UNIVERSIDAD CATÓLICA DEL MAULE
FACULTAD DE CIENCIAS DE LA INGENIERÍA
ESCUELA INGENIERÍA CIVIL INFORMÁTICA

Implementación de una solución Business Intelligence utilizando Talend y Snowflake

Luis Fernando Inostroza Flores
Profesor Guía: Dr. Felipe Tirado Marabolí

Proyecto de título para optar al Título Profesional de Ingeniero Civil Informático

TALCA, SEPTIEMBRE 2023

UNIVERSIDAD CATÓLICA DEL MAULE

FACULTAD DE CIENCIAS DE LA INGENIERÍA

ESCUELA INGENIERÍA CIVIL INFORMÁTICA

Proyecto de título para optar al Título Profesional de Ingeniero Civil Informático.

Implementación de una solución de Business Intelligence utilizando Talend y Snowflake

Luis Inostroza Flores

COMISIÓN EXAMINADORA	**FIRMA**
Profesor de comisión interna Mg. Carlos Castro Bustamante Universidad Católica del Maule	_____
Profesor de comisión externa Mg. Natalia Morales Rosales Universidad Católica del Maule	_____
Profesor Guía Mg. Felipe Tirado Marabolí Universidad Católica del Maule	_____

NOTA FINAL EXAMEN DE TÍTULO _____

TALCA, SEPTIEMBRE 2023

Agradecimientos

En primer lugar, quiero agradecer a mi familia, quienes de una u otra manera, me ayudaron en este largo camino y a superar todas las adversidades que existieron. A mi madre María Flores, que ha dedicado su vida a cuidarme y estar siempre pendiente de mí. A mis hermanas Isabel y Paola, por ser mis segundas madres, y a sus familias que también son las mías.

También quisiera agradecer a los profesores que me apoyaron, en especial al profesor Felipe Tirado que me guio en el desarrollo de este proyecto. Así como a todos los que formaron parte de mi vida durante este período universitario.

MUCHAS GRACIAS.

Luis Fernando Inostroza Flores.

Septiembre, 2023.

Resumen

En los últimos años, la inteligencia empresarial se ha convertido en un pilar fundamental dentro de cada empresa, ayudando en las tomas de decisiones a nivel gerencial y al crecimiento de estas. Esto ha impulsado el desarrollo de nuevas tecnologías de integración de datos, que permitan flexibilizar la gobernanza de datos, y darles otro aspecto a los almacenes de datos masivos.

Uno de los retos más complicados a los que un profesional del área de datos se enfrenta al diseñar sistemas de información, son los sistemas heredados, los cuales deben ser migrados a sistemas más avanzados a medida que surgen nuevas tecnologías y se desarrollan dominios de aplicación, con el fin adaptarse a una funcionalidad mejorada y a modelos de negocio rediseñados. En este contexto nacen herramientas como Talend y Snowflake, que permiten dar un salto importante en el manejo de datos masivos.

Este proyecto de titulación tiene por objetivo implementar una solución de inteligencia empresarial utilizando dichas herramientas, realizando una migración de datos masivos, transformándolos a un nuevo modelo relacional y moldeándolos hacia una nueva tecnología de punta, de tipo datamart, que utilice tecnología cloud para su funcionamiento. Todo esto complementado con investigaciones que ayuden a su desarrollo.

Para la metodología, el proyecto adopta un enfoque basado en la metodología de Kimball, esto con el fin de darle un orden formal al desarrollo de la solución planteada, la cual constó de 6 etapas. Se realizó una toma de requisitos para evaluar necesidades del sistema. A partir de ello, se realizó una arquitectura técnica de la solución para poder dar una respuesta a los requerimientos. Posteriormente se diseñó el modelo relacional y finalmente la solución fue implementada, asegurando con ello su factibilidad.

Los resultados del proyecto incluyen informes de datos masivos obtenidos a raíz de la implementación de la solución y la compatibilidad de análisis de estos datos en base a los KPI requeridos.

Finalmente, el modelo propuesto garantiza un datamart a la vanguardia, dentro del cual sea posible mantener y acceder a una gran cantidad de datos, aprovechando las ventajas que la tecnología cloud brinda actualmente y su evolución para el futuro.

Abstract

In the last few years, business intelligence has become a mainstay within each company, supporting the decision-making at managerial level and helping the growth of it. This has driven the development of new data integration technologies that allows the flexibilities of data governance and giving data warehouse a whole new aspect.

In addition, one of the greatest challenges that a data professional must deal with when designing information systems are outdated systems, which must be migrated to newer systems as new technologies emerge and application domains develop with the purpose of adapting it to a better functionality and redesigned business models. In this context, technologies like Talend and Snowflake are born, which allows to take a big leap in big data management.

Therefore, the aim of this degree project is to implement a business intelligence solution using such tools, doing a big data migration, transforming them to a new relational model, and molding them towards cutting-edge technology, a datamart, that uses cloud for its functioning. All this complemented with scientific research that helps to its development.

Regarding to the methodology, the project embraces an approach based on Kimball's methodology, with the objective of giving a formal order to the development of the proposed solution, which consists of six stages. Firstly, a requirement gathering was carried out to evaluate the system needs. Since then, a technical architecture of the solution was made to respond to the requirements. Thereafter, the relational model was designed and lastly the solution was implemented, ensuring its feasibility.

The results of this project include big data reports obtained because of solution implementation and the compatibility of analysis of this data based on the required KPIs.

Finally, the proposed model grants a cutting-edge datamart, within which it is possible to provide and access to a big quantity of data, taking advantage of what cloud technologies gives us nowadays and its future evolutions.

Índice general

1. Introducción 1
1.1. Problema u oportunidad 2
1.2. Objetivos 2
1.2.1 Objetivo general 2
1.2.2 Objetivos específicos 2
1.3. Resultados esperados o contribuciones 2
1.4. Estructura del Trabajo 3
2. Estado del Arte 4
2.1. Introducción 5
2.2. Migración de datos y Talend Studio 5
2.3. ETL vs ELT 5
2.4. Integración de la nube a la fase ETL/ELT 7
2.5. Data Warehouse tradicional vs Data Warehouse cloud 8
2.6. El Data Warehouse elástico “Snowflake” 9
2.7. Conclusiones del capítulo 11
3. Marco Teórico 12
3.1. Introducción 13
3.2. Bases de datos relacional 13
3.3. Stored Procedures 13
3.4. Esquema copo de nieve 13
3.5. Business Intelligence 14
3.6. Big Data 14
3.7. Data Warehouse 14
3.8. Datamart 14
3.9. ETL 15
3.10. Talend Studio 15
3.11. Talend Remote Engine 16
3.12. Talend Cloud 16
3.13. Snowflake 17
3.14. ELT 17
3.15. Power BI 18

3.16. Metodología Kimball 19
3.13.1. Definición de requerimientos del negocio 19
3.13.2. Diseño de arquitectura técnica 19
3.13.3. Modelado dimensional 20
3.13.4. Especificación de aplicaciones de BI 20
3.13.5. Implementación 20
3.13.6. Mantenimiento y Crecimiento 20
3.17. Conclusiones del capítulo 20
4. Método Propuesto 21
4.1. Introducción 22
4.2. Requerimientos del negocio 22
4.2.1. Requerimientos 22
4.2.2. Indicadores de gestión 22
4.3. Diseño de arquitectura técnica 23
4.4. Modelado dimensional 24
4.4.1. Diagrama Lógico 24
4.4.2. Diagrama entidad-relación 25
4.5. Especificación de aplicaciones BI 27
4.6. Implementación 28
4.6.1. Extracción 28
4.6.2. Carga 29
4.6.3. Transformación 30
4.6.4. Análisis 31
4.7. Mantenimiento y crecimiento 32
4.8. Conclusiones del capítulo 32
5. Antecedentes 31
5.1. Introducción 32
5.2. Antecedentes técnicos y normativos 32
5.2.1. Antecedentes técnicos 32
5.2.2. Antecedentes normativos 32
5.3. Antecedentes económicos y financieros 33
5.4. Antecedentes Sociales y Ambientales 34

5.4.1. Impacto 34
5.5. Conclusiones 35
6. Planificación 36
6.1 Plan de trabajo 37
6.2 Carta Gantt 37
7. Resultados 39
7.1 Resultados Cualitativos 40
7.2 Resultados Cuantitativos 40
7.2.1 Integridad de las transacciones a nivel nacional 41
7.2.2 Incidentes ATM a nivel nacional 43
8. Discusión y Conclusiones 46
8.1. Discusión 47
8.2. Conclusiones 47
8.3. Limitaciones y posibles soluciones 48
8.4. Trabajos futuros 49
Bibliografía 42
9. Glosario 53
A. Detalle Jobs y SP 55
A.1. Jobs 56
A.1.1. Maestros 56
A.1.2. Vistas 62
A.2. Stored Procedures 64

Índice De Figuras

Figura 3.1: UI Talend Studio 16
Figura 3.2: UI Talend Cloud 17
Figura 3.3: UI Snowflake 17
Figura 3.4: Diagrama de flujo de trabajo ELT 18
Figura 3.5: UI de Power BI 18
Figura 3.6: Metodología Ralph Kimball. 19
Figura 4.1: Diagrama de arquitectura técnica. 24
Figura 4.2: Diagrama lógico. 25
Figura 4.3: Diagrama entidad relación. 26
Figura 4.4: Aplicaciones BI a utilizar (elaboración propia). 27
Figura 4.5: Proceso de extracción. 28
Figura 4.6: Proceso de carga de datos. 29
Figura 4.7: Proceso de Transformación. 30
Figura 4.8: Funcionamiento Stored Procedures. 31
Figura 4.9: Proceso de análisis 31
Figura 7.1: Modelo Analítico en Snowflake 40
Figura 7.2: Dashboard transacciones. 41
Figura 7.3: Dashboard transacciones por tipo. 42
Figura 7.4: Dashboard transacciones por región. 43
Figura 7.5: Resumen Incidentes. 44
Figura 7.6: Incidentes por downtime. 45
Figura 7.7: Incidentes v/s downtime 45
Figura A.1: Job Aprobado_Rechazado en Talend Studio 56
Figura A.2: Job BIN en Talend Studio 57
Figura A.3: Job CANAL en Talend Studio 57
Figura A.4: Job FECHA en Talend Studio 58
Figura A.5: Job INSTITUCION_ADQUIRIENTE en Talend Studio 59
Figura A.6: Job INSTITUCION_EMISORA en Talend Studio 59
Figura A.7: Job METODO_AUTORIZACIÓN en Talend Studio 60
Figura A.8: Job TIEMPO en Talend Studio 60
Figura A.9: Job TIPO_PROCESAMIENTO en Talend Studio 61

Figura A.10: Job TIPO_TRANSACCIÓN en Talend Studio 62
Figura A.11: Job DWH en Talend Studio 62
Figura A.12: Job SP_TRANSACCIONES en Talend Studio 63
Figura A.13: Job PROVIEW-ATM en Talend Studio 63
Figura A.14: Job PROVIEW-EVENTOS en Talend Studio 64
Figura A.15: Diagrama funcionamiento SP_CARGARATM 65
Figura A.16: Diagrama funcionamiento SP_CARGARLOCALIDAD 65
Figura A.17: Diagrama funcionamiento SP_CARGARCANAL 66
Figura A.18: Diagrama funcionamiento SP_CARGARATMACANAL 66
Figura A.19: Diagrama funcionamiento SP_CARGARLOCALIDADACANAL 67
Figura A.20: Diagrama funcionamiento SP_CARGARTRANSACCIONES 67
Figura A.21: Diagrama funcionamiento SP_CARGARBINTRX 68
Figura A.22: Diagrama funcionamiento SP_CARGARMETODOAUTORIZACIONTRX.... 68
Figura A.23: Diagrama funcionamiento SP_CARGARCANALTRX 69
Figura A.24: Diagrama funcionamiento SP_CARGARTIPOTRANSACCIONTRX..... 69
Figura A.25: Diagrama funcionamiento SP_CARGARPROCESAMIENTOTRX 70
Figura A.26: Diagrama funcionamiento SP_UPDATEIDFECHA 70
Figura A.27: Diagrama funcionamiento SP_UPDATETIEMPO 71
Figura A.28: Diagrama funcionamiento SP_CARGARTIEMPOTRX 71
Figura A.29: Diagrama funcionamiento SP_CARGARATM_ACTUAL 71
Figura A.30: Diagrama funcionamiento SP_CARGARINCIDENTES 72

Índice de Tablas

Tabla 2.1: Comparativa ETL vs ELT. [4]..... 7
Tabla 2.2 : Comparativa DWH tradicional vs DWH cloud..... 9
Tabla 5.1: Costo de personal..... 33
Tabla 5.2: Costo de licencias..... 34
Tabla 6.1: Carta Gantt del proyecto..... 38

Capítulo 1

1. Introducción

En este capítulo se presenta el tema, problema y oportunidad, se situará en un contexto donde se describirán los objetivos (general y específicos) y lo que se espera del trabajo o contribuciones de este.

1.1. Problema u oportunidad

En la actualidad, y como viene siendo una tendencia hace varios años, la demanda de datos está creciendo a pasos agigantados, y seguirá en ese camino debido a que la información es la base de cada organización. Es por esto por lo que las empresas buscan mejorar sus sistemas de análisis de datos, y para lograrlo deben modernizarse constantemente, debido a la compatibilidad de nuevas tecnologías en el mercado.

La problemática del proyecto reside en que una empresa del rubro bancario tiene actualmente un reporte de datos contables, contenido en un modelo relacional heredado y poco ágil, lo cual les ralentiza el proceso de análisis de datos masivos. Dicha compañía ha solicitado mejorar su procesamiento y capacidad analítica, hoy limitada por capacidades de la infraestructura tecnológica actual, además de no contar con un modelo con verdadera orientación analítica dentro de las herramientas en uso por la gerencia y la compañía.

1.2. Objetivos

1.2.1 Objetivo general

Implementar una solución de Business Intelligence (BI) utilizando Talend y Snowflake para recopilar, almacenar y visualizar datos de manera eficiente, para mejorar la toma de decisiones en áreas clave dentro de la compañía.

1.2.2 Objetivos específicos

En orden de lograr el objetivo general, se plantean los siguientes subobjetivos:

1. Llevar a cabo un levantamiento y análisis de requerimientos que la empresa posee, junto con ello sus indicadores de gestión.
2. Realizar un modelado de datos, diseñando nuevas tablas de información y utilizando aplicaciones que nos permitan integrar los datos mediante una estructura definida.
3. Implementar un repositorio analítico de tipo datamart en Snowflake, minimizando los riesgos de pérdida de información y facilitando el acceso a los datos requeridos.
4. Creación de reportes visuales de datos de tipo dashboard, esenciales para la empresa y basados en los indicadores de gestión requeridos.

1.3. Resultados esperados o contribuciones

La implementación de una plataforma que utilice la inteligencia empresarial para ajustar el funcionamiento de la empresa en cuestión favorecerá principalmente el acceso a los datos, lo

que facilitará el estudio de la información contable. Junto con ello la reducción de costos, ya que el tener una buena plataforma de manejo de información reducirá la cantidad de personal especializado que deba contratarse para analizar estos datos. Además, el proyecto facilitará la colaboración entre diferentes áreas dentro de la organización, lo que apoyará a la junta directiva en sus decisiones.

Finalmente, el desarrollo de este sistema y la documentación que salga de él apoyará a largo plazo los proyectos futuros que presenten requisitos similares, siendo así un puntapié inicial para futuros analistas e ingenieros de datos.

1.4. Estructura del Trabajo

Este documento, se descompone como sigue:

- En el capítulo 2, se presenta el estado del arte, el cual permite dar un contexto claro sobre las tecnologías a utilizar en el mercado mundial, identificando investigaciones relevantes que apoyen al proyecto actual mirando desde otras perspectivas.
- En el capítulo 3, se presenta el marco teórico, en donde se definirán conceptos importantes para la comprensión del documento.
- En el capítulo 4, se presenta la solución propuesta, explicando como el proyecto dará solución a la problemática y como es que la hipótesis será comprobada. Dando las explicaciones necesarias para clarificar la solución.
- En el capítulo 5, se presentan los antecedentes técnicos y normativos, económicos y financieros, sociales y ambientales.
- En el capítulo 6, se presenta la planificación, estipulando fechas y tareas utilizando una carta Gantt para que el proyecto se realice correctamente.
- En el capítulo 7, se presentan los resultados, discusión, conclusiones y trabajos futuros.
- En el capítulo 8, se describe la discusión y conclusiones del proyecto, dando las ideas principales obtenidas del proyecto y como se afrontarán las limitaciones de este en el futuro.
- Luego se presenta la bibliografía, dando enlaces a las fuentes y comprobación de referencias existentes en ellas.
- En el capítulo 9, se presenta el glosario, explicando términos técnicos con su significado.
- Finalmente, se presentan los anexos de la investigación, donde quedan en evidencia archivos que el proyecto necesite para evidenciar lo descrito en capítulos anteriores.

Capítulo 2

2. Estado del Arte

En este capítulo se presenta el estado de la técnica asociado al proyecto a desarrollar.

2.1. Introducción

A continuación, se dará un contexto de las tecnologías y técnicas a utilizar, y como estas se encuentran hoy en el mercado mundial, utilizando documentos científicos actualizados que ayuden a respaldar este trabajo de titulación.

2.2. Migración de datos y Talend Studio

La migración de datos es un método de transferir datos desde un sistema a otro. Es frecuentemente una subactividad de la implementación de una aplicación de BI. Como parte de esta actividad, es normal que las empresas orientadas al manejo de datos muevan sus desarrollos a Talend para gestionar de forma eficiente datos que conduzcan iniciativas de transformación digital. Los beneficios empresariales que conlleva usar Talend se extienden a mejorar la capacidad de integración de datos, la gobernación de datos, y también a todo lo relacionado con la integración de aplicaciones y API a los proyectos [1].

