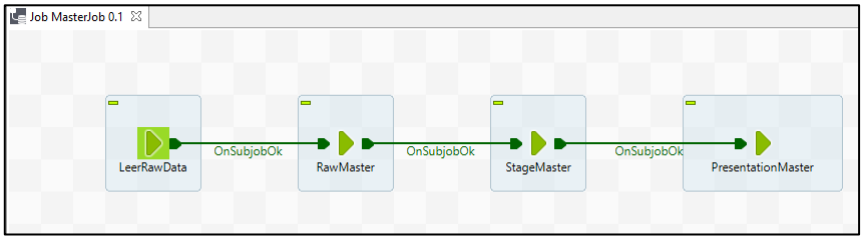


Ilustración 58: Procesos ETL - MasterJob

v. Plan de migración para migración de datos históricos

Para la puesta en marcha del sistema de Datawarehouse es necesario realizar la migración de los datos históricos de los años anteriores previo a la ejecución frecuente de los procesos ETL’s. Para ellos hay que tomar en cuente el orden en que deben migrarse y definir el día y hora en la cual realizarlo, dicho periodo debe estar fuera de los horarios que pudiera perjudicar a las actividades del día a día del negocio. Cabe mencionar que, para dicha migración, será hará uso, igualmente, de los procesos ETL´s de una manera más controlada o manual; es decir, no mediante Jobs anidados como podría hacerse posterior a la migración.

Debido a que las operaciones semanales de la empresa finalizan los días Sábados antes de la 1 pm y reanudan los días Lunes a las 8 am, se cuenta con una ventana de aproximadamente 43 hrs para realizar todo este trabajo. Según las pruebas realizadas en ambiente de prueba con copias más recientes de la base de datos de producción, se necesita de un periodo de 5 a 10 hrs para completar el trabajo. Esto nos deja con tiempo en caso que la migración presente inconvenientes y se requiera realizar trabajos de troubleshooting.

La recomendación es iniciar el proceso de migración un sábado a las 3 pm para contar con el día domingo para cualquier ajuste necesario y finalizar el uso de la base de datos de producción previo a la reanudación de la operación el día lunes por la mañana.

El orden de ejecución de los procesos es indistinto con relación a algunos datos históricos como: Articulos, Bodegas, Clientes, Cobradores, Proveedores, y Vendedores; sin embargo, hay tres condicionales que deben seguirse:

- Primeramente, se debe poblar la dimensión de Fechas mediante un proceso ETL particular para esta dimensión
- Debe respetarse el order de ejecución de grupos de Jobs (Zonas de datos)
    1. Leer de DB hacia la zona Raw (FromDBToRaw)
    2. Leer de la zona Raw hacia la zona Stage (FromRawToStage)
    3. Leer de la zona Stage hacia la zona de Presentation (FromStageToPresentation)
    4. Leer de Presentation hacia el DW en AWS Redshift (PresentationToRedShift)
- Los últimos datos que deben migrarse son los provenientes de FacturaLinea la cual es la base para la tabla de hechos

Dimensión de Fecha

La dimensión de Fecha es uno de los componentes críticos en el Data Warehouse planteado ya que permite a la empresa analizar los datos en función del tiempo. En este contexto, se hace uso de un insumo en formato Excel para la migración de dichos datos en la dimensión de Fecha a partir de un Job de carga incluyendo procesos de verificación y validación de los datos antes de la migración de las fechas en la dimensión. De esta manera, se evitan errores y se asegura la calidad y consistencia de los datos en la Dimensión de Fecha.

Este insumo de datos permite generar seis años completos de fechas y es totalmente parametrizable en base al campo full date, que se define según los requerimientos específicos de la organización en cuanto al periodo de tiempo establecido.

En conclusión, el uso de un insumo de datos en formato Excel para la migración de datos en la dimensión de Fecha es un proceso fundamental en un Data Warehouse. La estandarización, consistencia y verificación rigurosa de los datos son aspectos críticos para garantizar la calidad y precisión de las fechas generadas. Además, la definición cuidadosa y apropiada del parámetro es esencial para asegurar la coherencia de los datos en la dimensión de Fecha.

Otras dimensiones

Posterior a haber poblado la dimensión de Fecha, y aunque es indistinto para algunos datos, seguiremos un orden de migración con las demás dimensiones y la tabla de hechos.

Tomando en cuenta la segunda condicional, antes mencionada, debe migrarse en el siguiente orden:

1. Bodega
2. Vendedor
3. Proveedor
4. Cobrador
5. Cliente
6. Articulo

Lo anterior puede resumirse en: ejecutar uno a uno los Jobs del grupo 1 que corresponden a los datos de las entidades listadas anteriormente, luego los Jobs del grupo 2, 3, y finalmente 4.

Como resultado, tendremos las dimensiones de Bodega, Vendedor, Proveedor, Cobrador, Cliente, y Articulo pobladas en nuestro DW.

Como último paso, migrar la entidad (tabla) de FacturaLinea siguiendo, igualmente, el order de grupo de Jobs (zonas), ejecutando el job correspondiente a la entidad.

vi. Frecuencia y horarios para la ejecución de los ETLs

La frecuencia y horarios de ejecución de los ETL's dependerá de varios factores, como el volumen de los datos, la disponibilidad del sistema transaccional y la urgencia de la información.

Según los requerimientos de la empresa, se ha determinado que la mejor estrategia para la ejecución de los procesos ETL's será mediante tareas programadas en el servidor de la empresa. Los procesos se ejecutarán por zonas (Transaccional, Raw, Stage y Presentation) de manera semanal, el día sábado iniciando a las 2:00 pm y con una ventana de ejecución de 2 horas. Se deberá revisar periódicamente el rendimiento de los procesos ETL's y ajustar la estrategia de ejecución según las necesidades de la empresa.

En el caso se requiera la elaboración de informes de ventas bajo demanda, se deberá realizar la ejecución de los job maestros de forma manual para actualizar los datos del Datawarehouse con las transacciones más recientes. Se sugiere realizar esta tarea durante los periodos de menor actividad para evitar afectar el rendimiento del sistema transaccional.

En caso que los jobs maestros presentaran algún error, se sugiere ejecutar los process de forma individual para un mejor manejo de errores y correcciones.

b. Presupuesto de implementación

La información presentada a continuación puede variar ya que año por año puede cambiar (más probablemente aumentar) el espacio de almacenamiento de los archivos y horas de uso de los servicios.

Para el presupuesto de la implementación únicamente consideraremos el licenciamiento de Power BI y facturación de los servicios de AWS mensual.

Costeo de Amazon S3 (Simple Storage Service)

El costo del servicio de S3 cubre el espacio de almacenamiento que se utilice durante el tipo de vida de la solución, el número de peticiones realizadas de tipo PUT, COPY, POST y LIST, y de tipo GET y LIST. El estimado fue calculado en la calculadora del sitio web de Amazon (calculator.aws). A continuación, se detalla el cálculo:

Tiered price for: 10 GB

10 GB x 0.0230000000 USD = 0.23 USD

Total tier cost = 0.2300 USD (S3 Standard storage cost)

4,000 PUT requests for S3 Standard Storage x 0.000005 USD per request = 0.02 USD (S3 Standard PUT requests cost)

4,000 GET requests in a month x 0.0000004 USD per request = 0.0016 USD (S3 Standard GET requests cost)

2 GB x 0.0007 USD = 0.0014 USD (S3 select returned cost)

5 GB x 0.002 USD = 0.01 USD (S3 select scanned cost)

0.23 USD + 0.0016 USD + 0.02 USD + 0.0014 USD + 0.01 USD = 0.26 USD (Total S3 Standard Storage, data requests, S3 select cost)

S3 Standard cost (monthly): 0.26 USD

Subtotal: $0.26

Costeo de Amazon Redshift (Almacenamiento del modelo de estrella)

El costo del servicio de almacenamiento y procesamiento de datos de Redshift cubre el número de nodos (1), tipo de instancia (dc2.large), la de menor costo en este caso, y el número de horas de uso de la instancia durante el desarrollo (10 hrs. x 30 días = 300 hrs.). Esto implica que debe de programarse una rutina de encendido y apagado que se ejecute diariamente. A continuación, se detalla el cálculo:

1 instance(s) x 0.25 USD hourly x 300 hours in a month = 75.0000 USD

Redshift instance cost (monthly): 75.00 USD

Ilustración 59:Costo del servicio de Amazon Redshift - Implementación

Subtotal: $75.00

Costeo de Microsoft Power BI

El uso de la reportaría/visualización en Power BI, dentro de la empresa, únicamente es necesario para tres usuarios. El licenciamiento de Power BI Pro por usuario por mes tiene un costo de $9.99 (Licenciamiento Power BI)

Subtotal: $29.97

Total mensual: $105.23

Costo total anual: $1262.76

c. Análisis de resultados

En esta sección se brinda una breve descripción de los componentes de almacenamiento, organización, extracción, transformación y carga de datos transaccionales de la empresa al Data Warehouse, además, se muestran los resultados de la ejecución de la solución a través de capturas del aplicativo, organizadas en categorías según la tecnología utilizada y su rol en las diferentes zonas de procesamiento.

i. Bases de datos MS SQL Server

El gestor de bases de datos SQL Server tiene dos funciones en el procesamiento de datos, alojar los datos transaccionales producto de las actividades de la empresa y almacenar los datos de la zona Presentation temporalmente durante la eliminación de datos duplicados.

Base de datos transaccional

Fuente de datos transaccionales a los que se le aplicaran los procesos de extracción, transformación y carga.