Talend Studio ha sido adoptado por muchas organizaciones para propósitos de integración de datos. Esto muestra que la herramienta es muy confiable. La sincronización de datos, migración de datos de una base a otra, y el compartimiento de archivos entre sistemas están todos protegidos por esta herramienta. En el mundo actual, Talend se ha convertido en el enfoque de integración de datos más poderoso. Contiene cada característica necesaria para la integración de datos y su sincronización [2].

2.3. ETL vs ELT

ETL y ELT son dos conceptos hermanos, básicamente el mismo flujo de trabajo, pero en distinto orden. La razón por la cual ELT ha ido de a poco reemplazando a ETL, es por la existencia de warehouses en la nube cada vez más capaces de realizar distintas tareas aparte del almacenamiento, entre ellas la transformación de datos, como es el caso del proyecto actual.

En cuanto a su comportamiento, existen trabajos comparativos de estos dos flujos de trabajo, como, por ejemplo, la investigación realizada por Bharat Singal [3], quien realiza una comparación de ambos flujos en cuanto a rendimiento respecta.

Ambos flujos son útiles y dependen de la implementación y requerimientos de cada proyecto de inteligencia empresarial. La tabla 2.1 compara ambos flujos en distintos aspectos clave.

	ETL	ELT
Adopción de la tecnología	ETL ha estado en el mercado por casi 2 décadas y es relativamente fácil encontrar desarrolladores con experiencia en este flujo de trabajo.	ELT es una nueva tecnología que está enfocada en data warehouses basados en Cloud. Encontrar ingenieros adecuados para desarrollar flujos de trabajo ELT es tan fácil como para ETL.
Disponibilidad de datos	En una carga de trabajo ETL, solo los datos que son requeridos para el análisis son cargados al warehouse, dejando los datos innecesarios en el sistema tal como están.	Mientras que en un sistema ELT, tendemos a cargar absolutamente todos los datos hacia un warehouse o data lake desde donde puede ser analizado en otro punto de tiempo.
Campos calculados y transformaciones	Si, en ETL, podemos añadir o eliminar columnas específicas mientras transformamos los datos en la herramienta ETL. También podemos añadir columnas calculadas y cargarlas hacia el data warehouse.	En ELT, columnas adiciones son directamente añadidas al set de datos existente dentro del warehouse. Usualmente, no hay modificaciones de las columnas fuente.
Complejidades de transformación	En una carga de trabajo ETL, podemos implementar tantas transformaciones de datos complejas como y cuando sea requerido. Todas esas transformaciones ocurren en la memoria.	En una carga de trabajo ELT, el mayor foco es dado hacia analizar datos estructurados y no estructurados altamente variables que llegan en un ritmo rápido, más que complejidades de transformación.

Infraestructura	La mayoría de las herramientas tradicionales ETL necesitan estar instaladas on-premise lo que puede derivar en altos costos para la carga de datos analítica.	ELT, por otro lado, es mayormente basado en cloud y no requiere ser instalado on-premise.
Mantenimiento Postproducción	En un flujo ETL, que es instalado on-premise, el mantenimiento es requerido frecuentemente.	ELT, al ser basado en la nube o en un ambiente sin servidor, no requiere, o requiere muy poco mantenimiento.
Área de transformación	En un flujo de trabajo ETL, las transformaciones son aplicadas en la memoria en una capa de preparación antes de que los datos sean cargados hacia el data warehouse.	En ELT, las transformaciones son aplicadas una vez que los datos han sido cargados al warehouse o data lake. En este caso, usualmente, no se necesita una capa de preparación como en ETL.
Soporte para datos semiestructurados y no estructurados	Aunque una herramienta ETL es capaz de leer datos desde fuentes de datos semiestructurados y no estructurados, estos datos son usualmente transformados en la capa de preparación y son almacenados como una estructura adecuada en el data warehouse.	ELT está diseñado para aguantar todo tipo de estructuras de datos desde semiestructurados hasta no estructurados en los data lakes donde serán posteriormente analizados.

Tabla 2.1: Comparativa ETL vs ELT. [4]

2.4. Integración de la nube a la fase ETL/ELT

La adaptación de la fase ETL al contexto Big Data / Cloud, caracterizado por largos volúmenes de datos remotos y heterogéneos, se ha vuelto una necesidad. Hoy en día, el cloud se ha convertido en una prometedora solución para manejar datos a larga escala. Los investigadores de este contexto mencionan que uno de los grandes desafíos a los que se enfrenta este proceso ETL es el uso de tecnologías semánticas, y también de como la literatura actual

brinda soluciones que no consideran el dinamismo y la elasticidad de los datos contenidos en cloud. Por lo tanto, los mismos investigadores plantean trabajar en una herramienta ETL que explote las ventajas de la computación en la nube y a su vez considere las 3 características del Big Data (Volumen, Variedad y Velocidad) [5] . Esta es una de las principales razones por las cuales se espera que ELT experimente un crecimiento exponencial en cuanto a usuarios.

2.5. Data Warehouse tradicional vs Data Warehouse cloud

Un data warehouse (DWH) tradicional provee un soporte completo de SQL, pero carece de la característica de escalabilidad [6]. Se requiere mucho tiempo para configurar, optimizar y administrar el sistema. Hoy en día, las organizaciones están moviéndose hacia data warehouse cloud. La idea de un data warehouse cloud fue introducido para suplir dichas necesidades. Un data warehouse cloud es totalmente diferente a un sistema tradicional de data warehouse. Usando un DWH cloud los clientes pueden obtener datos desde diferentes zonas geográficas y resuelve muchos problemas que las tradicionales no, principalmente el acceso a los datos de manera remota. La tabla 2.2 ilustra diferencias clave entre DWH tradicional y DWH cloud [6] .

	DWH Tradicional	**DWH Cloud**
Planificación	La etapa de planificación para una DWH tradicional puede ser una tarea inmensa	No necesitas planificar tu DWH usando cloud
Flexibilidad y costo	Un DWH tradicional es inflexible y puede resultar en un sobre aprovisionamiento y pagos de más	Un DWH cloud es flexible y crece automáticamente cuando sus necesidades aumentan. Usar DWH cloud reduce costos
Efecto en queries	Un DWH tradicional afecta las queries cuando los datos crecen	Un DWH cloud no afecta a las queries cuando crecen los datos
Disminución por inactividad	Un DWH tradicional no se encoge cuando es poco utilizado	Un DWH cloud se encoge automáticamente cuando se utiliza poco para ahorrar gastos
Criterio de selección	El criterio para seleccionar un	El criterio para seleccionar un

	DWH tradicional es un volumen de datos inmenso sin aumentar el número de usuarios	DWH cloud es un volumen dinámico y elástico de datos que incremente el número de usuarios
Ampliación de datos	Usando un DWH tradicional no podemos ampliar y reducir datos en cualquier momento. Toma horas o días configurar el hardware, software y la infraestructura. Además, no puede manejar el incremento de usuarios	DWH construido para cloud puede "ampliar y reducir datos en cualquier momento". No tomará horas o días para configurar hardware, software e infraestructura. Además, puede manejar el incremento de usuarios.
Mantenimiento postproducción	Un DWH tradicional es caro y no es fácil de acomodar y mantener cuando la cantidad de datos crece	Un DWH construido para cloud es sencillo y barato para acomodar y mantener
Soporte para tipos de datos	Un DWH tradicional no puede manejar tipos de datos diversos. Solo maneja datos estructurados	DWH cloud es capaz de manejar distintos tipos de datos, es decir, estructurados y semiestructurados
Flexibilidad de espacio de trabajo	Un DWH tradicional te obliga a comprar computación y almacenamiento juntos	DWH cloud te permite cambiar el tamaño de los clústeres de computación (solo pagas por lo que necesitas, cuando lo necesitas)

Tabla 2.2: Comparativa DWH tradicional vs DWH cloud.

2.6. El Data Warehouse elástico "Snowflake"

La llegada de la tecnología cloud marca una salida de entrega de software y su ejecución en servidores locales, hacia centros de datos compartidos y soluciones SaaS (software-as-a-service) ofrecidos por proveedores de plataforma tales como Amazon, Google o Microsoft. Pero no solo las plataformas han cambiado, también los datos. Anteriormente los datos de un data warehouse

solían venir desde la misma empresa, por lo que la estructura, volumen y proporción de los datos eran todos fácilmente predecibles y con límites bien conocidos. En la actualidad los datos generalmente vienen de fuentes externas y de difícil control, además de poseer datos poco estructurados y sin esquema definido. En este contexto nace el data warehouse elástico "Snowflake" o simplemente Snowflake, un DWH de tipo cloud que nace como alternativa a los ya existentes en el mercado, y que según un artículo científico realizado por los mismos desarrolladores de la plataforma [7], posee las siguientes características que lo diferencian del resto de competidores:

- **Experiencia Software-as-a-service (SaaS) pura:** Los usuarios no necesitan comprar máquinas, contratar a administradores de bases de datos, o instalar software. Los usuarios pueden tener desde un principio sus datos en la nube o cargarlos. Luego pueden inmediatamente manipular o consultar sus datos usando la interfaz gráfica de Snowflake o interfaces estandarizadas como por ejemplo ODBC. En contraste con otros servicios de Database-as-a-Service (Daas), el aspecto de Snowflake se extiende a la experiencia de usuario completa. No hay perillas de sintonización, ni diseño físico, ni tareas de preparación de almacenamiento por parte del usuario.
- **Relacional:** Snowflake posee soporte comprensivo para transacciones SQL. La mayoría de los usuarios son capaces de migrar cargas de trabajo existentes con muy pocos o sin cambios.
- **Semiestructurado:** Snowflake ofrece funciones integradas y extensiones SQL para poder realizar atravesamiento, aplanado y unión de datos semiestructurados, con soporte para formatos populares como JSON y AVRO.
- **Elástico:** Almacenamiento y recursos computacionales pueden ser escalados independiente y perfectamente, sin ningún impacto en la disponibilidad de datos o el desempeño en las consultas concurrentes del sistema.
- **Altamente disponible:** Snowflake tolera fallas de los nodos, clústeres, e incluso fallas completas del centro de datos. No hay tiempo muerto para el usuario durante la actualización de software o hardware.
- **Duradero:** Snowflake está diseñado para una extrema durabilidad, con protecciones adicionales contra pérdida de datos accidental: clonación, deshacer, y respaldos entre regiones.

- **Costo-Eficiente:** Snowflake es altamente computacional-eficiente y todas las tablas de datos están comprimidas. Los usuarios pagan solo por el almacenamiento y recursos computacionales que ocupan.
- **Seguro:** Todos los datos que incluyen archivos temporales y tráficos de red son encriptados de manera end-to-end. Nunca habrá exposición de datos de usuario hacia la plataforma cloud. Adicionalmente, el control de acceso basado en roles les da a los usuarios la habilidad de ejercer un control de acceso detallado a nivel de SQL.

Lo anterior posiciona a Snowflake como uno de los principales DWH cloud en el mercado actual, y uno de los top 5 mejor catalogados para el año 2023 a nivel global [8], así como una muy buena elección para manejar proyectos de Big Data.

2.7. Conclusiones del capítulo

En esta sección se ha descrito la actualidad del paradigma investigativo en lo que a las tecnologías respecta, además de varias referencias que acrediten la vanguardia del proyecto. Dentro de ello tenemos dos aspectos claves que señalar.

El primero tiene que ver con la elección de Snowflake como DWH a utilizar, y es que hoy en día la tecnología en la nube crece a pasos agigantados, y además de brindarnos almacenamiento para usuarios convencionales también se adapta a las necesidades de datos masivos como los presentes en este proyecto, en palabras simples, Snowflake cubre todas las necesidades de la solución en cuanto a almacenamiento y compartición de la información a distintos usuarios, gracias a sus características mencionadas en el punto 2.6., además de ser un requisito funcional de la empresa cliente.

El segundo tiene que ver con la elección de un flujo de trabajo ELT sobre un ETL, ya que aprovechando la tecnología que nos brinda el DWH seleccionado, podemos realizar la transformación de datos una vez que ya se encuentren almacenados, por lo que un flujo de este tipo se adapta a las necesidades del proyecto, además de ser un requisito funcional de la empresa cliente.

Capítulo 3

3. Marco Teórico

En este capítulo, se presenta el Marco teórico que sustente los elementos para resolver la problemática, explicando conceptos importantes del área.

3.1. Introducción

A continuación, se definirán conceptos importantes que serán utilizados en el desarrollo del documento. Estos conceptos son importantes en el ámbito del análisis de datos, y servirán para dar una buena fuente de conocimiento técnico para entender de una mejor forma el modelo propuesto. Además, se utilizarán definiciones precisas para evitar confusiones.

3.2. Bases de datos relacional

Una base de datos relacional es un tipo de base de datos que almacena y proporciona acceso a puntos de datos relacionados entre sí. Las bases de datos relacionales se basan en el modelo relacional, una forma intuitiva y directa de representar datos en tablas. En una base de datos relacional, cada fila en una tabla es un registro con una ID única, llamada clave. Las columnas de la tabla contienen los atributos de los datos y cada registro suele tener un valor para cada atributo, lo que simplifica la creación de relaciones entre los puntos de datos.

El modelo relacional es sencillo pero muy potente, y lo utilizan organizaciones de todos los tipos y tamaños para una gran variedad de aplicaciones con datos. Las bases de datos relacionales se usan para rastrear inventarios, procesar transacciones de comercio electrónico, administrar cantidades enormes y esenciales de información de clientes y mucho más. Las bases de datos relacionales se pueden emplear para cualquier aplicación de datos en la que los puntos de datos se relacionen entre sí y deban gestionarse de forma segura, conforme a normas y de un modo uniforme [9].

3.3. Stored Procedures

Un Stored Procedure (o procedimiento almacenado, abreviado SP) es un fragmento preparado de código en lenguaje SQL el cual tiene la característica de poder ser guardado para poder reusarse una y otra vez, además de poder recibir parámetros y así actuar de acuerdo con estos parámetros recibidos.

Los Stored Procedures son creados para llevar a cabo una o más operaciones de manipulación de datos en una base de datos. En palabras simples, no es nada más que un grupo de sentencias SQL que aceptan entradas en forma de parámetros y realizan una tarea, la cual, dependiendo de su funcionamiento, devuelve o no un valor [10].

3.4. Esquema copo de nieve

Un esquema de copo de nieve es un modelo de datos multidimensional que es una extensión

de un esquema de estrella, donde las tablas de dimensiones se descomponen en subdimensiones. Los esquemas de copo de nieve se utilizan comúnmente para inteligencia empresarial e informes en almacenes de datos OLAP, almacenes de datos departamentales y bases de datos relacionales.

En un esquema de copo de nieve, los ingenieros descomponen las tablas de dimensiones individuales en subdimensiones lógicas. Esto hace que el modelo de datos sea más complejo, pero puede ser más fácil de manejar para los analistas, especialmente para datos masivos [11].

3.5. Business Intelligence

Business Intelligence o inteligencia empresarial, abreviado BI, combina análisis de negocios, minería, visualización, herramientas e infraestructura de datos, además de prácticas recomendadas para ayudar a las empresas a tomar decisiones basadas en los datos. En la práctica, implementar la inteligencia de negocios moderna implica contar con una vista integral de todos los datos de la organización. Además, consiste en usar estos datos para impulsar el cambio, eliminar las ineficiencias y adaptarse rápidamente a los cambios del mercado o la demanda [12].

3.6. Big Data

El término “Big Data” abarca datos que contienen una mayor variedad y que se presentan en volúmenes crecientes y a una velocidad superior. Esto también se conoce como “las tres V”.

Dicho de otro modo, el big data está formado por conjuntos de datos de mayor tamaño y más complejos, especialmente procedentes de nuevas fuentes de datos. Estos conjuntos de datos son tan voluminosos que el software de procesamiento de datos convencional sencillamente no puede gestionarlos. Sin embargo, estos volúmenes masivos de datos pueden utilizarse para abordar problemas empresariales que antes no hubiera sido posible solucionar [13].

3.7. Data Warehouse

Un data warehouse (DWH) es una base de datos centralizada y consolidada que se utiliza para almacenar grandes cantidades de datos históricos y actuales de diferentes fuentes de datos dentro de una organización. El objetivo principal de un data warehouse es proporcionar un entorno de almacenamiento optimizado y estructurado para facilitar el análisis de datos y la toma de decisiones empresariales [14].

3.8. Datamart

Un datamart es un subconjunto de un data warehouse que está diseñado para satisfacer las necesidades de un departamento o área específica dentro de una organización. Proporciona datos preprocesados, organizados y estructurados para facilitar el análisis y la toma de decisiones en ese ámbito en particular [15].

3.9. ETL

ETL, que significa extraer, transformar y cargar, es un proceso de integración de datos que combina datos de múltiples fuentes de datos en un único almacén de datos consistente que se carga en un almacén de datos u otro sistema de destino. Un flujo ETL se recomienda frecuentemente para repositorios de datos más pequeños y que requieren actualizaciones menos frecuentes, mientas que otros métodos, como el ELT, se utilizan para integrar volúmenes mucho mayores de datos [16].

3.10. Talend Studio

Talend Studio (o Talend como tal) es una herramienta de integración de datos de código abierto y una plataforma de integración de datos. Proporciona una interfaz gráfica intuitiva y potente para diseñar, desarrollar y administrar flujos de trabajo de integración de datos. Talend es compatible con diversos sistemas de bases de datos, aplicaciones empresariales y tecnologías de almacenamiento, lo que lo convierte en una elección popular para la extracción, transformación y carga de datos. Tiene una versión Open Source y de paga. Transforma Jobs a código Java automáticamente [17]. En la figura 3.1 se aprecia la interfaz de usuario de Talend Studio.