Results | Messages

	ARTICULO	PLANTILLA_SERIE	DESCRIPCION	CLASIFICACION_1	CLASIFICACION_2	CLASIFICACION_3	CLASIFICACION_4	CLASIFICACION_5	CLASIFICACION_6	TIPO	ORIGEN_CORP	ULTIMA_SALIDA
1	MPBA10	NULL	BOLSA SONRISA SNACK LIG. SALADA REF.904 4 GALLETAS	009	9002	14	NULL	NULL	NULL	T	T	2020-12-31 20:39:17.740
2	E56	NULL	ESMALTE LOLITA ULTRA COLOR No.56	001	1001	01	ES	NULL	NULL	T	T	2016-05-19 14:14:25.167
3	E10	NULL	ESMALTE LOLITA ULTRA COLOR No.10 763041002104	001	1001	01	ES	NULL	NULL	T	T	2021-03-05 11:04:47.597
4	E100	NULL	ESMALTE LOLITA ULTRA COLOR NATURAL 763041002005	001	1001	01	ES	NULL	NULL	T	T	2021-10-27 14:10:57.740
5	E101	NULL	SECADOR ESMALTE LOLITA ULTRA COLOR 763041003040	001	1001	01	ES	NULL	NULL	T	T	2022-07-02 11:22:36.543
6	E102	NULL	ESMALTE LOLITA ULTRA COLOR No. 102 763041001022	001	1001	01	ES	NULL	NULL	T	T	2018-01-23 12:16:45.323
7	E103	NULL	ESMALTE LOLITA ULTRA COLOR No. 103 763041001039	001	1001	01	ES	NULL	NULL	T	T	2017-12-05 12:26:59.010
8	E104	NULL	ESMALTE LOLITA ULTRA COLOR No. 104 763041001046	001	1001	01	ES	NULL	NULL	T	T	2018-01-23 12:16:45.443
9	E105	NULL	KIT FRENCH LOLITA BLACK 763041003644	001	1010	01	NULL	NULL	NULL	T	T	2019-06-13 16:52:44.803
10	E106	NULL	ESMALTE LOLITA ULTRA COLOR No. 106 763041001060	001	1001	01	ES	NULL	NULL	T	T	2018-10-04 10:02:23.840
11	E107	NULL	ESMALTE LOLITA ULTRA COLOR No. 107 763041001077	001	1001	01	ES	NULL	NULL	T	T	2018-01-24 16:41:33.783
12	E108	NULL	ESMALTE LOLITA ULTRA COLOR No. 108 763041001084	001	1001	01	ES	NULL	NULL	T	T	2018-10-04 10:02:24.050
13	E12	NULL	ESMALTE LOLITA ULTRA COLOR No.12 763041002128	001	1001	01	ES	NULL	NULL	T	T	2019-10-16 09:06:11.360
14	E14	NULL	ESMALTE LOLITA ULTRA COLOR No.14 763041002142	001	1001	01	ES	NULL	NULL	T	T	2020-09-22 12:13:23.463
15	E15	NULL	ESMALTE LOLITA ULTRA COLOR No.15 763041002159	001	1001	01	ES	NULL	NULL	T	T	2021-01-25 08:15:32.823
16	E2	NULL	ESMALTE LOLITA ULTRA COLOR No. 2 763041002029	001	1001	01	ES	NULL	NULL	T	T	2017-11-10 10:53:43.133
17	E23	NULL	ESMALTE LOLITA ULTRA COLOR No.23 763041002234	001	1001	01	ES	NULL	NULL	T	T	2020-01-14 14:16:43.363
18	E24	NULL	ESMALTE LOLITA ULTRA COLOR No.24 763041002241	001	1001	01	ES	NULL	NULL	T	T	2019-10-16 09:06:10.840

Ilustración 60: MSSQL - Tabla de artículos en DB transaccional

Base de datos staging

Cuenta con la misma estructura de tablas que la base de datos en Redshift, sin embargo, no posee lleves primarias, foráneas o restricciones, ya que se espera que la data cargada tenga valores duplicados o desactualizados, únicamente se crean índices no agrupados para que la búsqueda de los registros más recientes y la eliminación de duplicados sea más eficiente.

	articulo_id	articulo_tipo	articulo_descripcion	clasificacion1_grupo	clasificacion2_subgrupo	clasificacion3_marca	fecha_modificacion
1	MPBA10	Terminado	BOLSA SONRISA SNACK LIG. SALADA REF.904 4 GALLETAS	MATERIA DE EMPAQUE	BOLSAS DE GALLETAS	DELIRICE	2020-12-31 20:39:17.740
2	E56	Terminado	ESMALTE LOLITA ULTRA COLOR NO.56	COSMETICOS	ESMALTES	LOLITA	2020-12-30 09:15:35.840
3	E10	Terminado	ESMALTE LOLITA ULTRA COLOR NO.10 763041002104	COSMETICOS	ESMALTES	LOLITA	2022-06-28 18:06:41.307
4	E100	Terminado	ESMALTE LOLITA ULTRA COLOR NATURAL 763041002005	COSMETICOS	ESMALTES	LOLITA	2022-06-21 10:58:44.480
5	E101	Terminado	SECADOR ESMALTE LOLITA ULTRA COLOR 763041003040	COSMETICOS	ESMALTES	LOLITA	2022-07-02 11:22:36.557
6	E102	Terminado	ESMALTE LOLITA ULTRA COLOR NO. 102 763041001022	COSMETICOS	ESMALTES	LOLITA	2022-06-28 18:05:16.813
7	E103	Terminado	ESMALTE LOLITA ULTRA COLOR NO. 103 763041001039	COSMETICOS	ESMALTES	LOLITA	2020-12-30 09:15:35.840
8	E104	Terminado	ESMALTE LOLITA ULTRA COLOR NO. 104 763041001046	COSMETICOS	ESMALTES	LOLITA	2022-06-28 18:07:36.227
9	E105	Terminado	KIT FRENCH LOLITA BLACK 763041003644	COSMETICOS	KITS DE ESMALTES	LOLITA	2022-06-28 18:08:02.310
10	E106	Terminado	ESMALTE LOLITA ULTRA COLOR NO. 106 763041001060	COSMETICOS	ESMALTES	LOLITA	2022-06-28 18:08:19.463
11	E107	Terminado	ESMALTE LOLITA ULTRA COLOR NO. 107 763041001077	COSMETICOS	ESMALTES	LOLITA	2022-06-28 18:08:37.950
12	E108	Terminado	ESMALTE LOLITA ULTRA COLOR NO. 108 763041001084	COSMETICOS	ESMALTES	LOLITA	2020-12-30 09:15:35.840

Ilustración 61: MSSQL - Tabla de artículos en DB de staging durante la limpieza

Al finalizar el proceso de limpieza y carga de datos en Redshift (dimensiones o tabla de hechos), se truncan todas las tablas para de staging para prepararlas para la siguiente ejecución del proceso ETL.

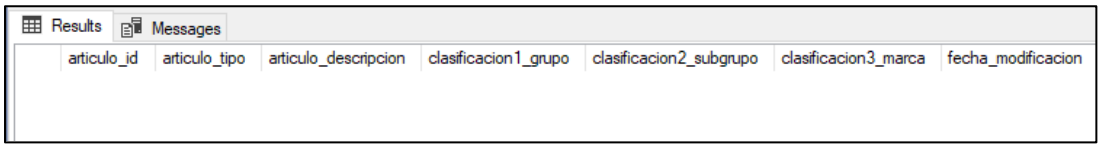


Ilustración 62: MSSQL - Tabla de artículos en DB de staging finalizado el procesamiento

ii. Bucket de S3

El bucket de S3 es el componente principal de almacenamiento y organización de archivos, hace la función de Data Lake y contiene dos categorías de archivos, No Procesados y Archivados.

No procesados: Contiene archivos creados producto de la ejecución de los ETL en cada una de las zonas de procesamiento. Se llaman no procesados ya que, al momento de cargarlos en el Data Lake se ubican en la zona posterior a la que fueron creados y se ponen a disposición para ser transformados más adelante.

Archivados: Contiene los archivos que ya procesados por los ETL en una zona particular, se utiliza para llevar un histórico de las cargas de datos extraídos y transformados.

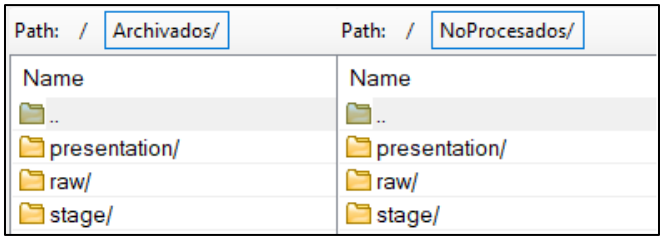


Ilustración 63: Datalake en S3 - Estructura de carpetas

iii. Carpetas de archivos temporales

Se utilizan para almacenar temporalmente archivos descargados de S3 o generados por los ETL y a la espera de ser cargados al Data Lake. Al final de cada etapa de procesamiento independientemente del resultado de la ejecución, se eliminan todos los archivos creados o descargados.

Tiene la misma estructura del Data Lake.

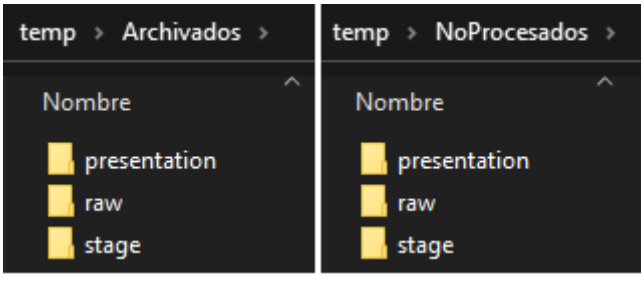
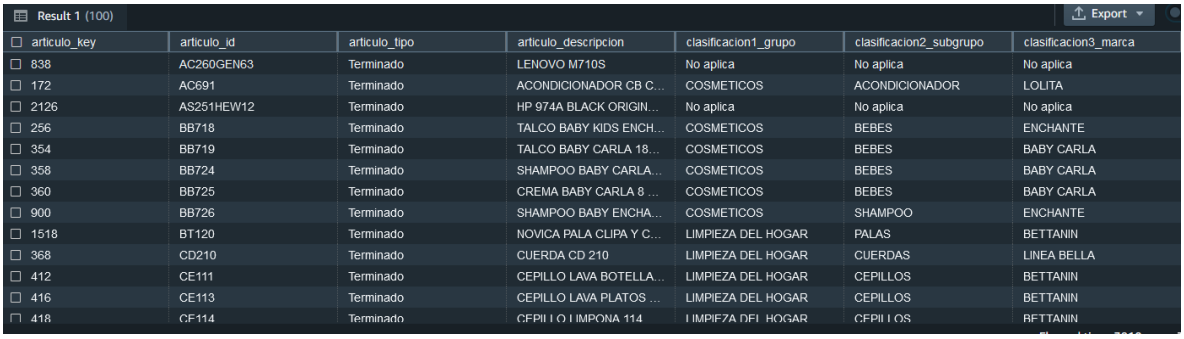


Ilustración 64: Estructura de carpetas locales

iv. Redshift

Componente principal de almacenamiento de los datos del Data Warehouse, la base de datos está construida utilizando un esquema de estrella que consta de 7 dimensiones y una tabla de hechos, este modelo de datos estructurados le permite aprovechar el procesamiento de datos en paralelo y facilita la lectura eficiente de millones de registros de ventas, además, permite el consumo de la información y la visualización su comportamiento en Power BI.