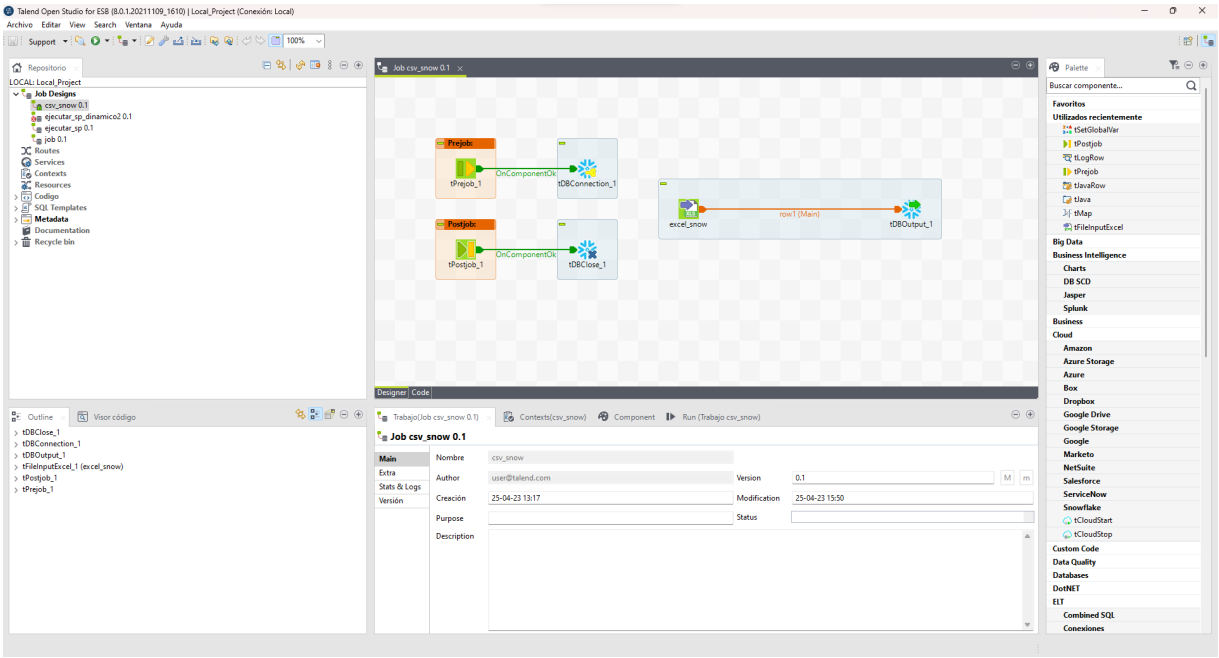


Figura 3.1: UI Talend Studio

3.11. Talend Remote Engine

La plataforma Talend Remote Engine, ofrecida por la empresa Talend, permite ejecutar flujos de trabajo de integración de datos y procesamiento de datos en entornos remotos y distribuidos. Actúa como un servidor de ejecución que se comunica con los diferentes componentes y servicios de Talend para llevar a cabo tareas de integración de datos de manera eficiente y escalable [18].

3.12. Talend Cloud

Talend Cloud es una plataforma de integración de datos y administración del ciclo de vida de datos basada en la nube ofrecida por Talend. Proporciona una amplia gama de capacidades para la integración de datos, el procesamiento de datos en tiempo real, la calidad de datos y la gestión de API. Talend Cloud permite a los usuarios diseñar, desarrollar y desplegar flujos de trabajo de integración de datos de manera ágil y escalable en un entorno en la nube [19]. En la figura 3.2 podemos ver la interfaz de la aplicación web.

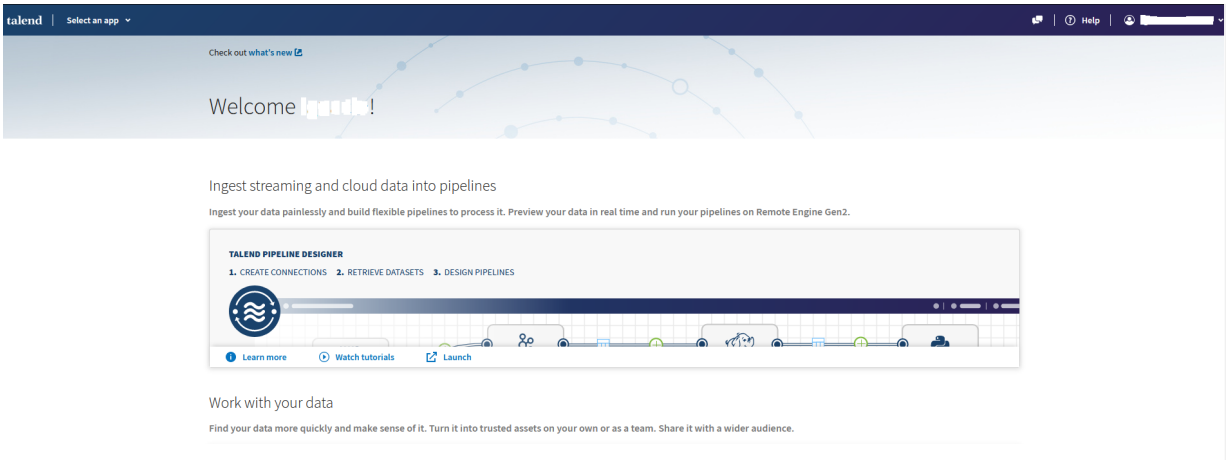


Figura 3.2: UI Talend Cloud

3.13. Snowflake

Snowflake es una base de datos en la nube diseñada específicamente para Cloud. Proporciona un entorno de almacenamiento y procesamiento escalable y eficiente para grandes volúmenes de datos. Es conocido por su arquitectura de almacenamiento separada de cómputo, que permite escalar los recursos de cómputo y almacenamiento de forma independiente según las necesidades. Cuenta con un sistema de pago-uso, en donde el cliente del servicio costea sus gastos a medida que ocupa capacidad de cómputo dentro del mismo software [20]. En la figura 3.3 podemos ver su interfaz.

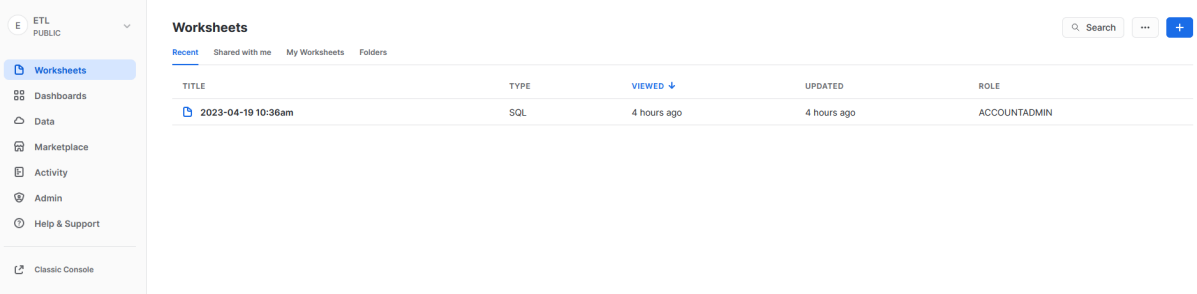


Figura 3.3: UI Snowflake

3.14. ELT

ELT, que significa extracción, carga y transformación es un enfoque de procesamiento de datos en el que los datos se extraen de las fuentes de origen, se cargan directamente en la plataforma de destino y luego se transforman dentro de la misma. A diferencia del enfoque tradicional ETL, donde los datos se transforman antes de cargarlos en el destino, ELT aprovecha el poder de procesamiento de la plataforma de destino para realizar las transformaciones después de la carga inicial. La figura 3.4 muestra visualmente el flujo de trabajo ELT [21].

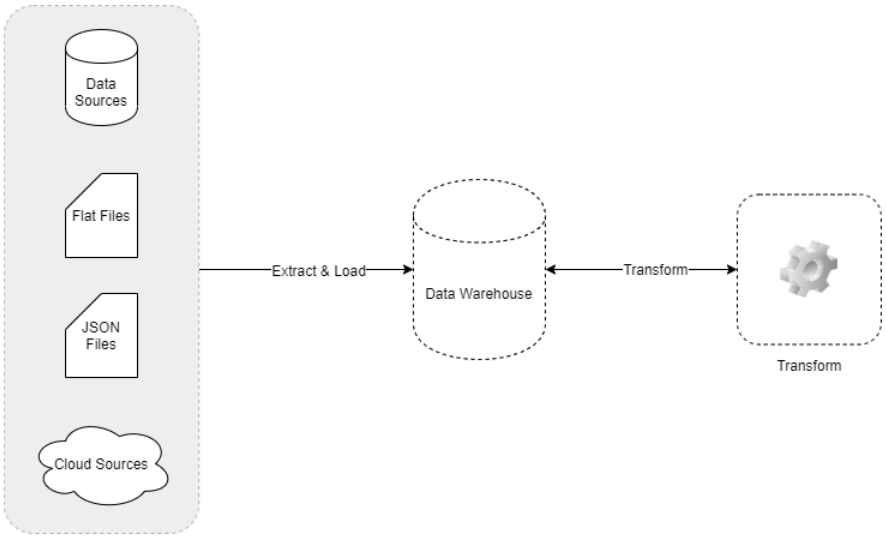


Figura 3.4: Diagrama de flujo de trabajo ELT.

3.15. Power BI

Power BI es una solución de análisis empresarial basado en la nube, que permite unir diferentes fuentes de datos, analizarlos y presentar un análisis de estos a través de informes y paneles. Con Power BI se tiene de manera fácil acceso a datos dentro y fuera de la organización casi en cualquier dispositivo. Cuenta con varias aplicaciones dentro de su estructura, pero la más importante y que se utilizará en el proyecto actual es Power BI Desktop, una aplicación gratuita de escritorio para transformar, visualizar datos y crear informes de estos [22]. La figura 3.5 muestra la interfaz de este software.

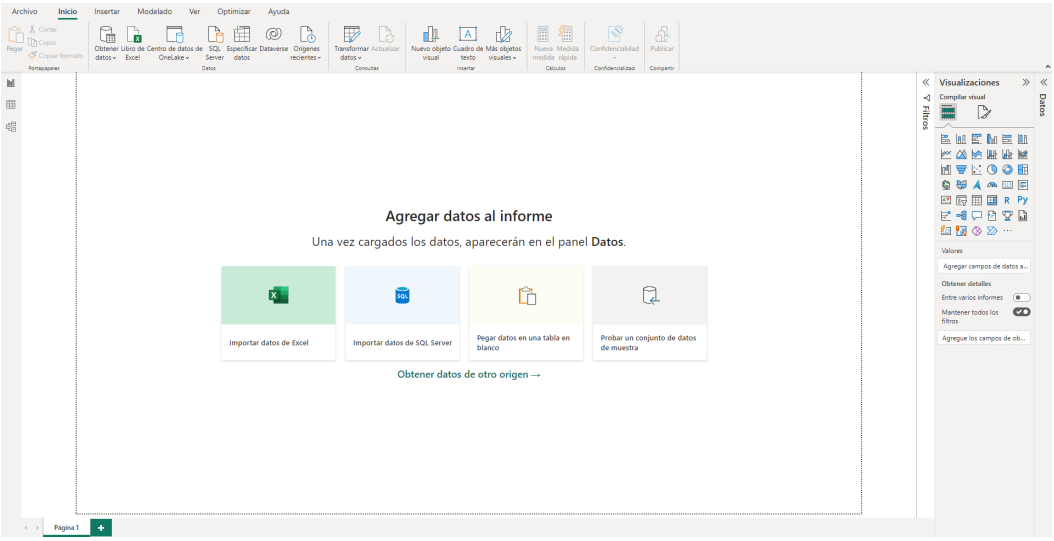


Figura 3.5: UI de Power BI

3.16. Metodología Kimball

Para asegurar el éxito de un proyecto, es necesario estructurar los procesos de manera adecuada, y es aquí donde las metodologías brindan este apoyo necesario. La metodología Kimball se basa en denominación del ciclo de vida dimensional del negocio, y facilita la creación de un DWH, ya que, de acuerdo con las distintas tareas de una organización, permite la creación de datamarts específicos, los cuales darán forma al futuro modelo general del sistema [23].

El proyecto cuenta con 4 etapas de esta metodología, las cuales se ilustran en la figura 3.6, y se explican a continuación [24].

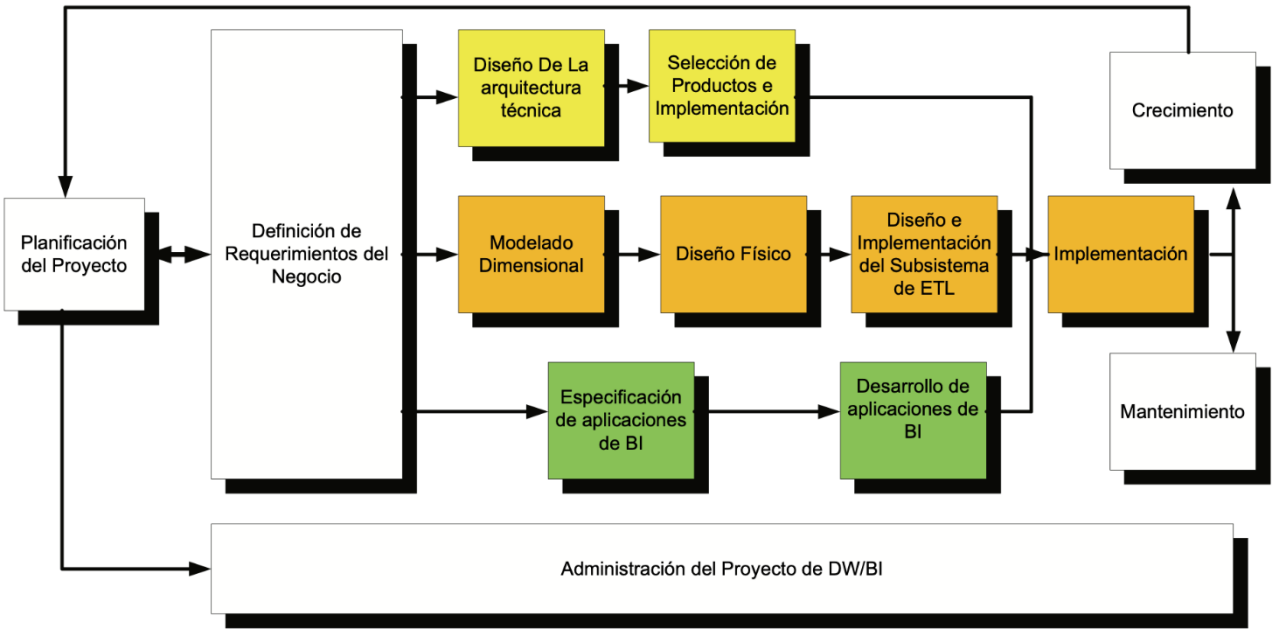


Figura 3.6: Metodología Ralph Kimball.

3.13.1. Definición de requerimientos del negocio

En este paso, se trabaja en estrecha colaboración con los usuarios y los interesados del negocio para comprender y documentar completamente los requisitos y necesidades que el data warehouse debe satisfacer. Se identifican las principales áreas de análisis, las preguntas clave que deben responderse y los indicadores clave de rendimiento (KPI) que se deben medir. Estos requerimientos del negocio servirán como base para el diseño del data warehouse y asegurarán que la solución entregada sea útil y relevante para los usuarios.

3.13.2. Diseño de arquitectura técnica

En este paso, se define la arquitectura técnica del data warehouse, lo que incluye la selección

de las tecnologías, herramientas y plataformas que se utilizarán en su construcción.

3.13.3. Modelado dimensional

El modelado dimensional es una técnica que se utiliza para diseñar la estructura de la base de datos del data warehouse. Se basa en el uso de esquemas de estrella o copo de nieve, que son estructuras optimizadas para facilitar el análisis y la consulta de datos. En este paso, se identifican las dimensiones (atributos descriptivos) y las medidas (datos cuantitativos) que serán utilizadas en el data warehouse, y se construyen las tablas dimensionales y las tablas de hechos que conformarán el esquema dimensional.

3.13.4. Especificación de aplicaciones de BI

En este paso, se definen y diseñan las aplicaciones de inteligencia de negocios (BI) que se construirán sobre el data warehouse. Estas aplicaciones pueden incluir paneles de control, informes, análisis ad hoc y otras herramientas que permitirán a los usuarios obtener información y conocimientos a partir de los datos almacenados en el data warehouse.

3.13.5. Implementación

En este paso, se lleva a cabo la construcción e implementación del data warehouse y las aplicaciones de BI definidas en los pasos anteriores. Se realiza el desarrollo y la configuración de la infraestructura técnica, la implementación de los procesos ETL para cargar los datos en el data warehouse, y la creación de las aplicaciones de BI según las especificaciones establecidas.

3.13.6. Mantenimiento y Crecimiento

En este último paso, se realiza una capacitación a los usuarios respecto al sistema construido, esto con el fin de evitar futuras fallas humanas en el uso del data warehouse, así como garantizar que el trabajo fue realizado fehacientemente y documentar posibles errores para que estos no afecten en su crecimiento a largo plazo [23].

Es importante señalar que las etapas de la metodología de Kimball no obedecen un orden estricto, y su funcionamiento se adapta a cada proyecto de Business Intelligence.

3.17. Conclusiones del capítulo

En esta sección se han descrito los principales conceptos técnicos para poder entender claramente el desarrollo del proyecto, a partir de este marco teórico, podemos ahora avanzar hacia la implementación y el análisis de nuestro proyecto con un enfoque informado y una comprensión profunda de los elementos clave involucrados en el proceso.

Capítulo 4

4. Método Propuesto

En este capítulo, se presenta el método propuesto como solución a la problemática.

4.1. Introducción

A continuación, se detalla cada una de las etapas la metodología que se utilizará para resolver la problemática del proyecto. Al ser una relativamente nueva tecnología, el flujo de trabajo ELT empleado en la solución no cuenta hoy en día con una metodología ampliamente aceptada, es por ello por lo que el método propuesto sigue la metodología Kimball, normalmente usada para ETL, la cual consta de 6 fases de desarrollo a ser explicadas paso a paso.

Es importante señalar que el método propuesto plantea una solución intermedia a la versión final, por lo que podría estar sujeta a cambios en un futuro.

4.2. Requerimientos del negocio

El primer paso de la solución conlleva a realizar un levantamiento de requerimientos, las que se traducen en reuniones con el cliente para estipular cuales son las necesidades específicas que tengan por su parte. Estas cuentan con la participación de la gerencia de proyectos de la empresa junto con el cliente que realiza el requerimiento.

4.2.1. Requerimientos

La empresa cliente manifestó su búsqueda por mejorar su sistema de almacenamiento actual, ya que el que utiliza actualmente se basa en un OLTP en base de datos Oracle, y no cuenta con un proceso de integración de datos, lo que les ocasiona lentitud en sus consultas hacia la información que esta base de datos contiene. Es por ello por lo que han requerido reducir sus tiempos de consulta, costos y flexibilizar su análisis. En otras palabras, lo que se busca mediante los requerimientos es tener un sistema que administre mejor y facilite el acceso a la información delicada de la empresa, con el fin de abaratar costos y cuidar la integridad de los datos.

Como respuesta a esta necesidad el cliente tomó la decisión estratégica de emplear tecnología en la nube para el almacenamiento de datos. Esta elección se fundamenta en la búsqueda de una solución eficiente y escalable que responda a los desafíos actuales del sistema heredado. Para abordar el proceso de carga de datos masivos, se identificó la necesidad de contar con un software que cumpla con el proceso, llevando a la elección de Talend como la herramienta idónea para esta función. En cuanto a las visualizaciones, el cliente expresó su preferencia por mantener Power BI como su software de confianza para obtener información gráfica.

4.2.2. Indicadores de gestión

Principalmente la empresa en cuestión se dedica a procesar transacciones de diferentes bancos a nivel nacional, así como a administrar cajeros automáticos en nuestro país, por lo tanto

manejan información relativa a estos puntos, y por ello se acordó la existencia de dos grandes KPI para direccionar mejor la etapa de visualización para ambas partes:

- Integridad de las transacciones a nivel nacional: El enfoque de este indicador es medir la fiabilidad que poseen las transacciones bancarias en nuestro país, dando información contable de diferentes ámbitos enfocados en monitorizar la garantía de que las transacciones se llevan a cabo de manera segura, íntegra y sin alteraciones.
- Incidentes de cajeros automáticos (ATM) a nivel nacional: Este indicador busca recopilar información relativa a los incidentes que ocurren en los ATM en nuestro país, especialmente enfocado en el tiempo de caída y la causa del incidente, con el fin de tomar medidas rápidamente frente a cualquier inconveniente en las máquinas.

4.3. Diseño de arquitectura técnica

La solución propuesta para la problemática consta de un datamart que evolucione el almacenamiento de datos de la empresa, a uno que posea las características necesarias para cumplir con los requerimientos y muchas más facilidades con respecto al manejo de datos, ya que, al ser una plataforma en la nube, abarata problemas de data warehouse locales y utiliza los que Snowflake provee, dándole así mucha más vida útil al sistema de almacenamiento, esto gracias a la característica de elasticidad de Snowflake, específicamente a su capacidad para escalar de manera flexible y eficiente, tanto en términos de capacidad de almacenamiento como de potencia de procesamiento, con el objetivo de satisfacer las necesidades cambiantes de la organización.

Se observa en la figura 4.1, el diagrama de proceso para lograr la arquitectura técnica que posee el datamart propuesto como solución a la problemática.

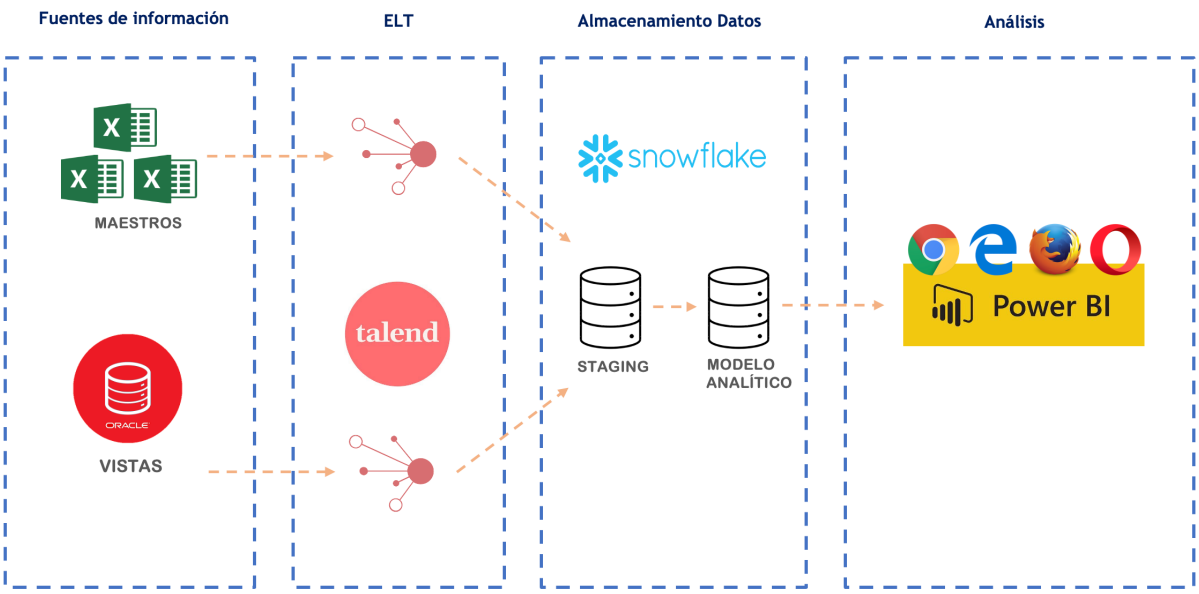


Figura 4.1: Diagrama de arquitectura técnica.

En la figura se observan las cuatro fases en las que la solución se desglosa. Recibe una cantidad masiva de datos, provenientes de vistas y archivos csv, los cuales pasan hacia la fase ELT, donde mediante la aplicación Talend la información es transmitida de manera cruda hacia Snowflake, en el cual se realiza la transformación desde el esquema de llegada "Staging" hacia el esquema final "Modelo Analítico", que contendrá el nuevo modelo de datos. Finalmente, se crearán visualizaciones de datos en Power BI, accesibles desde cualquier navegador.

4.4. Modelado dimensional

Para el modelado, se realizó dos tipos de diagramas, en los cuales se puede visualizar como quedará compuesto el datamart en cuanto a tablas de información respecta. El diagrama lógico nos muestra la estructura del sistema de manera general, mientras que el diagrama entidad-relación describe a más detalle cómo se distribuyen los datos dentro de cada tabla del modelo. Ambos diagramas se ilustran a continuación seguido de la explicación del modelado dimensional.

4.4.1. Diagrama Lógico

El diagrama lógico del sistema se aprecia en la figura 4.2.

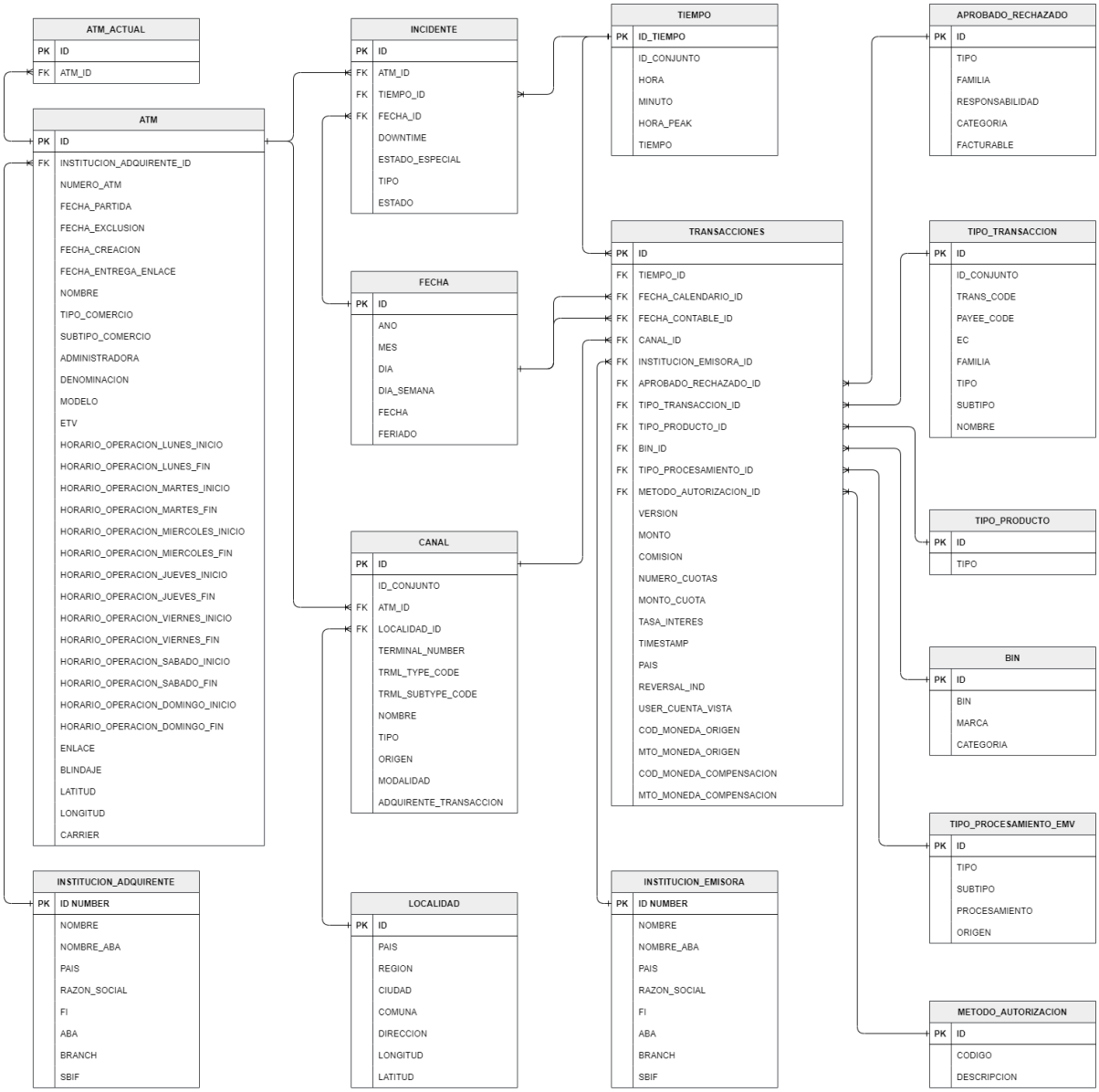


Figura 4.2: Diagrama lógico.

4.4.2. Diagrama entidad-relación

El diagrama entidad relación propuesto para el sistema se aprecia en la figura 4.3.

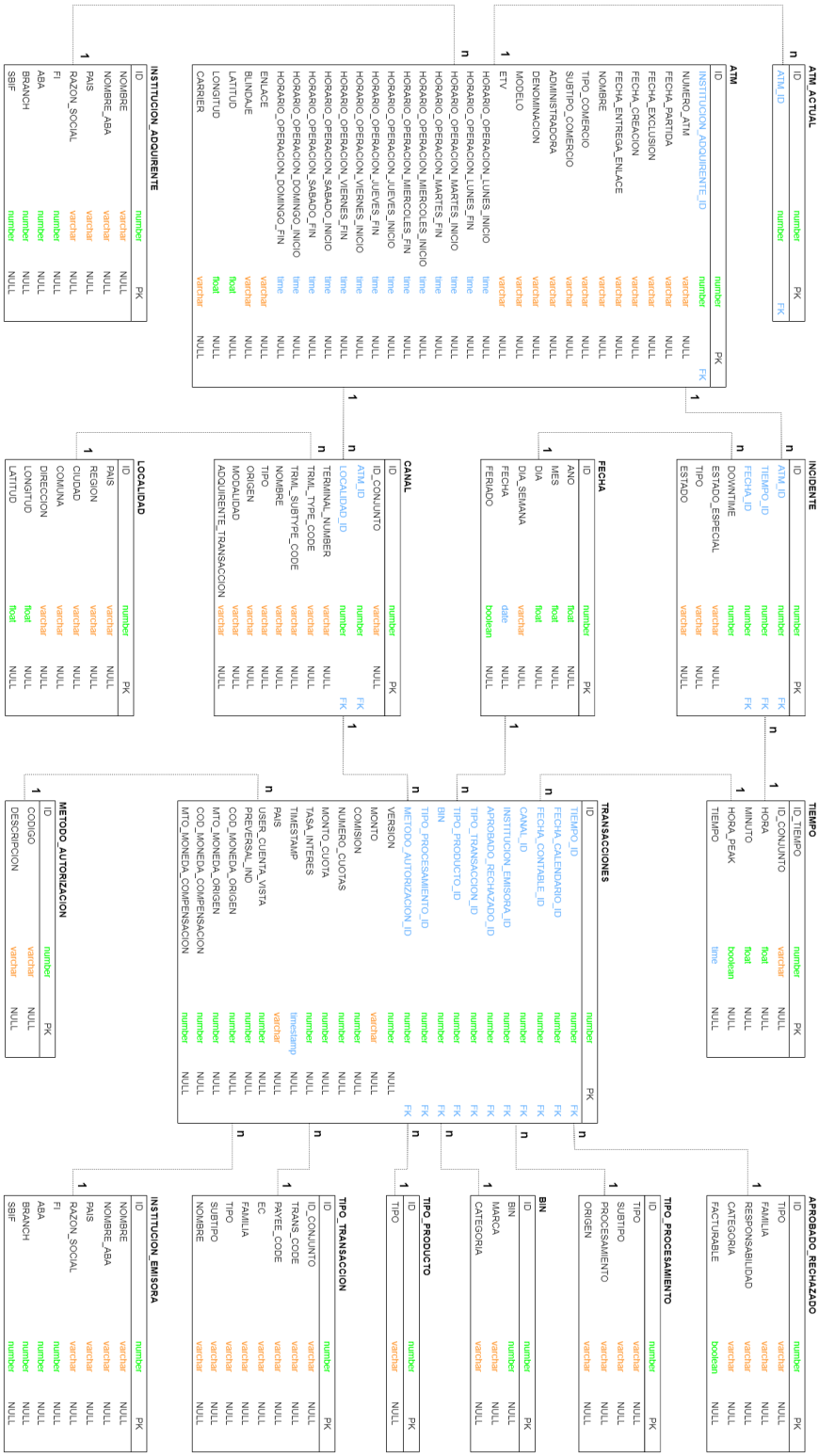


Figura 4.3: Diagrama entidad relación.

El enfoque dimensional se basa en un esquema tipo copo de nieve, debido a que las dimensiones de este se dividen en subtablas. En esta estructura existe una tabla de hechos principal llamada “transacciones”, dentro de la cual se alojan datos de tipo ID como claves foráneas provenientes de las diferentes dimensiones del diseño y una clave primaria tipo ID, esto con el fin de normalizar y ordenar los datos para su acceso y análisis. Esta tabla de hechos contiene información importante en modo resumen de las dimensiones que la componen. En cuanto a las dimensiones, existen un total de 11 con 4 subdimensiones, las cuales contienen atributos con información confidencial. En el lado izquierdo del modelo se alojan todas las dimensiones relativas a los ATM, como la institución adquiriente, la información contable y todos los detalles acerca de los posibles incidentes. Por otro lado, el lado derecho del modelo contiene a todas las dimensiones que se relacionan directamente con la información acerca de las transacciones, el tipo de procesamiento, el tiempo que toma cada transacción, la institución emisora, etc., todo esto con el fin de estructurar de manera clara y precisa los datos contenidos.

4.5. Especificación de aplicaciones BI

Las aplicaciones BI a utilizar en el proyecto son tres, Talend (Studio, Remote Engine y Cloud), Snowflake, y Power BI, las cuales se resumen en la figura 4.4.

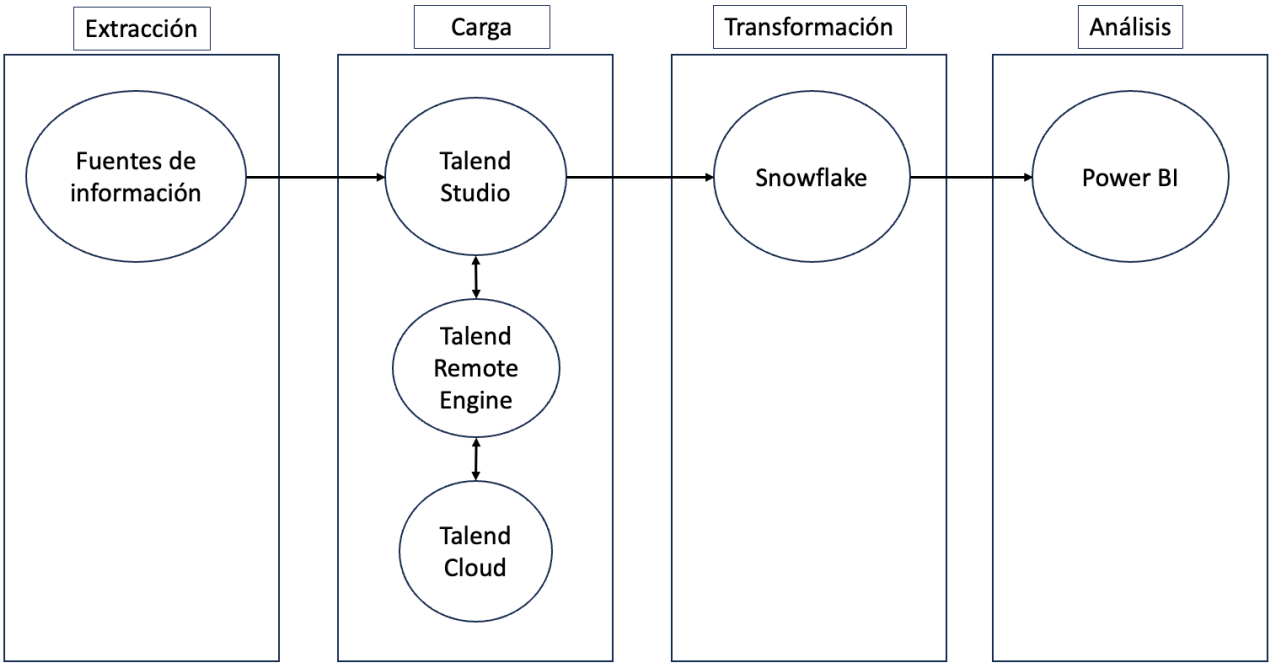


Figura 4.4: Aplicaciones BI a utilizar (elaboración propia).