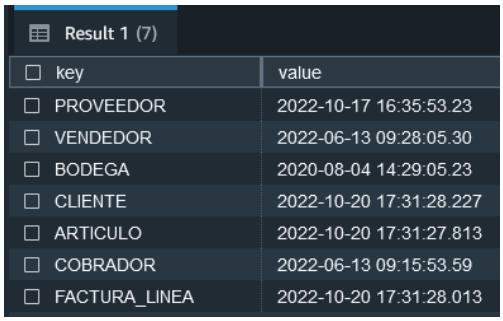


Result 1 (100) — Export

articulo_key	articulo_id	articulo_tipo	articulo_descripcion	clasificacion1_grupo	clasificacion2_subgrupo	clasificacion3_marca
838	AC260GEN63	Terminado	LENOVO M710S	No aplica	No aplica	No aplica
172	AC691	Terminado	ACONDICIONADOR CB C...	COSMETICOS	ACONDICIONADOR	LOLITA
2126	AS251HEW12	Terminado	HP 974A BLACK ORIGIN...	No aplica	No aplica	No aplica
256	BB718	Terminado	TALCO BABY KIDS ENCH...	COSMETICOS	BEBES	ENCHANTE
354	BB719	Terminado	TALCO BABY CARLA 18...	COSMETICOS	BEBES	BABY CARLA
358	BB724	Terminado	SHAMPOO BABY CARLA...	COSMETICOS	BEBES	BABY CARLA
360	BB725	Terminado	CREMA BABY CARLA 8 ...	COSMETICOS	BEBES	BABY CARLA
900	BB726	Terminado	SHAMPOO BABY ENCHA...	COSMETICOS	SHAMPOO	ENCHANTE
1518	BT120	Terminado	NOVICA PALA CLIPA Y C...	LIMPIEZA DEL HOGAR	PALAS	BETTANIN
368	CD210	Terminado	CUERDA CD 210	LIMPIEZA DEL HOGAR	CUERDAS	LINEA BELLA
412	CE111	Terminado	CEPILLO LAVA BOTELLA...	LIMPIEZA DEL HOGAR	CEPILLOS	BETTANIN
416	CE113	Terminado	CEPILLO LAVA PLATOS ...	LIMPIEZA DEL HOGAR	CEPILLOS	BETTANIN
418	CE114	Terminado	CEPILLO LIMPONA 114	LIMPIEZA DEL HOGAR	CEPILLOS	BETTANIN

Ilustración 65:Redshift - Datos de la tabla DimArticulo

También se almacena el registro de las cargas delta realizadas en la etapa del ETL. Posee un registro por cada tabla transaccional de la que se extraen datos.



Result 1 (7)

key	value
PROVEEDOR	2022-10-17 16:35:53.23
VENDEDOR	2022-06-13 09:28:05.30
BODEGA	2020-08-04 14:29:05.23
CLIENTE	2022-10-20 17:31:28.227
ARTICULO	2022-10-20 17:31:27.813
COBRADOR	2022-06-13 09:15:53.59
FACTURA_LINEA	2022-10-20 17:31:28.013

Ilustración 66: Redshift - Registros de cargas delta

v. Procesos ETL en Talend

A continuación, se describen los trabajos más importantes encargados de realizar el proceso ETL o apoyar en el mismo.

En cada zona de procesamiento se tiene un proceso (job) por tabla transaccional, además de un proceso maestro que se encarga de ejecutarlos.

Aclaración: Cada imagen correspondiente a un job muestra el mismo siendo ejecutado o luego de una ejecución exitosa. Y para Jobs que tienen un estructura y flujo de información similar, y que pertenecen a la misma zona, se ha colocado únicamente el considerado más relevante; en la descripción del mismo se detalla que jobs poseen la similitud.

Trabajos comunes

Son los trabajos encargados de realizar tareas comunes como listar, descargar, cargar, copiar, mover y eliminar archivos, y apoyan a las tareas de transformación en las zonas de procesamiento.

Actualizar delta

Se encarga de actualizar los registros de las cargas delta realizadas por el aplicativo, así la siguiente vez que se ejecute una tarea de extracción solo se seleccionarán los datos con fecha mayor a la que contenga el registro correspondiente.

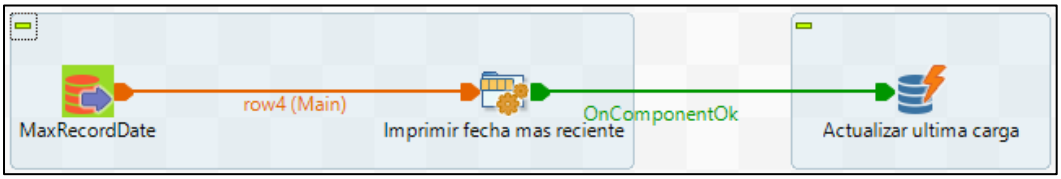


Ilustración 67: Proceso ETL - Actualizar delta

Descargar archivo

Comprueba la conexión a S3 y descarga el archivo que se le pasa como parámetro

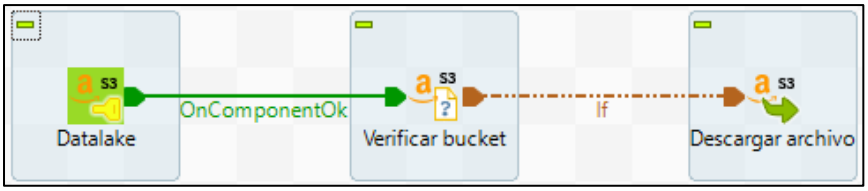


Ilustración 68: Proceso ETL - Descargar archivo de S3

Descargar lista de archivos

Hace uso del job anterior para descargar múltiples archivos basado en un prefijo Ej: “vendedor_raw”.

Ilustración 69: Proceso ETL - Descargar lista de archivos

Subir archivo

Carga los archivos generados por los ETL a S3.

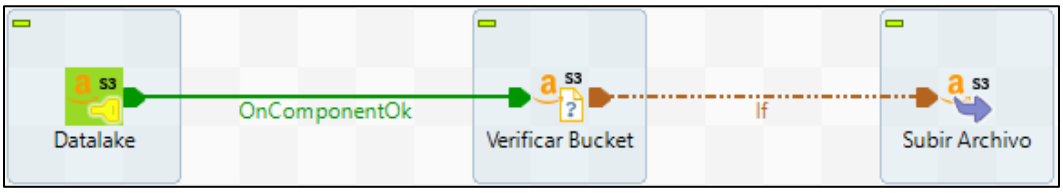


Ilustración 70: Proceso ETL - Subir archivo al datalake

Mover archivo

Traslada los archivos ya procesados hacia la carpeta “Archivados” luego de la ejecución de una tarea de transformación.

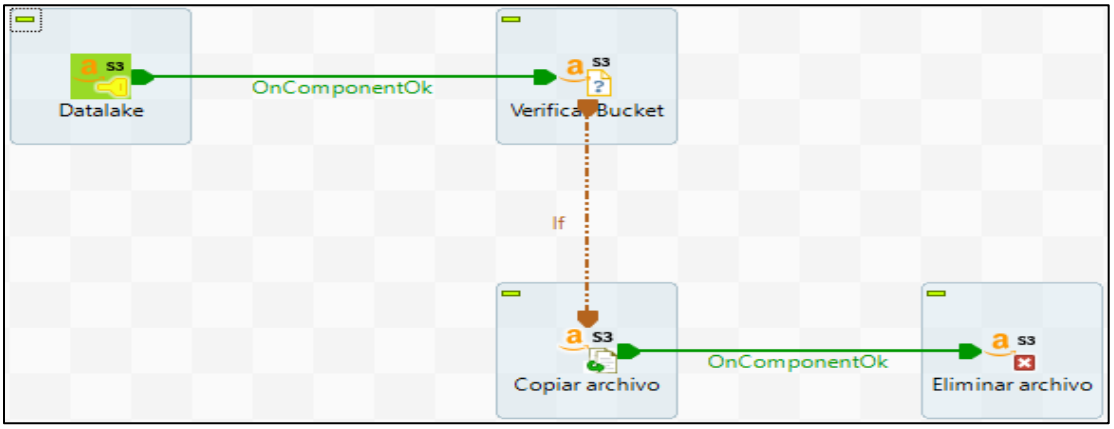


Ilustración 71: Proceso ETL - Mover Archivo procesado en S3

Eliminar archivo local

Al finalizar un proceso ETL automáticamente eliminar los archivos locales generados o descargados de S3.

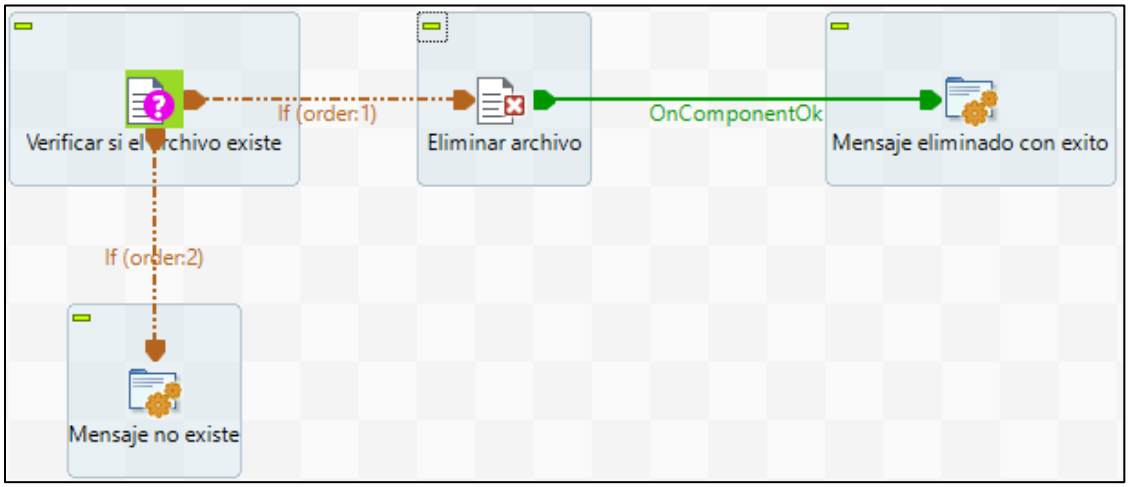


Ilustración 72: Proceso ETL - Eliminar archivo local

Extracción de datos transaccionales

Son los procesos encargados de extraer los datos del proceso de ventas de la empresa y almacenarlos en el Data Lake.