La figura separa el proyecto en cuatro fases, esto para ayudar a clarificar la función específica que cumple cada aplicación dentro de la solución, desglosada de la siguiente manera.

- **Talend Studio**: En esta herramienta se crearán los Jobs, que servirán para cargar datos hacia Snowflake.
- **Talend Remote Engine**: Aplicación conjunta de Talend Studio, que actuará como un conector de este con Talend Cloud.
- **Talend Cloud**: Aplicación alojada en la nube que permitirá automatizar el proceso ELT ejecutando los Jobs mediante triggers programados en horarios establecidos.
- **Snowflake**: DWH Cloud que permitirá alojar los datos crudos para posteriormente realizar una transformación dentro del mismo software y mantenerlos almacenados.
- **Power BI**: Representación visual del datamart, alojará datos visuales del modelo analítico final en formato dashboard basados a los KPI pedidos por la empresa para su posterior análisis.

4.6. Implementación

En esta etapa se detalla cómo se realiza la etapa de implementación, correspondiente al flujo completo ELT, divididos en cuatro secciones de trabajo.

4.6.1. Extracción

La extracción de datos del modelo propuesto consta de dos fuentes de datos, los cuales en conjunto forman las fuentes de información del datamart. La figura 4.5 representa visualmente el proceso de extracción de datos.

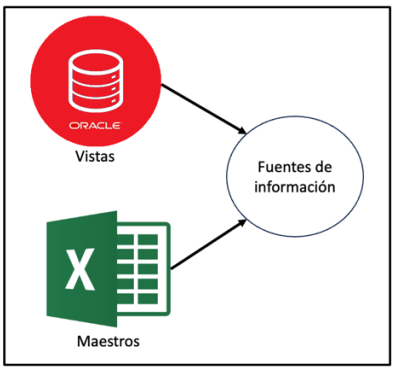


Figura 4.5: Proceso de extracción.

Vistas

Son vistas generadas desde la base de datos Oracle actual de la empresa cliente. Estas vistas son obtenidas mediante un componente específico de Talend Studio.

Maestros

Los maestros son una serie de archivos Excel que contienen información contable de la

empresa cliente.

4.6.2. Carga

La figura 4.6 representa visualmente la carga automatizada utilizando las tres herramientas de Talend.

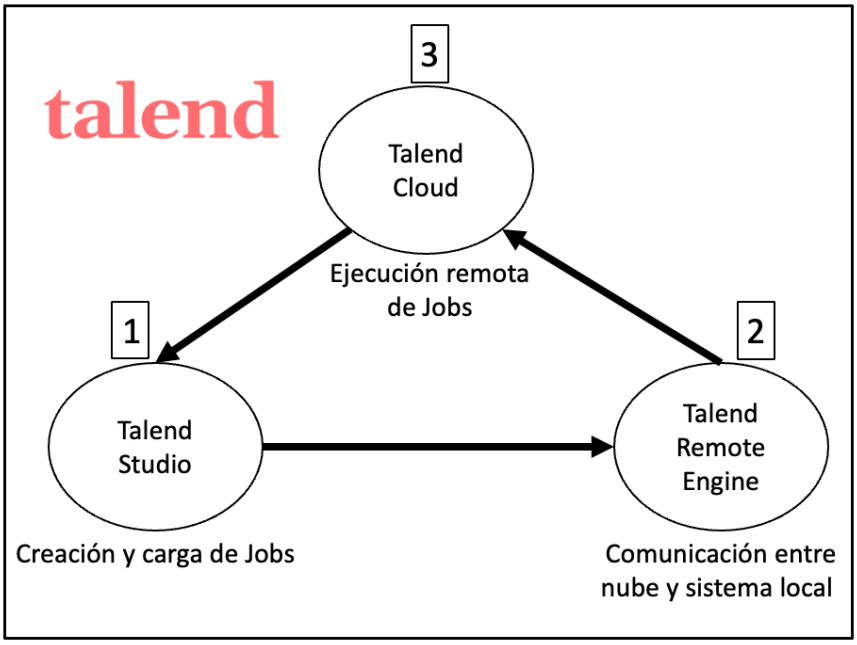


Figura 4.6: Proceso de carga de datos.

- Primera fase, Talend Studio: Una vez obtenidas las fuentes de información, se procede a crear un total de 14 Jobs en Talend Studio. Estos Jobs toman las fuentes de extracción de datos y los cargan directamente hacia Snowflake. La función principal de estos Jobs es tomar todos los datos presentes en las bases de datos de Oracle de la empresa y llevarlas hacia Snowflake, para crear el modelo analítico final, acompañado de queries que contienen los Stored Procedures. Para conocer el funcionamiento de estos 14 Jobs a detalle, revisar el Anexo A de este documento.
- Segunda fase, Talend Remote Engine: La siguiente etapa es conectar el entorno local desarrollado en Talend Studio con la nube, esto se realiza instalando y configurando la aplicación Talend Remote Engine en la misma máquina donde se aloja Talend Studio, para posteriormente interconectarlos gracias a una llave de conexión que brinda Talend Cloud.
- Tercera fase, Talend Cloud: Finalmente, una vez se tenga el entorno local y nube conectados, los Jobs deben subirse a la nube y una vez ahí, programar la carga de datos

en un periodo de tiempo acordado, en este caso de lunes a domingo a las 00:30, esto con el fin de que la información esté actualizándose constantemente.

4.6.3. Transformación

La figura 4.7 representa visualmente el proceso de transformación de datos.

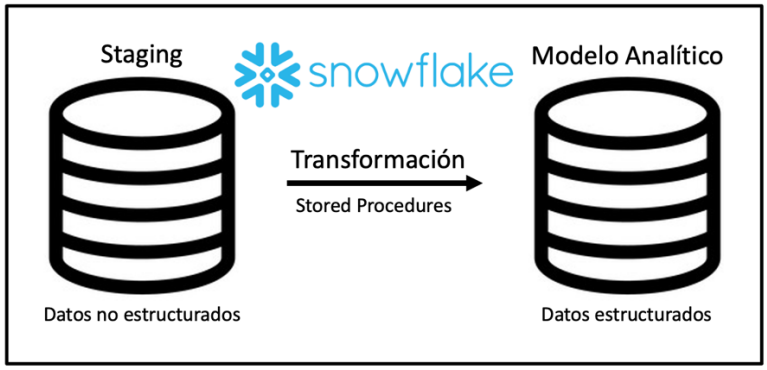


Figura 4.7: Proceso de Transformación.

Los datos llegan a la plataforma cloud de almacenamiento Snowflake, sin estructura y directamente al esquema de nombre “Staging”, desde donde son transformados al nuevo modelo relacional visible en el punto 4.4 de este documento, el cual quedará almacenado en el esquema final llamado “Modelo Analítico”.

La transformación de datos es realizada mediante quince procedimientos almacenados, que son ejecutados automáticamente desde Talend Cloud, gracias a los Jobs mencionados en el punto 4.6.2. La figura 4.8 muestra un resumen del funcionamiento de los Stored Procedures (SP).

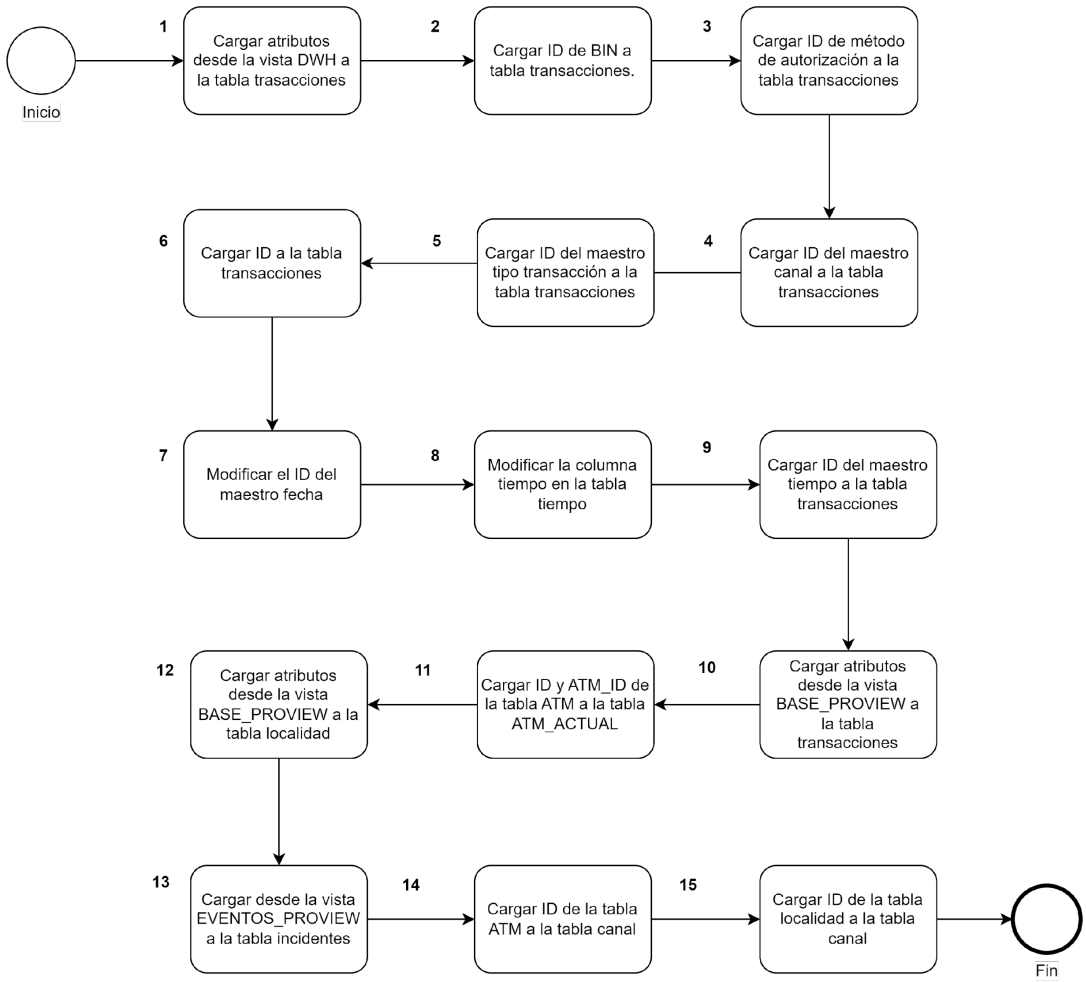


Figura 4.8: Funcionamiento Stored Procedures.

Para revisar a detalle el funcionamiento de cada Stored Procedure, se recomienda revisar el Anexo A del presente documento.

4.6.4. Análisis

La figura 4.9 representa visualmente el proceso de análisis de datos.

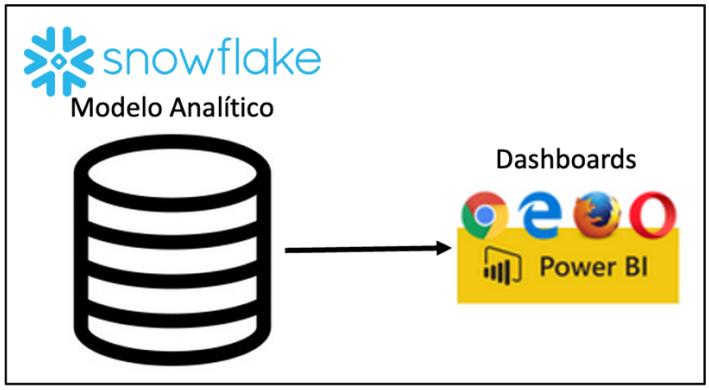


Figura 4.9: Proceso de análisis

En esta etapa se crean los dashboards necesarios para el análisis, los cuales deben cumplir con los KPI estipulados en el punto 4.1, así como con cubrir las áreas de interés de la empresa, en otras palabras, deben reflejar los datos que la empresa requiera para su posterior análisis. Los dashboards creados apuntan hacia el esquema "Modelo Analítico" alojado en Snowflake para obtener los datos y generar las visualizaciones.

4.7. Mantenimiento y crecimiento

Para el desarrollo de esta etapa, se utilizaron documentos en formato pdf con el fin de crear manuales instructivos de cómo implementar el datamart propuesto, además de cómo afrontar errores poco comunes, los cuales fueron desarrollados a lo largo del avance del proyecto. Luego, mediante una capacitación del sistema por parte del desarrollador hacia la empresa cliente, se asegurará que el sistema sea conocido y absorbido de buena manera. Estos documentos mencionados son confidenciales y no se puede dar evidencia explícita de su contenido.

4.8. Conclusiones del capítulo

Gracias a la metodología Kimball, fue posible dividir el proyecto en seis fases, lo que permitió tener una estructura organizada y detallada para llevar a cabo cada uno de los pasos y procedimientos. Además, la ejecución secuencial de estos procesos ha brindado un alto nivel de control sobre las diferentes etapas.

Capítulo 5

5. Antecedentes

En este capítulo, se presentan los antecedentes técnicos y normativos, económicos y financieros y sociales y ambientales del proyecto.

5.1. Introducción

En la presente sección se detallan los diferentes tipos de antecedentes que el proyecto conlleva, explicando en profundidad en cada uno de ellos. Se exploran tanto los aspectos técnicos y normativos del proyecto, haciendo hincapié en la importancia de la integridad ética profesional. Asimismo, se detallan los aspectos económicos y financieros, incluyendo la definición del presupuesto del proyecto, desglosado en personal y licencias. Por último, se analizan los factores sociales y ambientales que tienen relevancia en el proyecto.

5.2. Antecedentes técnicos y normativos

5.2.1. Antecedentes técnicos

La mayoría de los elementos técnicos del proyecto se basan en conocimientos y tecnologías de gran importancia en la actualidad del mundo de los datos. La comprensión del problema del acceso a la información y su sistema heredado, el estado actual de la literatura y las técnicas de modelado y migración de datos entregan una extensa visión del panorama, así como los conceptos técnicos necesarios para comprender mejor el contexto investigativo. Las tecnologías y herramientas utilizadas durante la implementación del proyecto fueron presentadas en el capítulo 4, dentro del cual se proporcionaron descripciones detalladas de cada uno de estos elementos técnicos que se usaron en distintas etapas del proyecto. Estos elementos abarcaron aspectos como almacenamiento masivo de datos, flujos de trabajo, y especialmente las herramientas de tipo software que se utilizaron para resolver la problemática.

5.2.2. Antecedentes normativos

Si bien no existe una ley que regule el análisis y migración de datos como tal en nuestro país, si existe una ley de protección de datos de los ciudadanos chilenos. La ley de protección de datos N°19.628. Dicha ley promulgada en 1999 establece normas para la protección de los datos personales de distinta índole, pero anticuada en muchos aspectos [24]. Es por ello por lo que, en el año 2018, con la modificación del artículo 19 Nº4 de la constitución, se estableció la protección de datos personales como derecho fundamental. Según estas leyes, se considera datos de carácter o datos personales, los relativos a cualquier información concerniente a personas naturales, identificadas o identificables. Desde el R.U.T., nombre y apellidos hasta dirección de correo electrónico, teléfono, lugar de residencia, edad, etc.

Las empresas deben tomar en cuenta la Ley de protección de datos en Chile y garantizar el cumplimiento de sus obligaciones. Lo anterior afecta al proyecto, ya que al estar presente en el rubro bancario y realizar un análisis de datos exhaustivo y periódico, los datos bancarios de las personas deben estar resguardados tras un firme sistema de confidencialidad, es por ello también que, mediante acuerdo entre empresa y desarrollador, los datos de prueba del proyecto constan de información antigua y reducida, además de no contener detalles sensibles de ciudadanos chilenos. Una vez que la solución entre en funcionamiento, la propia empresa maneja sus propias políticas de confidencialidad de datos.

5.3. Antecedentes económicos y financieros

En cuanto a los gastos económicos del proyecto, el desarrollo del proyecto cuenta con un profesional del área de ingeniería civil informática, el cual se dedicará de manera full-time al despliegue de la solución por un periodo de cuatro meses estimados en la duración del proyecto, dicha duración puede estar sujeta a retrasos. Los datos económicos de personal se resumen en la tabla 5.1.

Profesional	Valor Hora	Tiempo	Valor
Ingeniero Civil Informático[1]	$22.272	4 meses	$3.920.000
		Subtotal	$3.920.000

Tabla 5.1: Costo de personal.

En cuanto al costo de licencias y tecnologías, la solución contempla el uso de Talend, Snowflake y Power BI como servicios de plataforma. Talend cuenta con una compra de licencia mensual fija, pero en el caso de Snowflake el costo se debe calcular una vez los datos masivos entren en funcionamiento, con un valor basado en un sistema de créditos, con el cual se paga por el uso de la plataforma, es por ello por lo que este costo es variable. Además, se plantea la adquisición de la suscripción pro de Power BI para poder compartir los reportes con otros usuarios de la empresa, y/o externos en caso de ser necesario. Cabe señalar que las licencias a utilizar manejan precios en dólares, por lo que estos costos pueden variar a lo largo del tiempo. La tabla 5.2 resume los gastos de licencias del proyecto.

[1] Basado en sueldo promedio ing. Informático en chile, agosto 2023. https://cl.indeed.com/career/ingeniero-informático/salaries

Licencia	Costo Mensual	Tiempo	Valor
Talend[2]	$1.053.000	4 meses	$4.212.000
Snowflake Enterprise[3]	Gratuito	4 meses	$2.700 por crédito
Power Bi Pro[4]	$9.000	4 meses	$36.000
		Subtotal	$4.248.000 + variables

Tabla 5.2: Costo de licencias.

5.4. Antecedentes Sociales y Ambientales

El ambiente social moderno está marcado por la utilización de tarjetas bancarias para realizar todo tipo de trámites, y es que cada vez el dinero en efectivo va desapareciendo paulatinamente, esto trae consigo varios desafíos para mantener estos sistemas en funcionamiento y buscar reducir los errores de traspaso de montos e información delicada entre personas.