LeerDBMaster

Trabajo encargado de consultar las fechas de los últimos registros extraídos, cargarlas en variables de contexto y ejecutar los procesos que extraen los datos de la base transaccional.

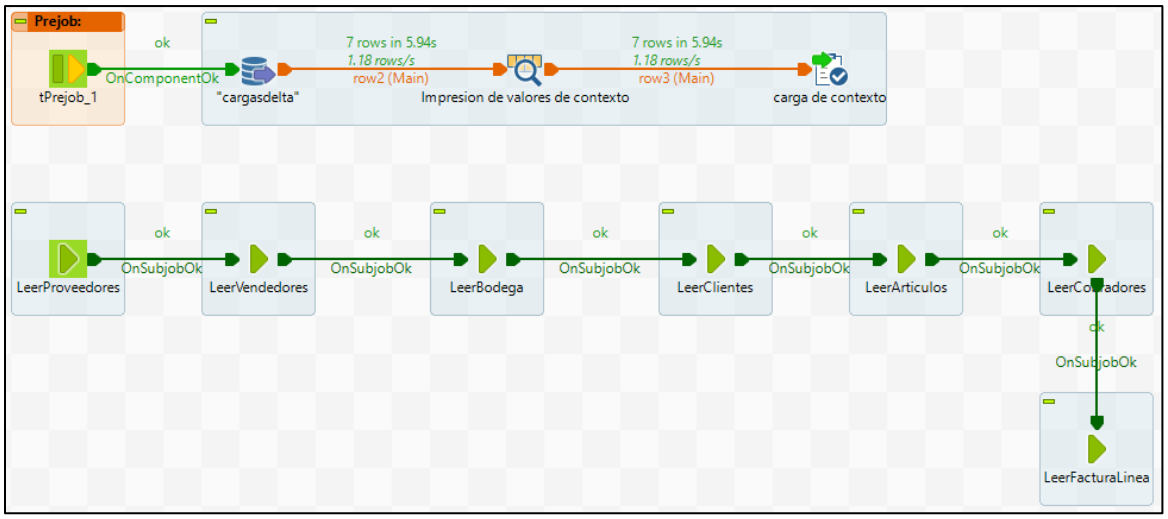


Ilustración 73: Procesos ETL – Ejecución del Proceso principal de extracción de la base transaccional

LeerArticulos

Proceso que extrae los datos transaccionales de la tabla Artículos. Asigna el nombre del archivo a cargar en S3 mediante código Java.

Este flujo también aplica para los procesos de las tablas: Cobrador.

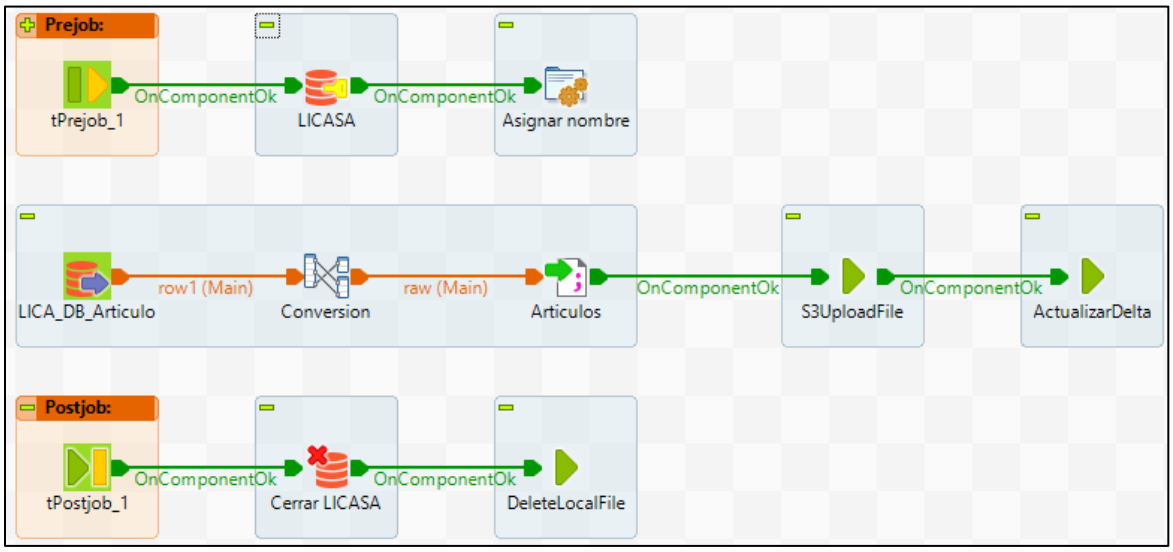


Ilustración 74: Procesos ETL - Extracción de registros de la tabla Artículos

LeerFacturaLinea

Proceso que extrae los datos transaccionales de la tabla FacturaLinea.

Hace lookup a las tablas Articulo y Factura para cambiar los id por descripciones fáciles de leer para los campos relacionados con el detalle de la factura y para consultar datos como: proveedor, vendedor, cobrador, cliente detallado en la factura.

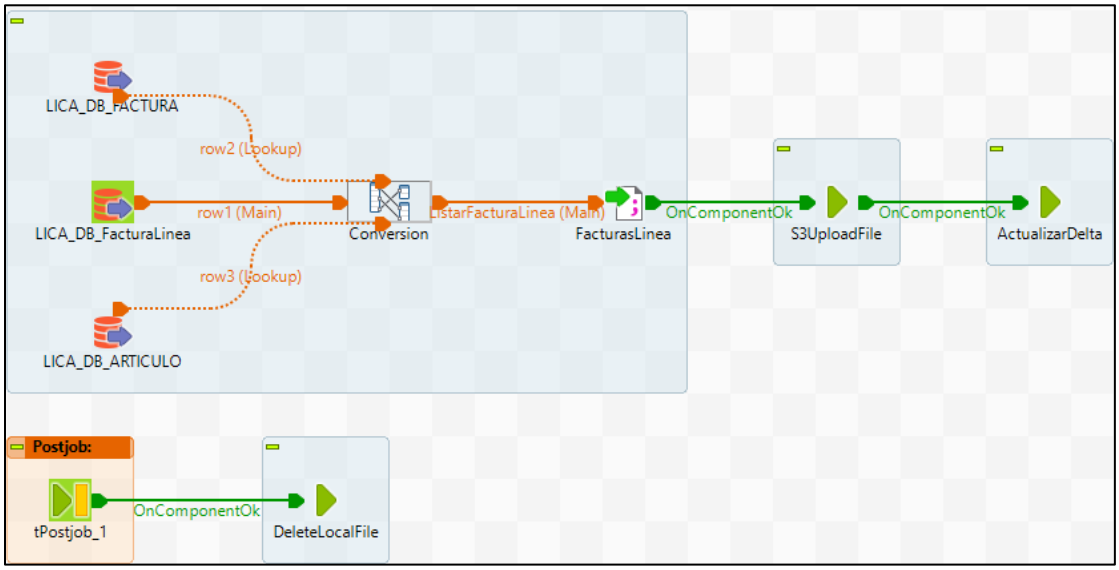


Ilustración 75: Procesos ETL - Extracción de registros de la tabla FacturaLinea

En la siguiente imagen se muestra el mapeo que consolida la información de las tres tablas.

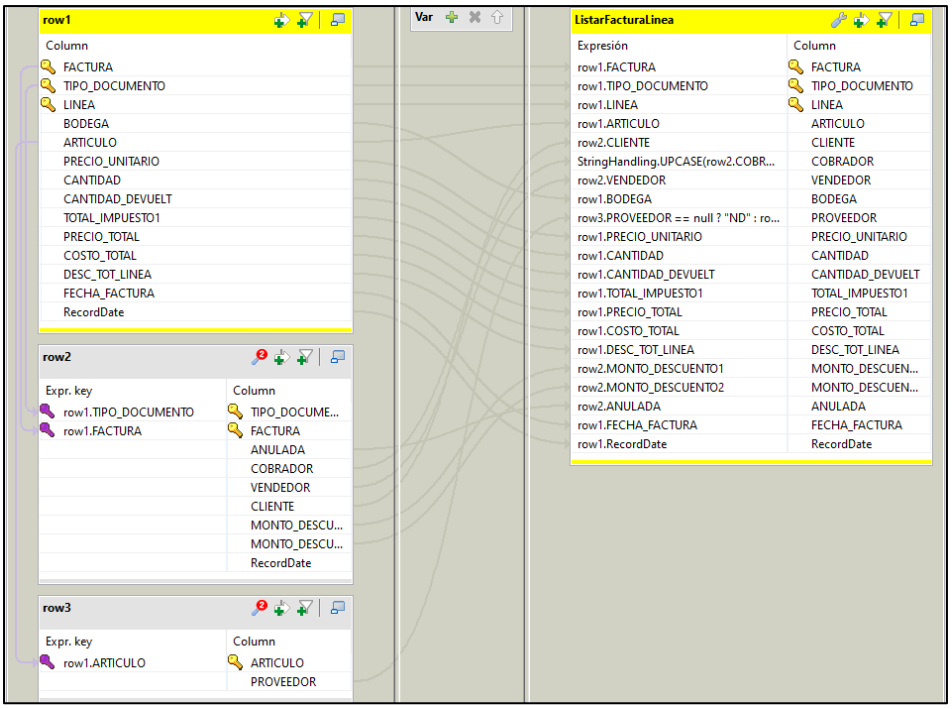


Ilustración 76: Procesos ETL - Consolidación de registros de Factura, FacturaLinea y Articulo

LeerBodega

Proceso que extrae los datos transaccionales de la tabla Bodega. Convierte a los nombres a mayúsculas para evitar inconsistencias.

Este flujo también aplica para los procesos de las tablas: Proveedor y vendedor.