La literatura relacionada al desarrollo de sistemas de manejo de datos en nuestro país es muy poco extensa, mucho más si esta incluye nuevas tecnologías, ya que por lo general al menos nacionalmente las empresas suelen ser muy esquivas al momento de actualizarse, por el miedo a perder datos de suma importancia, y a perder la velocidad de su servicio y prestigio.

Por otra parte, este mismo contexto de rapidez social trae consigo problemas de salud mental en personas dedicadas al análisis de estos datos, ya que el manejar una vida laboral muy intensa puede provocar estrés, agotamiento mental, problemas de sueño, entre otras patologías más severas.

5.4.1. Impacto

El proyecto trae consigo un beneficio social bastante importante, debido a que al ser destinado a una empresa del rubro bancario y con múltiples relaciones bancarias, el poder llevar a cabo un proyecto de esta magnitud beneficiará a una gran cantidad de ciudadanos del país, ya que en su mayoría las personas mayores de 18 años están afiliadas a una banca comercial. Se

[2] Basado en costo mensual en dólares, valor peso según fecha agosto 2023.
https://www.saasworthy.com/product/talend-cloud-data-integration/pricing
[3] Basado en costo por crédito en dólares, valor peso según fecha agosto 2023.
https://www.snowflake.com/en/data-cloud/pricing-options/
[4] Basado en costo mensual en dólares, valor peso según fecha agosto 2023.
https://powerbi.microsoft.com/en-us/pricing/

espera que los incidentes comerciales de todas las instituciones adquirientes puedan disminuir paulatinamente gracias al acceso rápido a los datos que el sistema brindará, aumentando la rapidez de cobertura. Asimismo, este proyecto realizará un aporte al área de la inteligencia de negocios, beneficiando a una buena cantidad de profesionales, como incentivo para extender la aplicación de tecnologías en la nube en sistemas que manejen datos masivos y aumentar la cantidad literaria. Finalmente, se espera que el tener un sistema mucho más avanzado tecnológicamente reduzca la carga laboral de los profesionales dedicados a analizar datos en él, mejorando su estado de salud.

5.5. Conclusiones

Los antecedentes en un proyecto son fundamentales para establecer una base sólida, proporcionar contexto, justificar la necesidad del proyecto y orientar su planificación y ejecución de manera efectiva. Ayudan a asegurarse de que el proyecto esté bien fundamentado y alineado con los objetivos y necesidades de las partes interesadas.

En lo que respecta a los antecedentes técnicos y normativos, fue posible identificar las tecnologías, conocimientos y herramientas esenciales para el desarrollo del proyecto, así como las regulaciones y leyes aplicables en nuestro país para este tipo de trabajos. En cuanto a los antecedentes económicos, a pesar de la existencia de costos variables, proporcionaron una visión sólida de los costos asociados con un proyecto de estas características. En lo que concierne a los antecedentes sociales y ambientales, el proyecto promete tener un impacto positivo que se espera pueda materializarse.

Capítulo 6

6. Planificación

En este capítulo, se presenta la planificación del trabajo que conlleva la realización del proyecto.

6.1 Plan de trabajo

En este capítulo se establece el plan de trabajo del proyecto, utilizando para ello un diagrama de Gantt, en el cual se dividirán por semanas los plazos de trabajo. En este caso, se toma los plazos de la carta Gantt en base a lo que el tesista participó, desde el inicio de abril, hasta el fin del proyecto que hasta ahora se estima como finales de Julio.

6.2 Carta Gantt

N°	Tiempo (semanas) / Actividad	Abril				Mayo				Junio				Julio			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1.	**Requerimientos**																
1.1.	Toma de requisitos del proyecto																
1.2.	Identificación de necesidades																
2.	**Arquitectura técnica**																
2.1.	Selección de tecnologías																
2.2.	Diseño de diagramas																
3.	**Modelado Dimensional**																
3.1.	Creación de tablas y dimensiones																
3.2.	Diseño de diagrama entidad-relación																
3.3.	Diseño de diagrama lógico																
4.	**Especificación de Aplicaciones BI**																
4.1.	Selección de aplicaciones BI a utilizar																
5.	**Implementación**																
5.1.	Extracción																
5.1.1.	*Creación de vistas*																
5.1.2.	*Recopilación de maestros*																
5.2.	Carga																
5.2.1.	*Creación de Jobs en Talend Studio*																

5.2.2.	*Subida de Jobs a Talend Cloud*																
5.2.3.	*Carga de datos hacia Snowflake*																
5.3.	Transformación																
5.3.1.	*Ejecución de SP en Snowflake*																
5.3.2.	*Creación de modelo analítico en Snowflake*																
5.4.	Análisis																
5.4.1.	*Creación de Dashboards*																
6.	**Mantenimiento y crecimiento**																
6.1.	Documentación del proceso completo																
6.2.	Capacitación a usuarios																

Tabla 6.1: Carta Gantt del proyecto.

Capítulo 7

7. Resultados

En esta sección las discusiones, conclusiones y trabajos futuros relacionados al trabajo realizado, son presentadas.

7.1 Resultados Cualitativos

Hoy en día la solución cumple con lo estimado, está presente dentro de la plataforma de Snowflake un datamart funcional, el cual se conecta exitosamente con los reportes de datos. Dicho datamart cuenta con 16 tablas, las cuales se relacionan de la misma manera en que la que fueron contempladas en el capítulo 4. En la figura 7.1. se observan las tablas en Snowflake, correspondiente al modelo analítico propuesto cargado en el Data Warehouse.

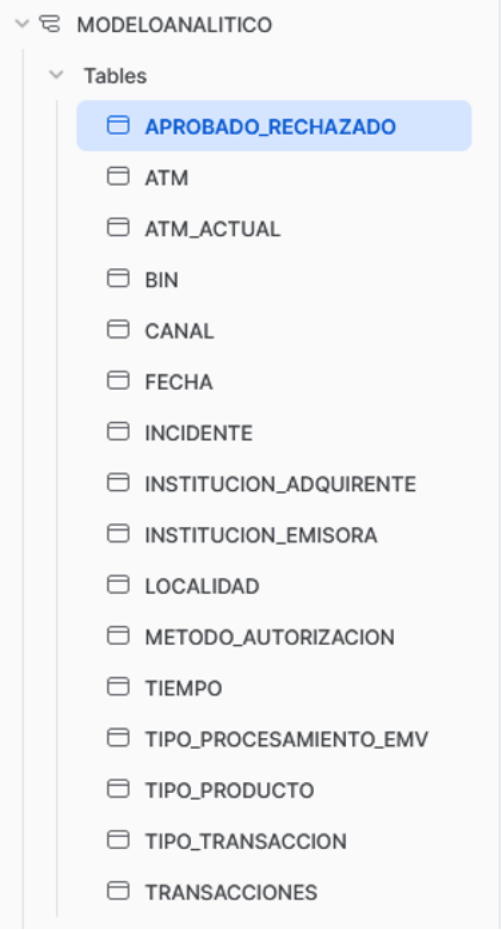


Figura 7.1: Modelo Analítico en Snowflake

Además, existe una amplia variedad de documentos que dan evidencia de lo realizado, de igual manera un plan de implementación para capacitar a los trabajadores de la empresa y obtener soluciones en caso de algún pormenor.

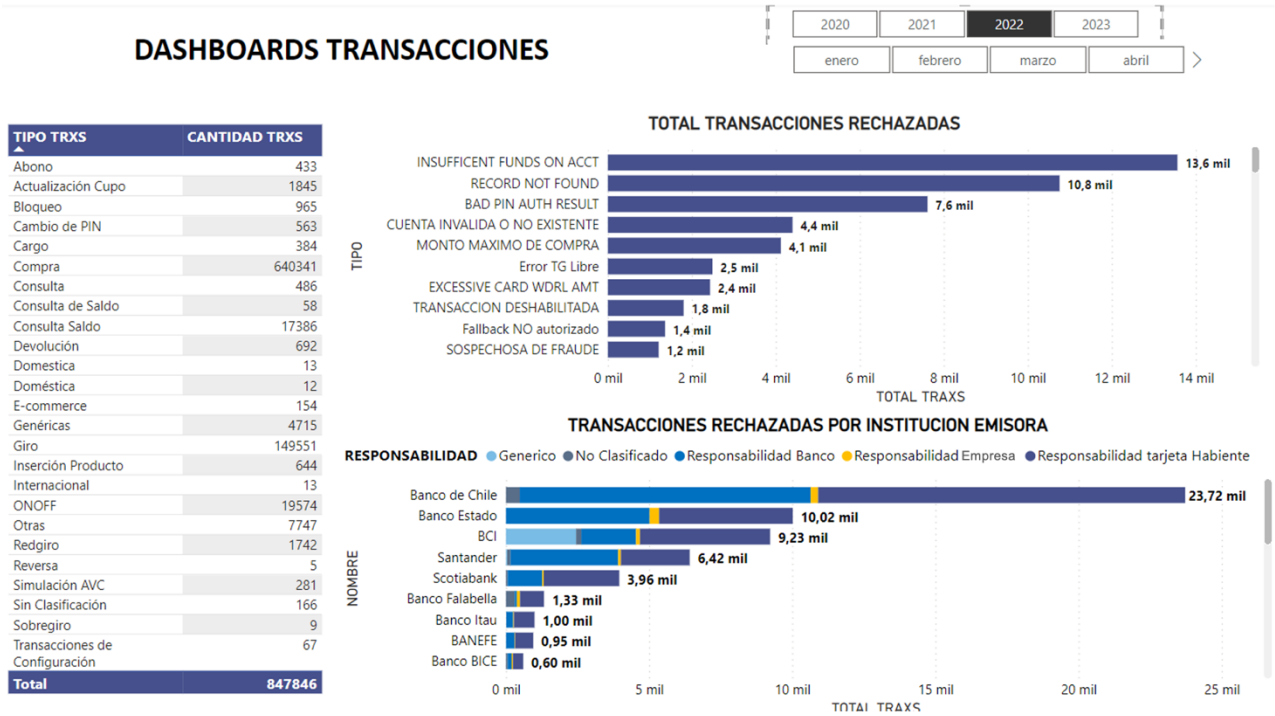
7.2 Resultados Cuantitativos

Los resultados cuantitativos del proyecto constan de dashboards dinámicos con información contable, los cuales se dividen en base a los KPI mencionados en el punto 4.2 y abarcan una

cantidad de datos reducida, ya que son datos de prueba, y la información real es confidencial para la empresa. La fuente de información de estos dashboards y se irá actualizando gracias a la implementación automatizada explicada en el punto 4 del documento. A continuación, se explicarán los dashboards resultantes de la implementación de la solución, detallando cómo pueden ser un aporte para resolver cuestiones relacionadas a los KPI.

7.2.1 Integridad de las transacciones a nivel nacional

Los dashboards con la leyenda "Transacciones" son una serie de visualizaciones realizadas en base al primer KPI, por lo que engloban información contable de transacciones realizadas a nivel nacional utilizando diferentes herramientas visuales, lo que ayudará a los analistas a resolver problemáticas de la empresa en base a los KPI acordados. A continuación, se da evidencia de ellos.



TIPO TRXS	CANTIDAD TRXS
Abono	433
Actualización Cupo	1845
Bloqueo	965
Cambio de PIN	563
Cargo	384
Compra	640341
Consulta	486
Consulta de Saldo	58
Consulta Saldo	17386
Devolución	692
Domestica	13
Doméstica	12
E-commerce	154
Genéricas	4715
Giro	149551
Inserción Producto	644
Internacional	13
ONOFF	19574
Otras	7747
Redgiro	1742
Reversa	5
Simulación AVC	281
Sin Clasificación	166
Sobregiro	9
Transacciones de Configuración	67
Total	**847846**

Figura 7.2: Dashboard transacciones.

Este dashboard principal presente en la figura 7.2 cuenta con información contable de todas las transacciones realizadas a nivel nacional, con filtro de fechas para obtener un rango determinado de tiempo para su análisis. Además, posee dos tipos de herramientas de visualización, tablas y gráficos, los que se separan en transacciones totales y rechazadas por cada institución, y el nombre del tipo de rechazo que obtuvo. Esto ayuda a la integridad de las transacciones, ya que podemos ver, por ejemplo, que bancos están recibiendo más errores en las

transacciones, y cuál es la descripción del error arrojado por el sistema, lo que da unos datos claros y precisos para solucionar los inconvenientes y mejorar la consistencia del servicio.

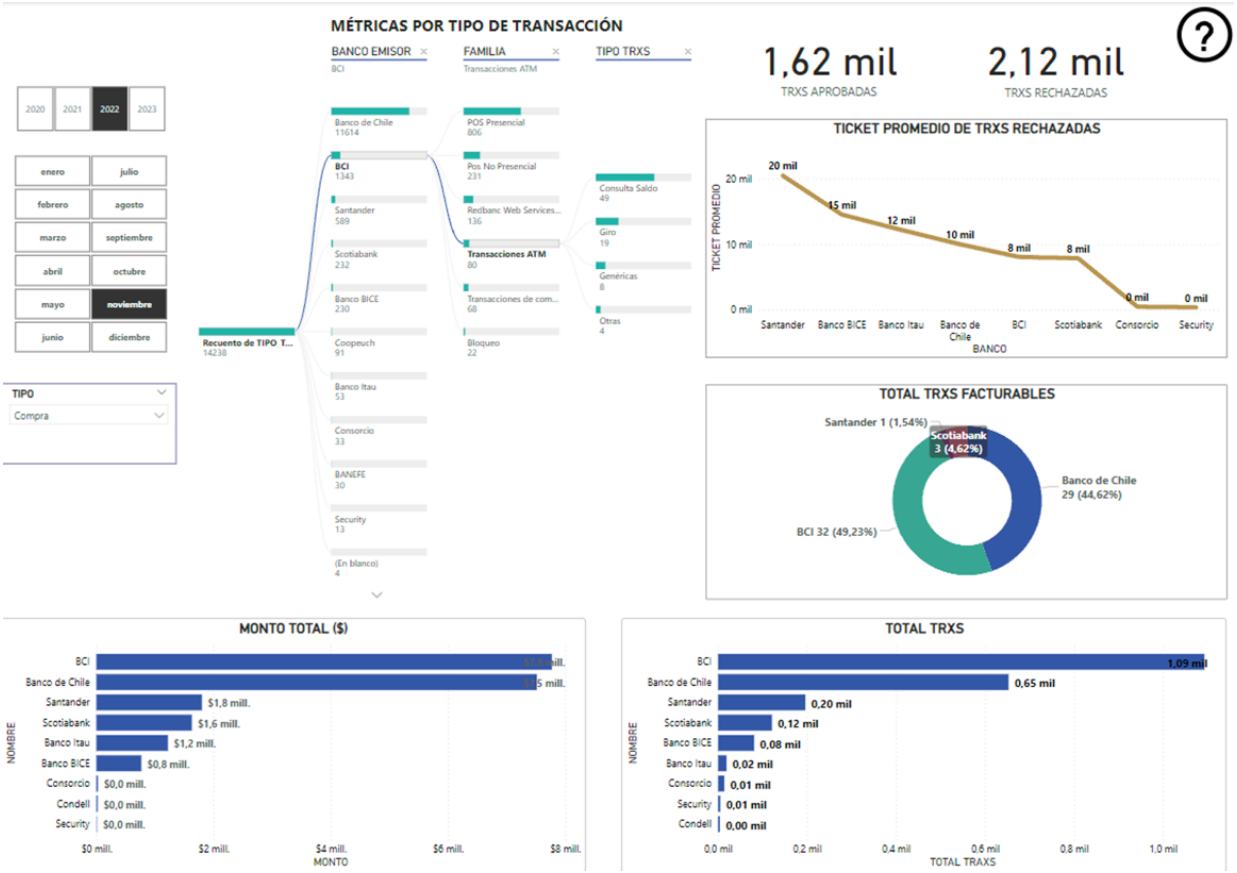


Figura 7.3: Dashboard transacciones por tipo.

El siguiente dashboard presente en la figura 7.3 posee un filtro de tipo de transacción, desde donde podemos ver un tipo específico de operación y conocer la información contable de esta clase seleccionada. Esto ayuda a la integridad de las transacciones de distintas maneras, por ejemplo, distinguir las transacciones de tipo “compra” con las de tipo “giro” y revisar el ticket promedio de rechazo de esta clase de operaciones, el total facturable, el monto total obtenido de estas, etc. Lo que ayudará a los analistas a identificar diferentes indicadores que afecten a la integridad de la información.