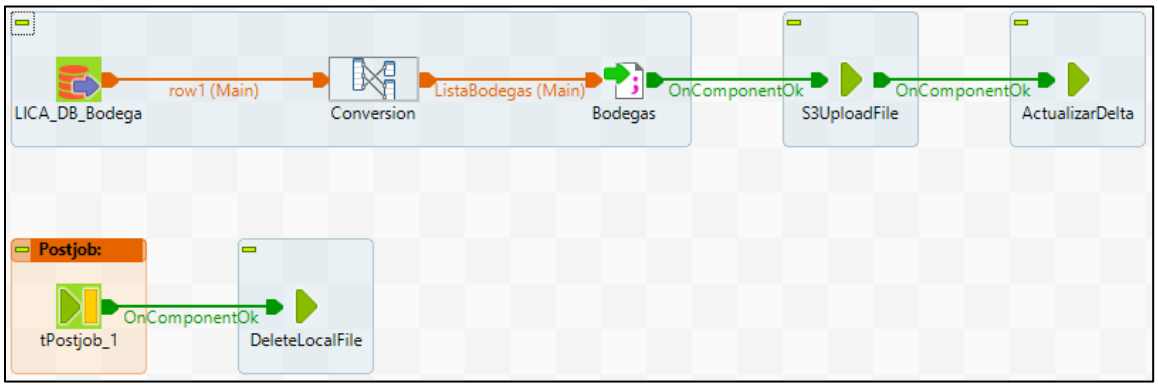


Ilustración 77: Procesos ETL - Extracción de registros de la tabla Bodega

LeerClientes

Proceso que extrae los datos transaccionales de la tabla Clientes. Hace lookup a las tablas Zona y Categorías_Cliente para cambiar los id por descripciones fáciles de leer. Cambia valores nulos por "No definido".

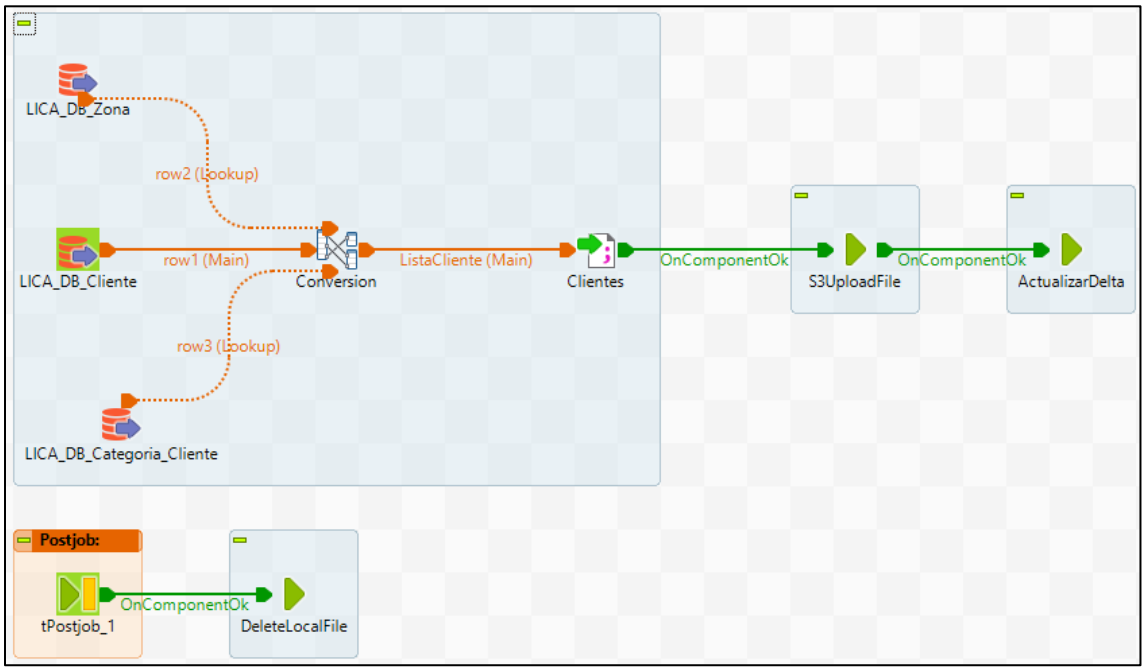


Ilustración 78: Procesos ETL - Extracción de registros de la tabla Clientes

Zona Raw

RawMaster

Trabajo encargado de ejecutar los procesos que transforman los datos de la zona Raw y cargan a la zona Stage en S3.

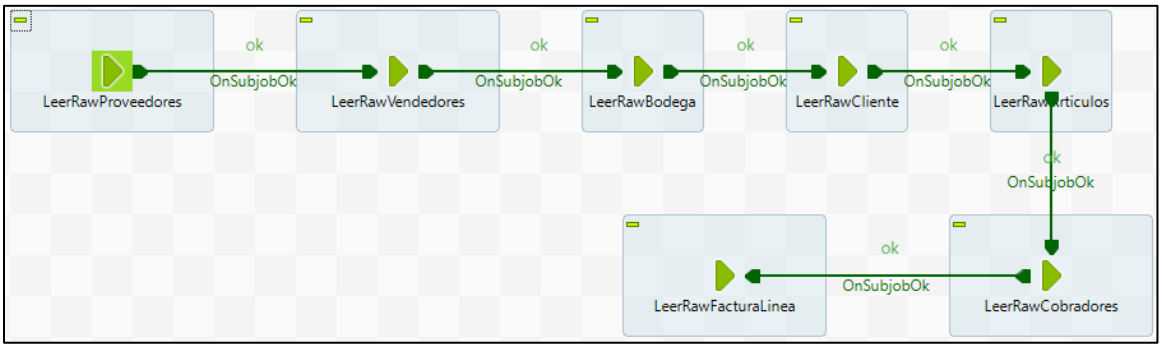


Ilustración 79: Proceso maestro de la zona RAW

LeerRawVendedores

Transforma los tipos de datos de la DB transaccionales por los adecuados para la carga en el data warehouse Redshift

Este flujo también aplica para los procesos de: Bodega, Cobrador, Cliente, y Proveedor.

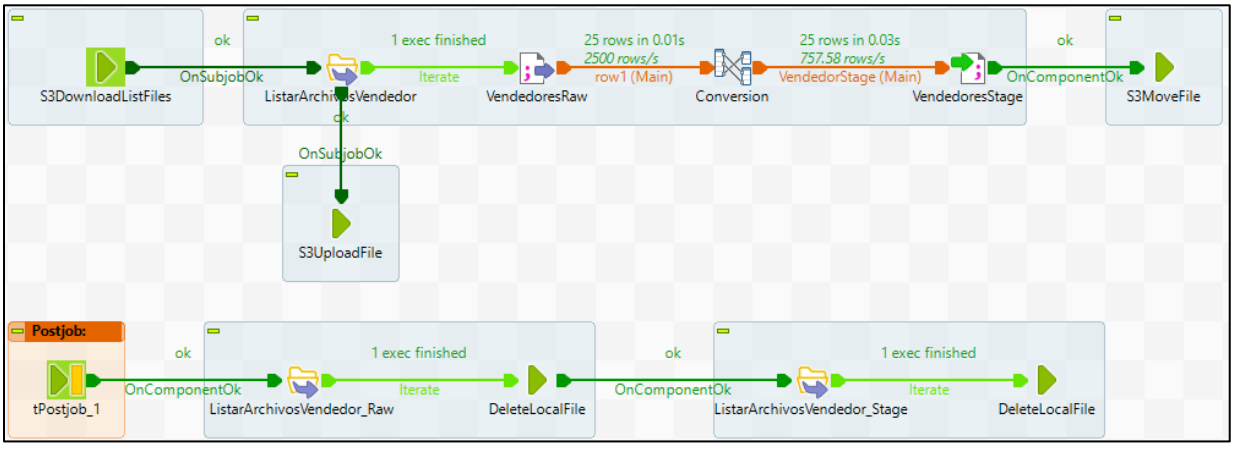


Ilustración 80: Procesos ETL - Transformación de datos de Vendedores

LeerRawFacturaLinea

Transforma los tipos de datos de la DB transaccionales por los adecuados para la carga en el Data Warehouse Redshift.

Realiza los cálculos del precio total de venta, la cantidad vendida, el descuento aplicado y la utilidad total. Además, cambia valores nulos por “No definido” o asigna valores por defecto.

Al finalizar el procesamiento, el archivo generado se comprime en formato ZIP para reducir el tamaño que ocupará en el Data Lake.

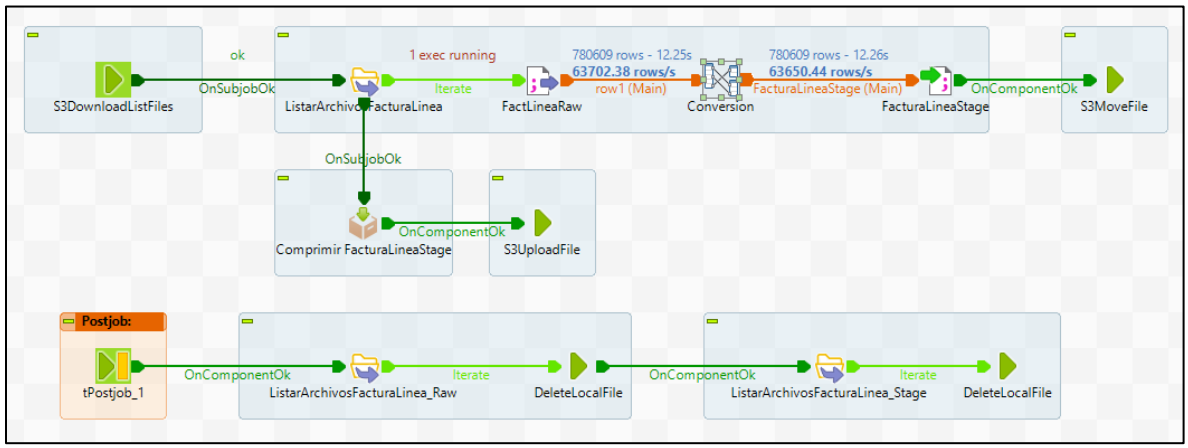


Ilustración 81: Procesos ETL - Transformación de la información sobre ventas

En la siguiente se página se muestran las transformaciones realizadas.

FacturaLineaStage

Expresión	Column
row1.FACTURA	factura_id
row1.TIPO_DOCUMENTO.equals("F") ? "FACTURA" : "DEVOLUC...	tipo_documento
row1.LINEA	linea_factura
row1.ARTICULO	articulo
row1.CLIENTE	cliente
row1.VENDEDOR	vendedor
row1.PROVEEDOR	proveedor
row1.BODEGA	bodega
row1.COBRADOR	cobrador
row1.PRECIO_UNITARIO	precio_unitario
1	cantidad_venta
row1.CANTIDAD.subtract(row1.CANTIDAD_DEVUELT)	cantidad_vendida
row1.PRECIO_TOTAL.add(row1.DESC_TOT_LINEA)	subtotal_vendido
row1.DESC_TOT_LINEA	total_descuento
row1.MONTO_DESCUENTO1.add(row1.MONTO_DESCUENTO2)	total_descuento_global
row1.TOTAL_IMPUESTO1	total_impuesto
row1.PRECIO_TOTAL	total_vendido
row1.COSTO_TOTAL	total_costo
row1.PRECIO_TOTAL.subtract(row1.COSTO_TOTAL)	utilidad_total
row1.ANULADA.equals("N") ? "NO":"SI"	anulada
row1.FECHA_FACTURA	fecha_factura
row1.RecordDate	fecha_modificacion

Ilustración 82: Procesos ETL - Mapeo y cálculo de métricas sobre el proceso de venta

LeerRawArticulos

Hace tres loopkup a la tabla clasificación, para asignar el grupo, subgrupo y marca de cada artículo, y reemplaza los id por nombres descriptivos.

Ahorrar memoria y procesamiento se utilizó un componente del tipo HashMap para hacer la referencia una única vez a los datos de la tabla.

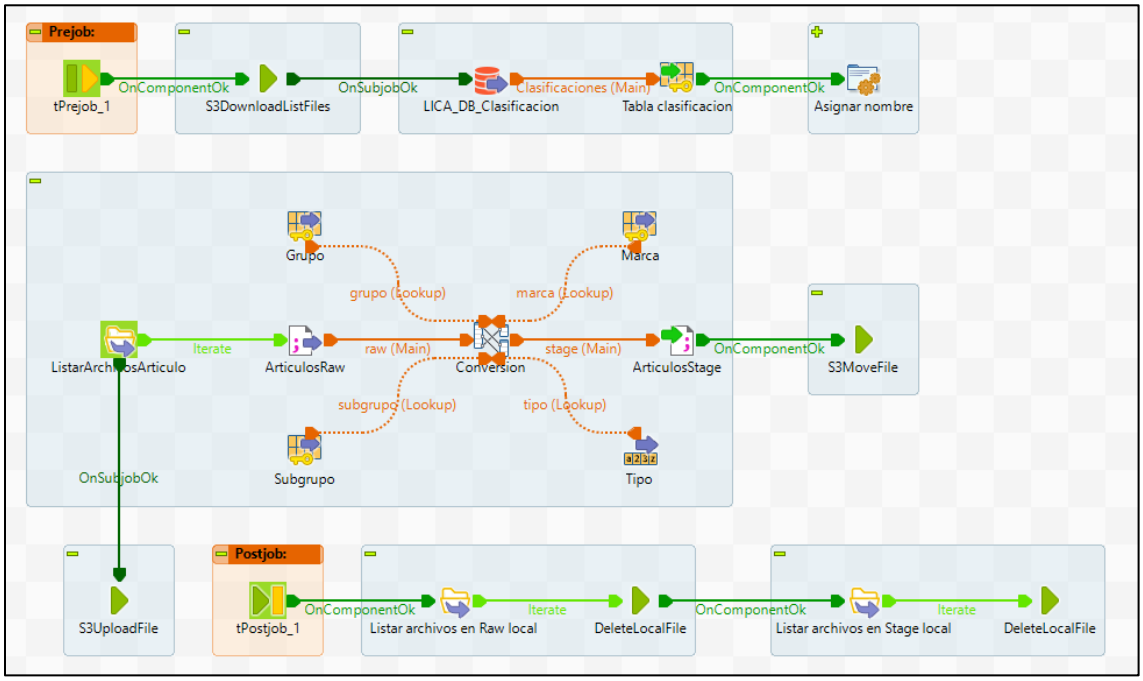


Ilustración 83: Procesos ETL - Transformación y lookup de información de artículos de venta

Zona Stage

StageMaster

Trabajo encargado de ejecutar los procesos que transforman los datos de la zona Stage y cargan a la zona Presentation en S3.

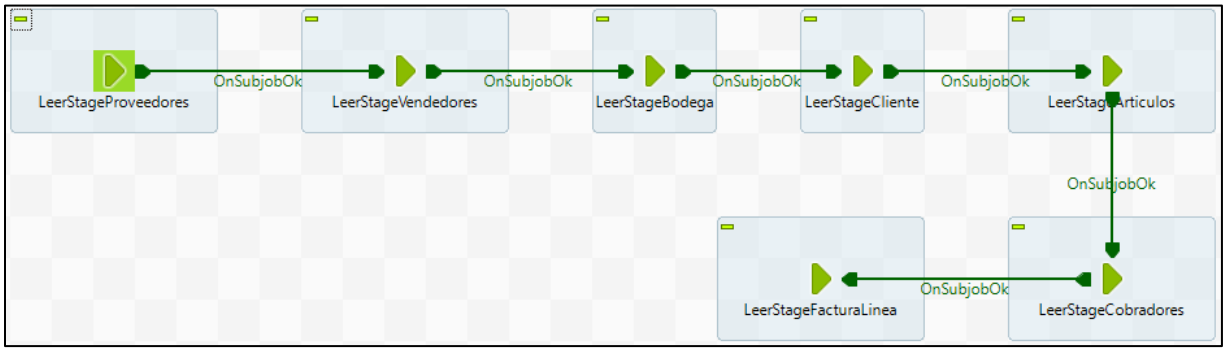


Ilustración 84: Procesos ETL - Proceso maestro de la zona STAGE

LeerStageFacturaLinea

Prepara los datos para ser cargados al data warehouse,

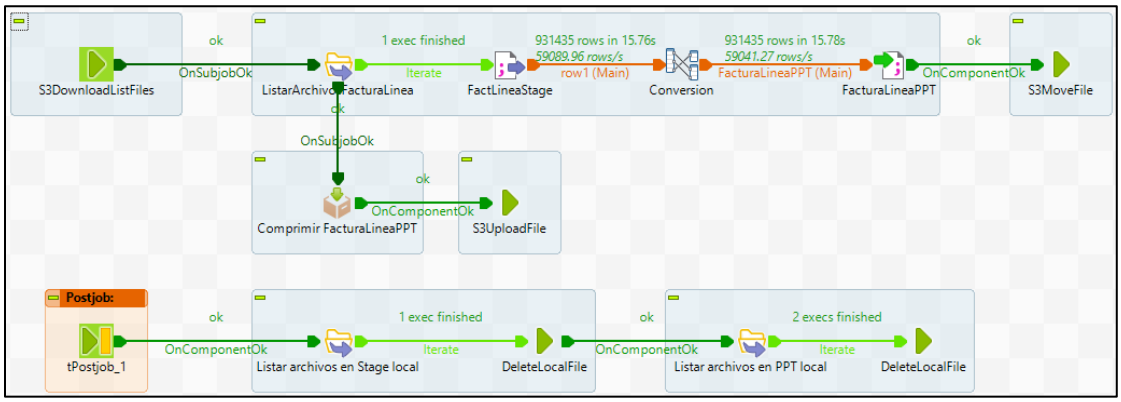


Ilustración 85: Procesos ETL - Preparación de datos de FacturaLinea para la zona de presentación

LeerStageProveedores

Prepara los datos para ser cargados al data warehouse y se mueven los archivos resultantes hacia la zona de presentación.

Este flujo es aplicable a datos de: Proveedores, Vendedores, Bodega, Artículos y Clientes.

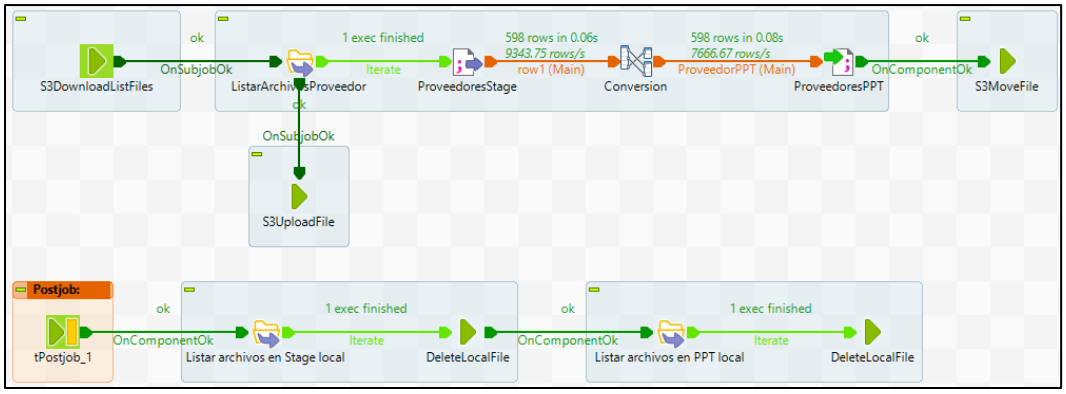


Ilustración 86: Procesos ETL - Preparacion de datos sobre proveedores para la zona de presentación

Zona Presentation

PresentationMaster

Trabajo encargado de ejecutar los procesos que cargan los datos de la etapa Presentation al data warehouse.

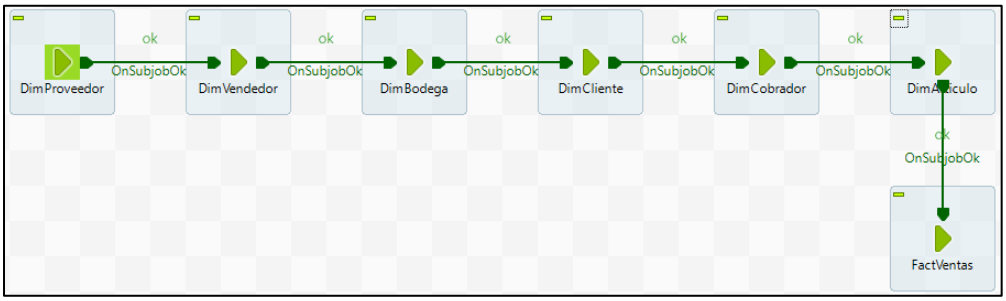


Ilustración 87: Procesos ETL - Proceso maestro de la zona PRESENTATION

Cargar datos en tabla de dimensión

La carga de datos a las tablas de dimensión esta estandarizado, por lo tanto, solo se muestra el resultado de una de las dimensiones. El proceso de carga se divide en 4 etapas:

La etapa preliminar consta se los componentes para establecer las conexiones a MS SQL Sever, Redshift y S3.