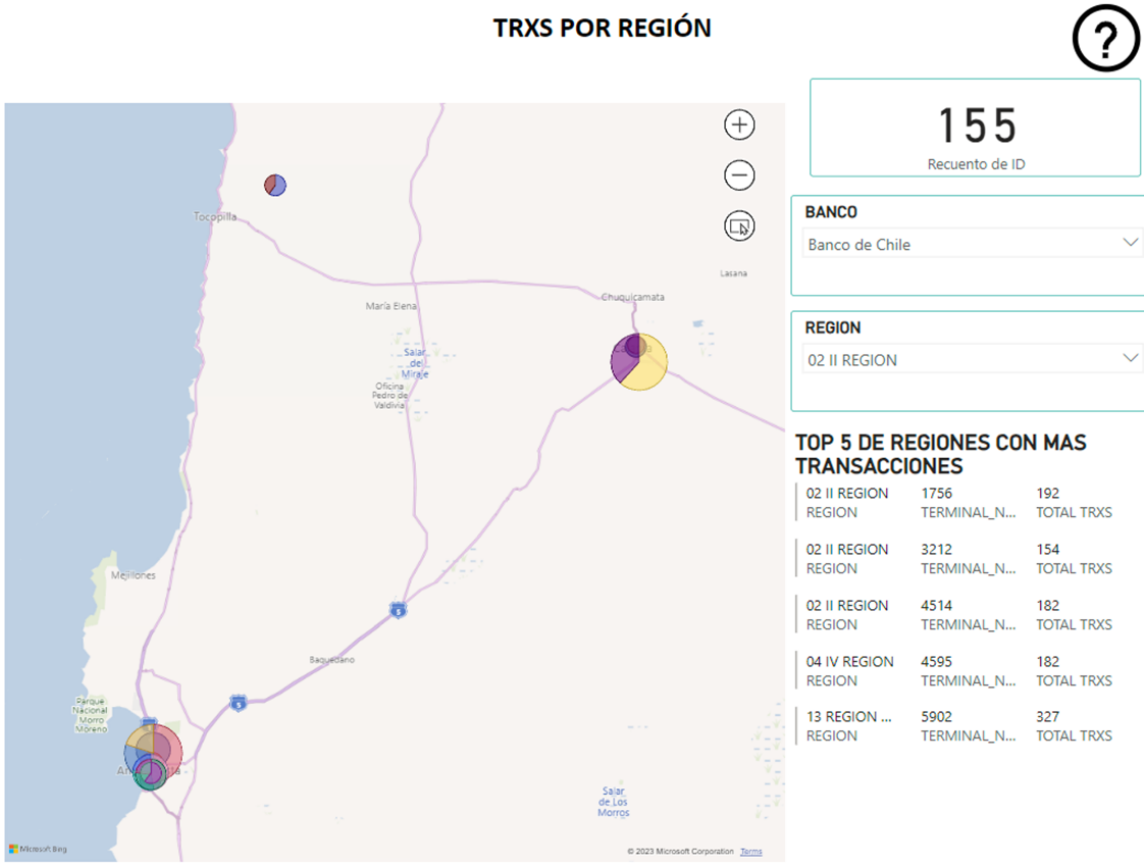


Figura 7.4: Dashboard transacciones por región.

El tercer dashboard presente en la figura 7.4 dispone de información geográfica de las transacciones en base a cada región de nuestro país, con un mapa topográfico y un top de regiones con más transacciones, lo que ayuda a reconocer geoespacialmente la información comercial. Gracias a esta herramienta podemos mejorar la integridad, por ejemplo, dirigiendo las miras de la compañía hacia las regiones que más transacciones procesan, lo que aumenta la probabilidad de problemas en estas.

7.2.2 Incidentes de cajeros automáticos (ATM) a nivel nacional

Los dashboards con la leyenda “incidentes” son visualizaciones enfocados en cubrir el segundo KPI, poseen información acerca de los incidentes en los cajeros automáticos a nivel nacional. A continuación, se ilustran, con su correspondiente descripción.

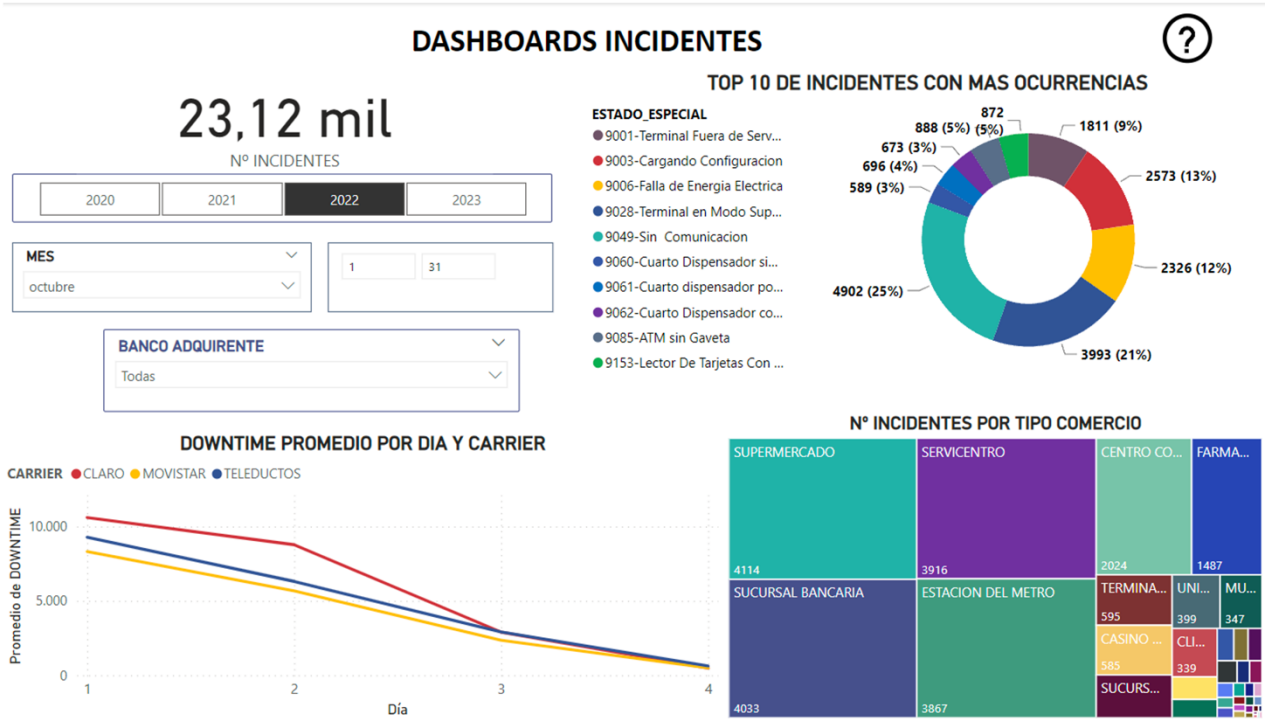


Figura 7.5: Resumen Incidentes.

Este dashboard principal presente en la figura 7.5 contiene información general acerca de los incidentes ocurridos en los distintos ATM a nivel nacional, posee un filtro de fechas indispensable para consultar un tiempo determinado, así como uno de bancos, para revisar individualmente cada entidad. Esta herramienta es muy útil para darnos una visión general y específica de donde y cuando existen fallas en el funcionamiento de los ATM.

Figura 7.6: Incidentes por downtime.

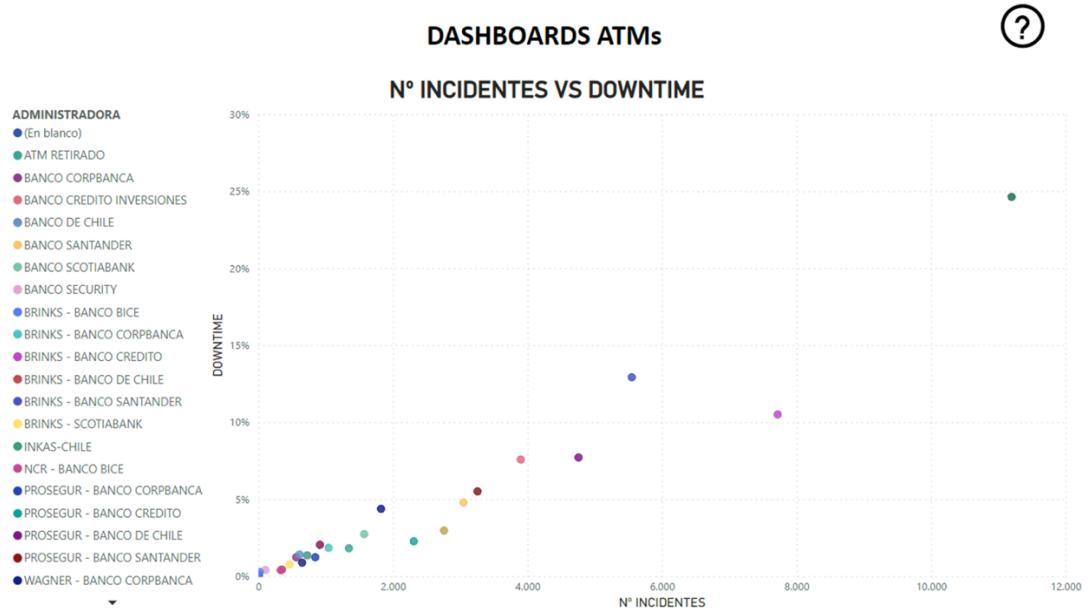


Figura 7.7: Incidentes v/s downtime

Los dashboards anteriores nos sirven para revisar cuanto tiempo estuvieron caídos los ATM, el primero, visible en la figura 7.6, especifica la razón por la que estuvo caído un ATM, mientras que el segundo, visible en la figura 7.7, nos da información acerca de la institución bancaria a la cual pertenece el ATM. Ambos dashboards nos son útil al momento de analizar cuales cajeros necesitan mantención más urgente, y para establecer tiempos de caída promedio de estos, así como varias otras métricas analíticas a lo largo del tiempo.

Capítulo 8

8. Discusión y Conclusiones

En este capítulo, se presenta el Glosario, en el que se indica el significado de los términos técnicos utilizados en el documento.

8.1. Discusión

Los resultados visibles en el capítulo 7 dan a conocer la implementación del modelo propuesto como solución intermedia a la problemática, esto se ve reflejado en los resultados cuantitativos, donde mediante la exposición de dashboards, se da cuenta del correcto acceso a los datos que el sistema posee. Si bien hoy en día se encuentra en una etapa en la que se utilizan datos no masivos, se espera que el datamart pueda alojar correctamente el big data posterior. La primera gran característica que separa a esta solución de otras comunes es la presencia del cloud, ya que gracias a las características de un software como Snowflake, no existirá problemas de acceso y modelado de datos una vez estén cargados en el sistema, sin alterar su inmutabilidad, ahorrando mucho tiempo en el análisis de la información contable. La segunda gran característica es el uso de ELT como flujo de trabajo, esto es un salto de calidad en cuanto al proceso de integración de datos, ya que como se evidenció en el capítulo 2, ELT posee una gran variedad de ventajas sobre ETL, siendo la más importante la velocidad. El proceso de ELT carga los datos directamente en el sistema de destino y los transforma en paralelo. Utiliza la potencia de procesamiento y la paralelización que ofrecen los almacenamientos de datos en la nube para ofrecer una transformación de datos en tiempo real o casi real para llevar a cabo análisis. Esto es fundamental al trabajar con datos masivos, debido a que el principal objetivo de este tipo de procesos es optimizar tiempo, que va de la mano con abaratar costos de producción.

En cuanto a la literatura, como se ha dejado evidenciar en el capítulo 2, no existen trabajos disponibles que utilicen las tecnologías que el proyecto posee, es por ello por lo que no se puede realizar una comparación de resultados. Lo anterior es debido a que, al ser herramientas relativamente nuevas, aún deben evolucionar en el mercado para obtener estudios de este tipo de solución propuesta.

8.2. Conclusiones

En el transcurso de este documento, se ha presentado una solución de Business Intelligence la cual cuenta con un novedoso sistema ELT para su funcionamiento, además de ser una mezcla de un par de tecnologías relativamente nuevas y con gran potencial dentro del campo del BI, como lo son Talend y Snowflake. Utilizando la metodología de Ralph Kimball, con algunas modificaciones, se logró dar una estructura para el desarrollo óptimo de cada una de las etapas. El modelo propuesto tuvo como raíces los requisitos de la empresa, gracias a los cuales se pudo

crear una arquitectura. Dentro de la arquitectura técnica, se desarrolló un sistema de datamart que cuenta con las bases principales de cualquier modelo analítico, desde la recolección de las fuentes de datos hasta la visualización de información relevante para la toma de decisiones. Los datos fueron estructurados gracias a un modelo de copo de nieve que organiza la información para mantener su integridad y facilitar la consulta hacia ésta. Además, fue necesario especificar cuales aplicaciones se dedicarían a cada tarea dentro de la metodología, esto para cubrir necesidades tecnológicas de la solución. Finalmente, se realizó la implementación del método propuesto, en el que se sintetiza todo lo anterior en acciones, explicando como el flujo ELT fue realizado y su posterior fase de análisis. Utilizando Talend fue posible migrar los datos desde Oracle, los cuales llegaron de manera cruda a Snowflake, donde se realizó un modelado de datos, para poder estructurarlos en el modelo final a donde apuntan los reportes de visualización. En definitiva, la elección de estas tecnologías y el posterior desarrollo de la solución para dar respuesta a la problemática cumple con el objetivo general del proyecto y se logra implementar una solución Business Intelligence utilizando Talend y Snowflake, cubriendo así las necesidades de la empresa, permitiendo recopilar, almacenar y visualizar datos de manera eficiente, así como una posible evolución a futuro, utilizando estrategias modernas de almacenamiento y modelado de datos, aprovechando las ventajas que el cloud tiene actualmente y su evolución para el futuro.

8.3. Limitaciones y posibles soluciones

Se establecen como limitaciones los siguientes puntos, que se espera sean resueltos por la persona que el liderazgo del proyecto a futuro, para asegurar su correcta evolución.

- Adaptar la plataforma para tratar con big data, obteniendo así un modelo datamart final, con una posterior evaluación de su funcionamiento. Esto debido a que actualmente el proyecto, al ser una solución intermedia, cuenta con datos de prueba, y en un futuro deberán ser ampliados a datos mucho más voluminosos. La solución final saldrá a luz a medida que se sigan realizando pruebas, y corrigiendo errores, en caso de existir, dentro del proceso productivo de la empresa.
- Evidenciar el ahorro de tiempo gracias al datamart propuesto. Una de las grandes premisas del proyecto es cuánto tiempo se ahorrará en un futuro, lo cual se traduce en grandes ahorros de dinero para la empresa. Para poder llevar a cabo esto, es necesario contar con todo el proceso descrito en el punto 4, pero de manera final, con

big data implementado. Una vez se tengan estos datos, crear métricas de tiempo para conocer cuánto tiempo se demora el proceso analítico actualmente dentro de la empresa y compararlo con el proceso analítico anterior, para finalmente crear gráficas que evidencien este tiempo ahorrado.

8.4. Trabajos futuros

Quedan propuestos como trabajos futuros los siguientes puntos, con los cuales se busca que el proyecto sea un aporte futuro para las investigaciones en el área de Business Intelligence.

- Replicar el modelo propuesto para otras áreas y/o empresas que lo requieran, dando un puntapié para que las organizaciones que manejen datos masivos aprovechen las nuevas tecnologías y se remodelen, al menos a nivel nacional.
- Integrar tecnología de inteligencia artificial al proyecto. Hoy en día, aún no ha sido desarrollado un sistema analítico completo que funcione utilizando inteligencia artificial y que sea compatible con las tecnologías utilizadas en este datamart. La expectativa mundial es que en los próximos años aparezcan estos modelos entrenados para realizar análisis y predicciones automatizadas, los cuales se esperan puedan ser adaptables a este tipo de proyectos que mezclan big data y business inteligence.

Bibliografía

[1] R. B. N. A. R. K. P. M. P. P. N Saranya, «IEEEXPLORE,» 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 03 06 2021. [En línea]. Available: https://ieeexplore.ieee.org/abstract/document/9441840/.

[2] J. S. N Prasath, «IEEEXPLORE,» 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 03 06 2021. [En línea]. Available: https://ieeexplore.ieee.org/abstract/document/9441898.

[3] B. Singhal, «IEEEXPLORE,» 2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), 24 02 2023. [En línea]. Available: https://ieeexplore.ieee.org/abstract/document/10046997.

[4] R. C. B. B. J. B. L. Frank Bell, «The Snowflake Data Cloud,» 2022. [En línea]. Available: https://doi.org/10.1007/978-1-4842-7316-6_1.

[5] A. B. S. N. Papa Senghane Diouf, «ieeexplore,» 2018 IEEE International Conference on Innovative Research and Development (ICIRD), 11 06 2018. [En línea]. Available: https://ieeexplore.ieee.org/abstract/document/8376308.

[6] U. A. S. M. Khawaja Ubaid Ur Rehman, «researchgate,» VAWKUM Transactions on Computer Sciences, 03 2018. [En línea]. Available: https://www.researchgate.net/publication/329704023_A_Comparative_Analysis_of_Traditional_and_Cloud_Data_Warehouse.

[7] T. C. M. Z. Benoit Dageville, «ACM Digital Library,» 14 06 2016. [En línea]. Available: https://dl.acm.org/doi/10.1145/2882903.2903741.

[8] A. Singh, «Shipyard,» 11 01 2023. [En línea]. Available: https://www.shipyardapp.com/blog/cloud-data-warehouse/.

[9] Oracle corp., «oracle,» 2023. [En línea]. Available: https://www.oracle.com/cl/database/what-is-a-relational-database/.

[10] GeeksForGeeks, «geeksforgeeks.org,» Patil Education and Welfare Trusts Group of

Institutions College of Engineering and Management PR Pote (PRPEGI) Amravati, 20 05 2023. [En línea]. Available: https://www.geeksforgeeks.org/what-is-stored-procedures-in-sql/.

[11] Databricks, «databricks.com,» 2023. [En línea]. Available: https://www.databricks.com/glossary/snowflake-schema#:~:text=A%20snowflake%20schema%20is%20a,data%20marts%2C%20and%20relational%20databases..

[12] Tableau, «tableau.com,» 2021. [En línea]. Available: https://www.tableau.com/es-es/learn/articles/business-intelligence.

[13] Oracle, «Oracle.com,» [En línea]. Available: https://www.oracle.com/cl/big-data/what-is-big-data/.

[14] PowerData, «powerdata.es,» PowerData, 2023. [En línea]. Available: https://www.powerdata.es/data-warehouse.

[15] IBM, «ibm.com,» IBM, 2023. [En línea]. Available: https://www.ibm.com/topics/data-mart.

[16] International Business Machines Corporation (IBM), «ibm.com,» 2023. [En línea]. Available: https://www.ibm.com/topics/etl.

[17] tutorialspoint, «tutorialspoint.com,» 2023. [En línea]. Available: https://www.tutorialspoint.com/talend/talend_talend_open_studio.htm.

[18] Talend, «talend.com,» 2023. [En línea]. Available: https://help.talend.com/r/en-US/Cloud/remote-engine-user-guide-linux/remote-engine-ug.

[19] Talend, «Talend Cloud,» Talend, 29 06 2023. [En línea]. Available: https://portal.us-west.cloud.talend.com/.

[20] SnapLogic, «Snap Logic,» 14 01 2022. [En línea]. Available: https://www.snaplogic.com/blog/snowflake-data-platform.