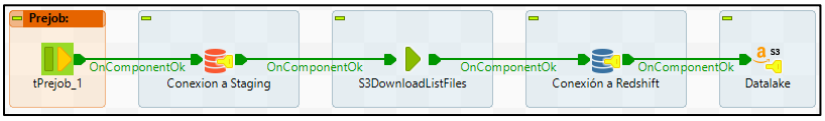


Ilustración 88: Carga de datos a tablas de dimensiones - Etapa preliminar

La primera etapa de procesamiento es la carga de datos descargados de S3 en la base de datos de Staging, luego se procede a identificar y eliminar los datos duplicados o desactualizados.

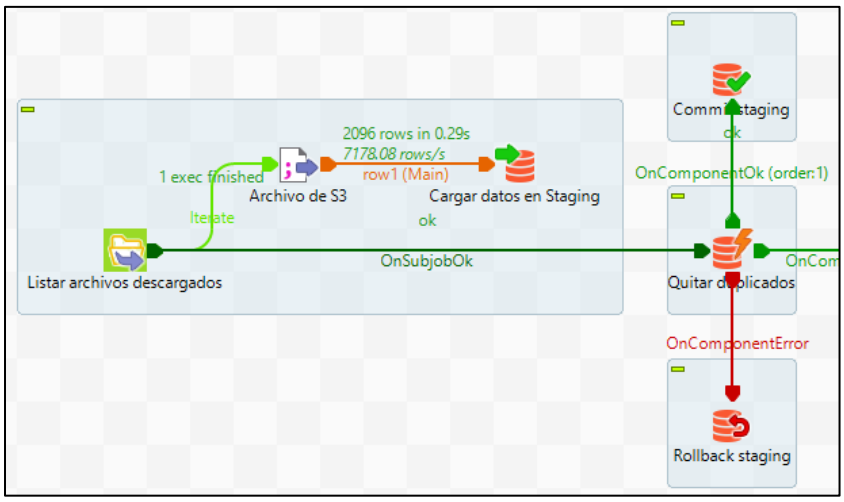


Ilustración 89: Carga de datos a tablas de dimensiones - Identificación y eliminación de duplicados

La segunda etapa se consultan los datos de staging y se clasifican los datos en nuevos y existentes, basándose en los registros actuales del data warehouse a través de un mapeo.

El resultado es un archivo que se utilizará para insertar los datos nuevos en la base de datos de Redshift mediante el comando “copy” y para datos que necesiten actualizarse se crea un componente que busca y actualiza los datos del data warehouse.

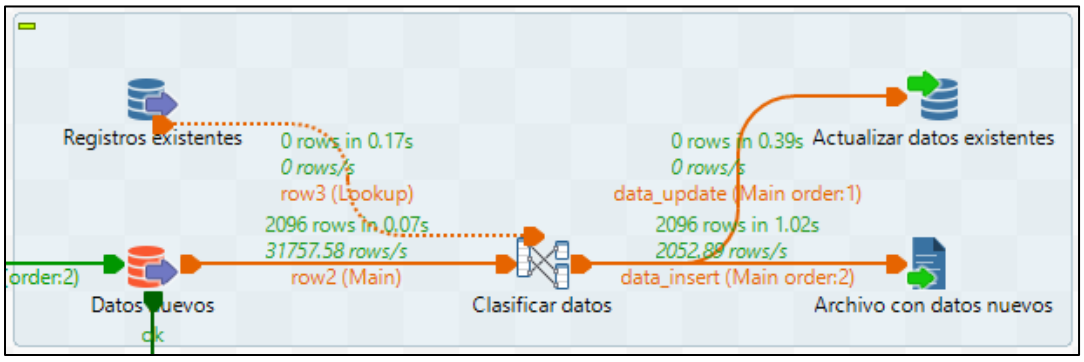


Ilustración 90: Carga de datos a tablas de dimensiones - Clasificación de datos

Una vez finalizada la clasificación se envía el comando desde Talend a Redshift para iniciar el proceso de inserción y actualización.

La creación del archivo para la carga de datos se condiciona a la existencia de datos nuevos, si únicamente se actualizarán registros existentes el job culmina su ejecución y procede a la última etapa para cerrar las conexiones y eliminar los archivos locales.

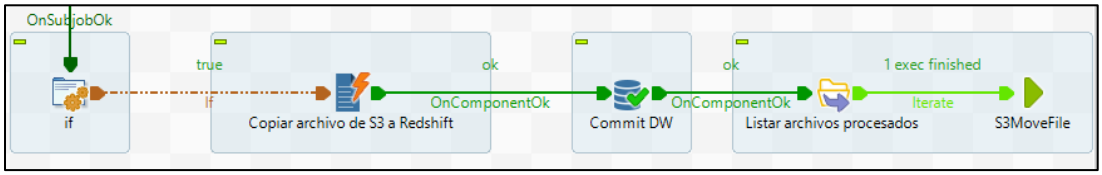


Ilustración 91: Creación del archivo para la inserción de datos en la dimensión

Finalmente, la tercera etapa o posterior se encarga de eliminar los archivos descargados o generados localmente y también cierra las conexiones establecidas.

Ilustración 92:Carga de datos a tablas de dimensiones - Etapa posterior a la ejecución

Cargar datos en tabla de hechos

La carga de datos para la tabla de hechos es similar a la de una dimensión, sobre todo la etapa preliminar y la identificación de duplicados.

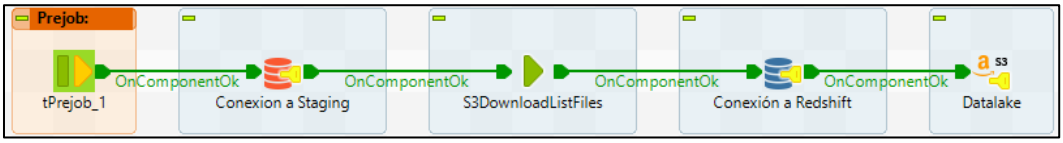


Ilustración 93: Etapa preliminar a carga de datos en tabla de hechos

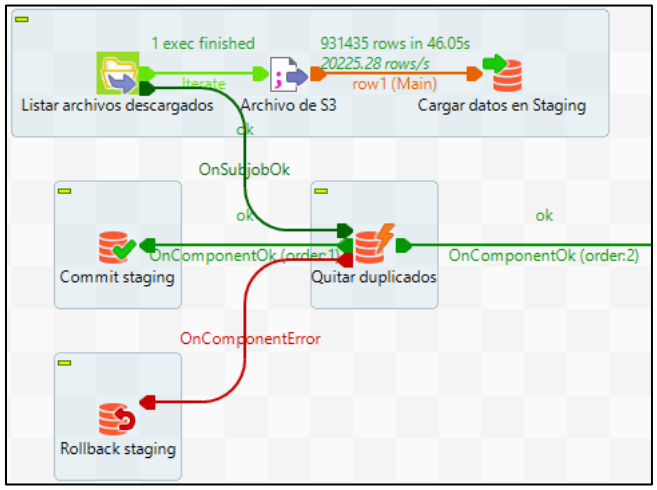


Ilustración 94: Eliminación de datos duplicados y desactualizados

La particularidad que tiene el proceso es que, previo a la etapa de clasificación de datos se realizan los lookups a las tablas de dimensiones para reemplazar las llaves de negocio con las llaves subrogadas.

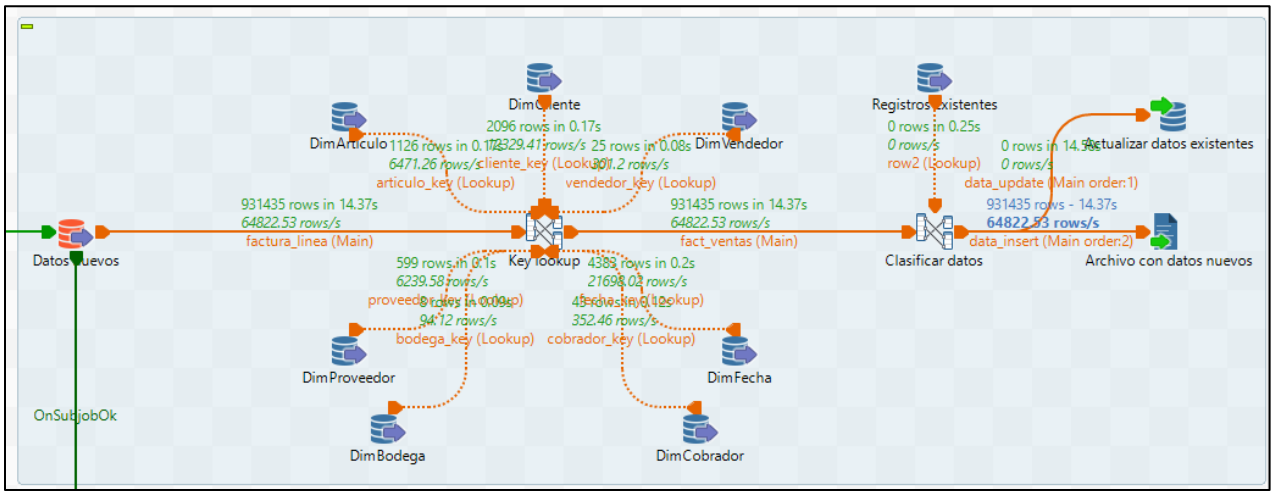


Ilustración 95: Lookup de llaves subrogadas para la tabla de hechos

Ilustración 96: Creación del archivo para realizar la carga de datos

Resultado del llenado de la tabla de hechos

Result 1 (100) — Export — Chart

articulo_key	cliente_key	fecha_key	vendedor_key	proveedor_key	bodega_key	cobrador_key
140	1056	20160409	22	158	2	30
144	1056	20160409	22	158	2	30
504	3692	20160413	26	450	2	2
510	3692	20160413	26	450	2	2
514	3692	20160413	26	450	2	2
522	3692	20160413	26	450	2	2
530	3692	20160413	26	450	2	2
530	3692	20160413	26	450	2	2
534	3692	20160413	26	450	2	2
506	3692	20160413	26	450	2	2
514	3692	20160413	26	450	2	2
504	3692	20160413	26	450	2	78

Ilustración 97: Registros en la tabla de hechos

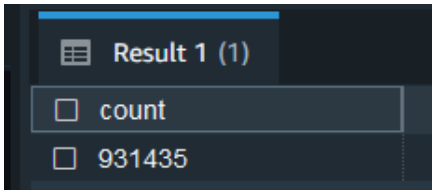


Ilustración 98: Total de registros en la tabla de hechos

vi. Power BI

A continuación, se muestran los dashboards/reportes/visualizaciones que detallaran de manera gráfica el análisis de los datos procesados y dan respuesta a las métricas identificadas durante el análisis y diseño del modelo dimensional

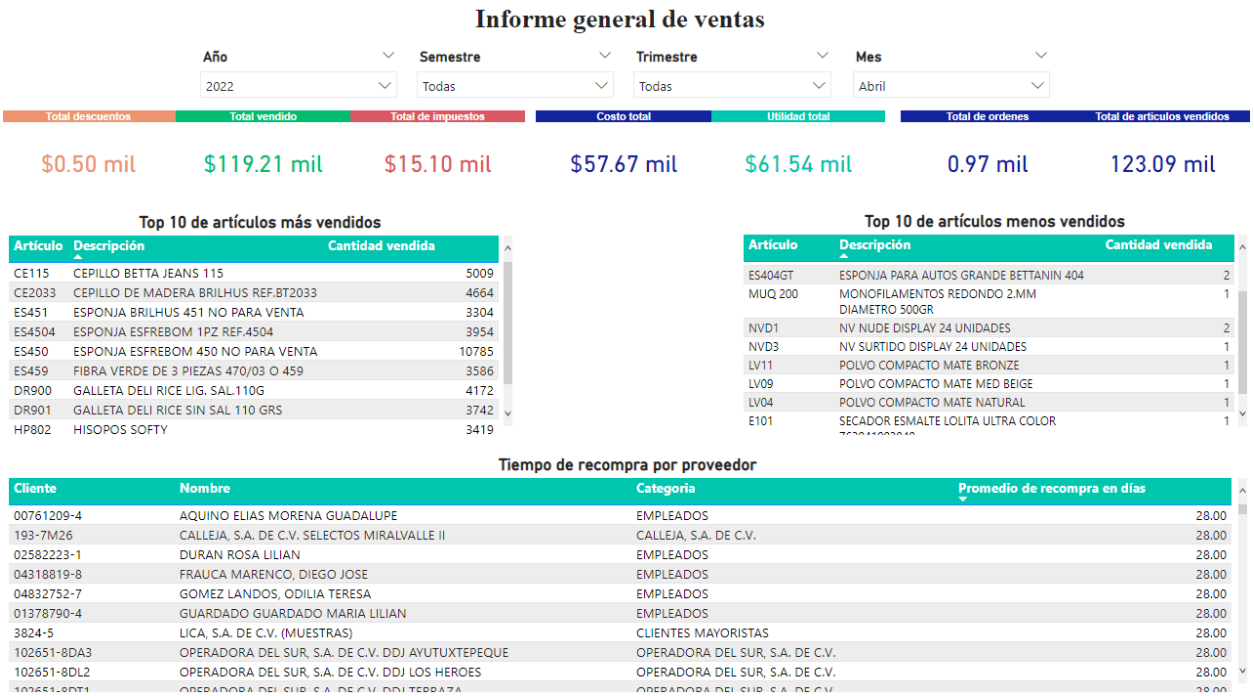


Ilustración 99: Informe general de ventas

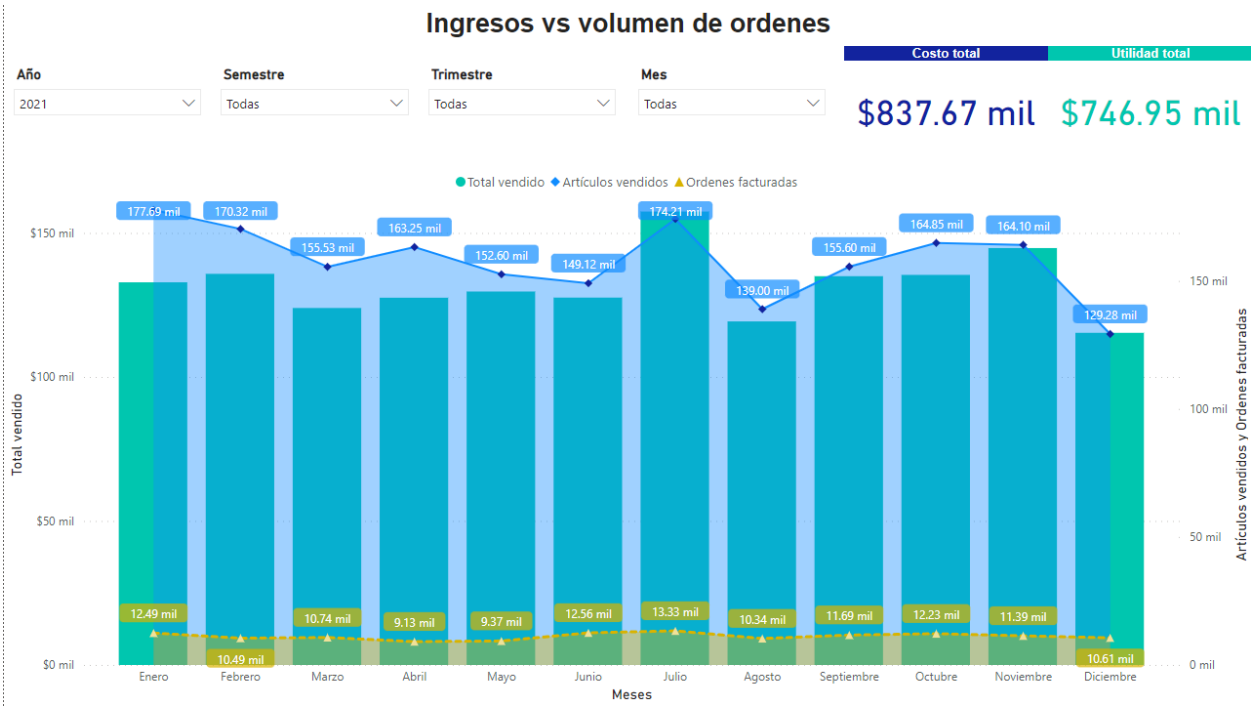


Ilustración 100: Informe de ingresos vs volumen de ordenes

Ilustración 101: Informe de rendimiento de vendedores

CONCLUSIONES

Por medio de la elaboración de este proyecto, con el fin de diseñar un model multidimensional de Data Warehouse que brindara soporte al proceso de ventas de la Empresa LICA SA de CV que lleva sus operaciones de ventas y facturación mediante el sistema de Softland, se logró conocer como es el funcionamiento del proceso de ventas en el ERP de Softland, analizar sus datos y realizar un perfilado mediante el uso de herramientas nuevas destinadas a brindar soporte para la construcción de soluciones de Data Warehouse.

El desarrollo del proyecto nos permitió crear un modelo multidimensional que solventara necesidades analíticas de las cuales la empresa carecía o le tomaba mucho tiempo procesar para generar información que le apoyara en la toma de decisiones en su negocio. Se construyeron procesos ETL's para trasladar los datos de la base de datos transaccional al data warehouse. Finalmente, con los datos en el data warehouse, se logró una integración con Power BI para la elaboración de visualizaciones que dieron respuesta a las interrogantes de la organización referente a su proceso de ventas y facturación.

RECOMENDACIONES

Por motivos de reducción de costos el proceso de filtrado de datos duplicados se realiza de forma local utilizando una base de datos de SQL Server, esto hace que las tareas de carga sean propensas a ejecuciones lentas cuando los archivos de datos de la zona de presentación poseen una gran cantidad de registros que requieren actualización, ya que se tiene como limitante la velocidad de conexión con los servicios de AWS.

Para continuar con el desarrollo del aplicativo y escalar su capacidad, así como su eficiencia se debe tomar en cuenta la creación de la base de datos de staging en Redshift, ya que el uso de comandos especializados como Merge y Copy agiliza el proceso de inserción y actualización de la información en el Data Warehouse, además, reduce la complejidad de las tareas del ETL.

Se debe considerar limitar el uso de archivos CSV en favor de archivos de texto plano para incrementar la velocidad de procesamiento.

Bibliografía

Amazon IAM, I. (s.f.). *¿Qué es IAM?* Obtenido de AWS Identity and Access Management.: https://docs.aws.amazon.com/es_es/IAM/latest/UserGuide/introduction.html

Amazon Redshift, I. (s.f.). *¿Qué es Amazon Redshift?* Obtenido de https://docs.aws.amazon.com/es_es/redshift/latest/mgmt/welcome.html

Amazon Web Services, I. (s.f.). *¿Qué es AWS?* Obtenido de https://aws.amazon.com/es/what-is-aws/

Amazon Web Services, I. (s.f.). *Almacenamiento de datos seguro en la nube (S3)*. Obtenido de https://aws.amazon.com/es/s3/?trk=5970b1e9-218b-48cc-9862-f23c151d81b2&sc_channel=ps&s_kwcid=AL!4422!3!590443989054!e!!g!!amazon%20s3&ef_id=CjwKCAiA8OmdBhAgEiwAShr40w7eD-j6IMZl0xF1bs9Ohh7sjCBsPfaYf06WrwMzxB0fcdg2rwjuMBoCzUIQAvD_BwE:G:s&s_kwcid=AL!4422!3!5

García Márquez, I. (2017). *Big data management. Springer International.*

KeepCoding, R. (23 de Mayo de 2022). *Talend Open Studio KeepCoding Tech School*. Obtenido de https://keepcoding.io/blog/talend-open-studio/

LICA. (2020). *LICA Distribuidores exclusivos*. Obtenido de https://www.lica.com.sv/

PowerBI, M. (s.f.). *¿Qué es Power BI? Definición y características*. Obtenido de https://powerbi.microsoft.com/es-es/what-is-power-bi/

Ross, K. &. (2013). *The data warehouse toolkit (3rd ed.)*.

S.L, S. I. (2020). *Softland*. Obtenido de https://www.softland.cl/#:~:text=Es%20un%20completo%20Sistema%20de,la%20empresa%20del%20siglo%20XXI.

ANEXOS

Anexo 1: Diagrama de estrella del Data Warehouse del proceso de ventas