[21] A. Das, «sqlshack,» SQL Shack, 21 04 2020. [En línea]. Available: https://www.sqlshack.com/an-overview-of-etl-and-elt-architecture/.

[22] Microsoft, «powerbi.microsoft.com,» [En línea]. Available: https://powerbi.microsoft.com/es-es/what-is-power-bi/.

[23] S.-G. J. A. Forero-Castañeda Deivy Alexander, «revistas.udistrital.edu.co,» Revista

Tecnol.Investig. Academia TIA, 2021. [En línea]. Available: https://revistas.udistrital.edu.co/index.php/tia/article/download/18082/17993/112771.

[24] MINISTERIO SECRETARÍA GENERAL DE LA PRESIDENCIA, «BCN,» Biblioteca del Congreso Nacional De Chile, 10 11 2022. [En línea]. Available: https://www.bcn.cl/leychile/navegar?idNorma=141599&idVersion=2020-08-26&idParte=.

[25] deloitte, «deloitte.com,» 2023. [En línea]. Available: https://www2.deloitte.com/es/es/pages/technology/articles/que-es-power-bi.html.

[26] G. R. Rivadera, «ucasal,» Universidad Católica de Salta. Facultad de Ingeniería, 30 12 2010. [En línea]. Available: http://bibliotecas.ucasal.edu.ar/opac_css/index.php?lvl=cmspage&pageid=24&id_notice=60536.

[27] J. Sreemathy, I. J. V. y S. Nisha, «Data Integration in ETL Using TALEND,» de *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, 2020.

[28] E. M. I. S. Blend Berisha, «Springer Link,» 06 08 2022. [En línea]. Available: https://link.springer.com/article/10.1186/s13677-022-00301-w#Sec2.

Capítulo 9

9. Glosario

En este capítulo, se presenta el Glosario, en el que se indica el significado de los términos técnicos y abreviaciones utilizadas en el documento.

Glosario

- **Trigger:** Gatilladores de ejecución, herramientas que permiten ejecutar procesos basados en criterios seleccionados por el usuario.
- **SP:** Stored Procedures, procedimientos almacenados. Estructura de lenguaje SQL que permite ejecutar secuencias de código.
- **DWH**: Data Warehouse. Almacén de datos.
- **Job**: Flujo de trabajo o proceso de integración de datos creado utilizando la herramienta de desarrollo de Talend Studio.
- **Log:** Registro de una ejecución, ya sea de un Jobs, queries u otro tipo de objeto ejecutable.
- **UI**: User Interface, interfaz de usuario.
- **On-premise**: Forma de implementación de software o sistemas en la que el software se instala y se ejecuta en la infraestructura local de una organización o empresa, en lugar de utilizar servicios o recursos externos en la nube.
- **Queries**: Consultas. Fragmentos de código SQL con distintas funcionalidades dentro de un DWH.
- **KPI**: Key Performance Indicator, indicador clave de rendimiento. Es una métrica cuantitativa que muestra cómo tu equipo o empresa progresa hacia tus objetivos empresariales más importantes.
- **Dashboard**: Es una herramienta personalizable de visualización de datos, que ayuda a conectar archivos, servicios, API o archivos adjuntos, y muestra estos datos como tablas, tipos de gráficas u otras visualizaciones de datos al espectador y reduce el esfuerzo manual.
- **Downtime**: Tiempo durante el cual un ATM se encuentra caído.
- **OLTP**: Procesamiento de transacciones en línea.

Anexo A

A. Detalle Jobs y SP

En este anexo se detalla cómo se distribuyen los distintos Jobs y Stored Procedures dentro del desarrollo del proyecto.

A.1. Jobs

A.1.1. Maestros

APROBADO_RECHAZADO

Se utiliza el componente “tFileInputExcel” de Excel para el origen, el cual se encargará de pasar todo lo que se encuentra en el maestro “APROBADO_RECHAZADO” hacia Snowflake, para el destino se utiliza el componente “tSnowflakeOutput”. Los dos tDBRow indican los logs si es que se realizó correctamente o fallo. La figura A.1 ilustra este proceso en Talend.

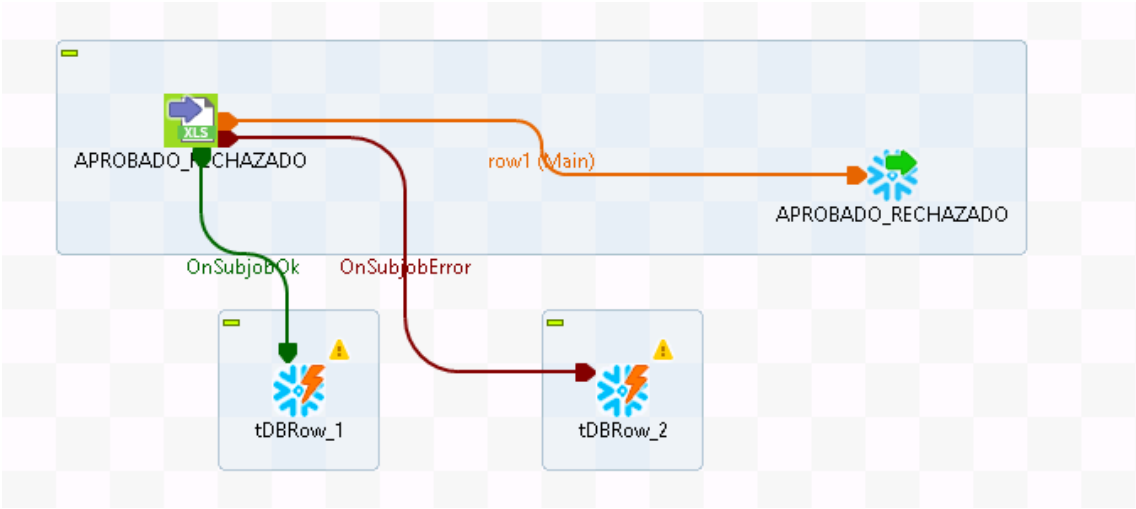


Figura A.1: Job Aprobado_Rechazado en Talend Studio

BIN

Se utiliza el componente “tFileInputExcel” de Excel para el origen, el cual se encargará de pasar todo lo que se encuentra en el Excel “BIN” hacia Snowflake, para el destino se utiliza el componente “tSnowflakeOutput” donde se llamará a la tabla del modelo (TABLA BIN) para cargar los datos que vienen del Excel. Los tDBRow finales son los encargados de ejecutar los Log. La figura A.2 muestra la construcción del Job en Talend Studio.

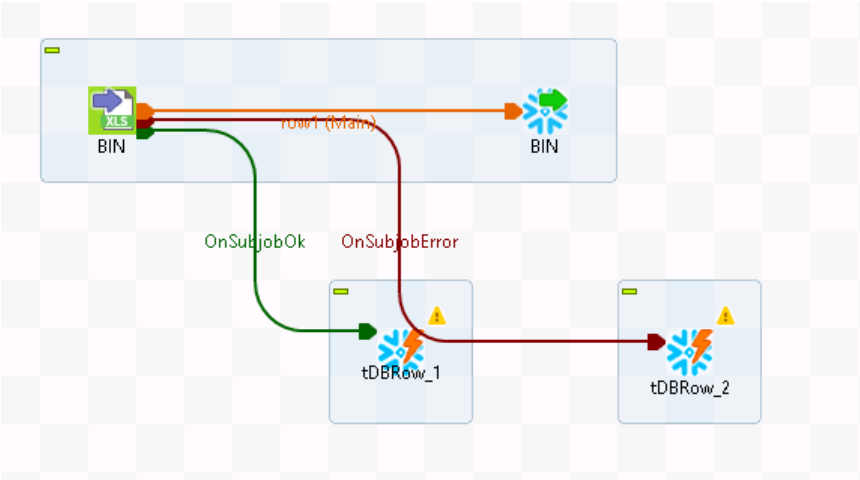


Figura A.2: Job BIN en Talend Studio

CANAL

Se utiliza el componente "tFileInputExcel" de Excel para el origen, el cual se encargará de pasar todo lo que se encuentra en el Excel "CANAL" hacia el componente "tMap", que se encarga de seleccionar las filas que se deseen, por ende, los únicos datos que se pasarán son los seleccionados. ATM_ID y LOCALIDAD_ID son identificadores que se cargarán mediante procedimientos almacenados. Para el destino se utiliza el componente "tSnowflakeOutput" para cargar todo a la tabla Canal.

Luego para llamar a los procedimientos almacenados se llamarán los componentes tSnowflakerow. Los tDBRow finales son los encargados de ejecutar los Log. La figura A.3 muestra la construcción del Job en Talend Studio.

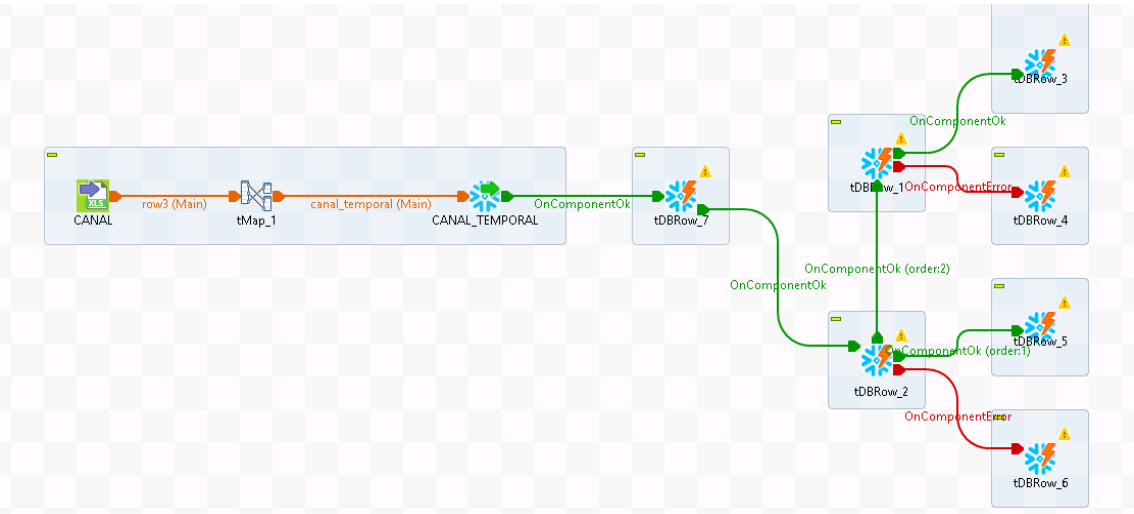


Figura A.3: Job CANAL en Talend Studio

FECHA

Se utiliza el componente “tFileInputExcel” de Excel para el origen, el cual se encargará de pasar todo lo que se encuentra en el Excel “FECHA” hacia el componente “tMap”, que se encarga de seleccionar las filas que se deseen, por ende, los únicos datos que se pasarán son los seleccionados, el único que no está relacionado es el ID que queda en blanco ya que en Snowflake el ID se calcula con procedimientos almacenados, esto quiere decir que se carga al ejecutar el SP_UPDATEIDFECHA. Y finalmente para el destino se utiliza el componente “tSnowflakeOutput”. El ultimo componente se encarga de llamar al procedimiento almacenado mencionado anteriormente. La figura A.4 muestra la construcción del Job en Talend Studio.

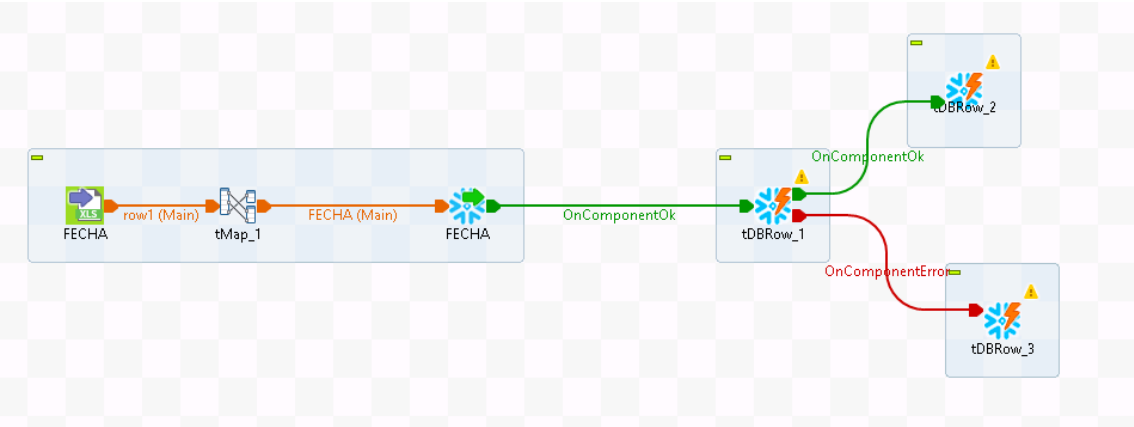


Figura A.4: Job FECHA en Talend Studio

INSTITUCION_ADQUIRENTE

Se utiliza el componente “tFileInputExcel” de Excel para el origen, el cual se encargará de pasar todo lo que se encuentra en el Excel “INSTITUCION_ADQUIRENTE” hacia Snowflake, para el destino se utiliza el componente “tSnowflakeOutput”. Los tDBRow finales son los encargados de ejecutar los Log. La figura A.5 muestra la construcción del Job en Talend Studio.

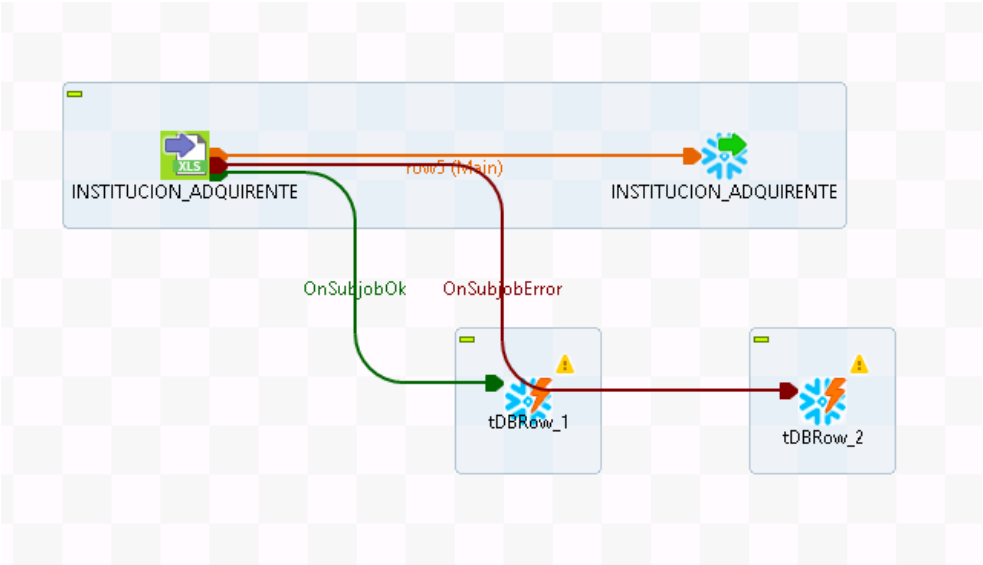


Figura A.5: Job INSTITUCION_ADQUIRIENTE en Talend Studio

INSTITUCION_EMISORA

Se utiliza el componente "tFileInputExcel" de Excel para el origen, el cual se encargará de pasar todo lo que se encuentra en el Excel "INSTITUCION_EMISORA" hacia Snowflake, para el destino se utiliza el componente "tSnowflakeOutput". Los tDBRow finales son los encargados de ejecutar los Log. La figura A.6 muestra la construcción del Job en Talend Studio.

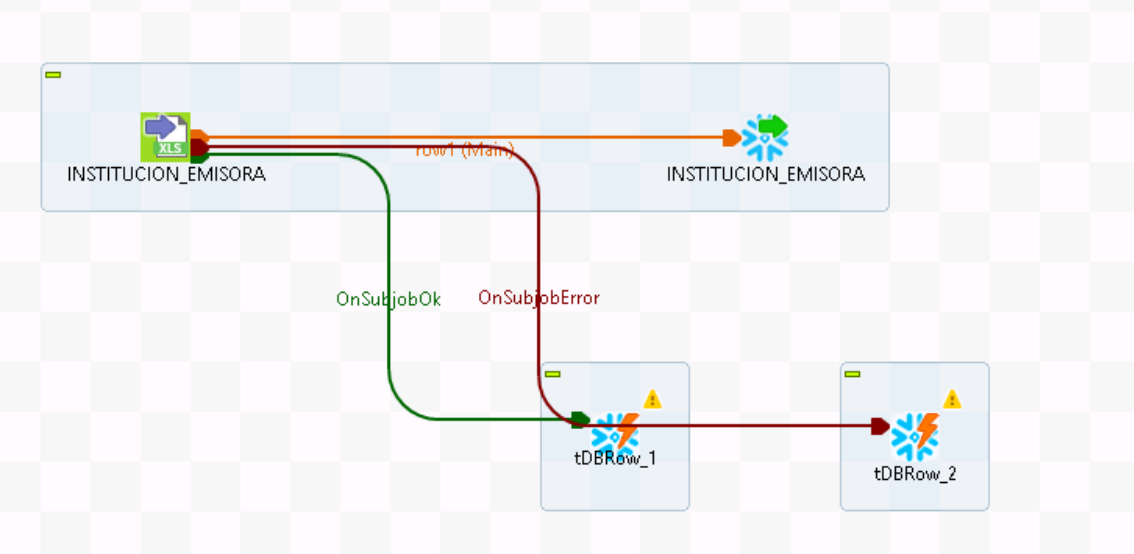


Figura A.6: Job INSTITUCION_EMISORA en Talend Studio

METODO_AUTORIZACION

Se utiliza el componente "tFileInputExcel" de Excel para el origen, el cual se encargará de pasar todo lo que se encuentra en el Excel "METODO_AUTORIZACION" hacia Snowflake, para el destino se utiliza el componente "tSnowflakeOutput". Los tDBRow finales son los

encargados de ejecutar los Log. La figura A.7 muestra la construcción del Job en Talend Studio.

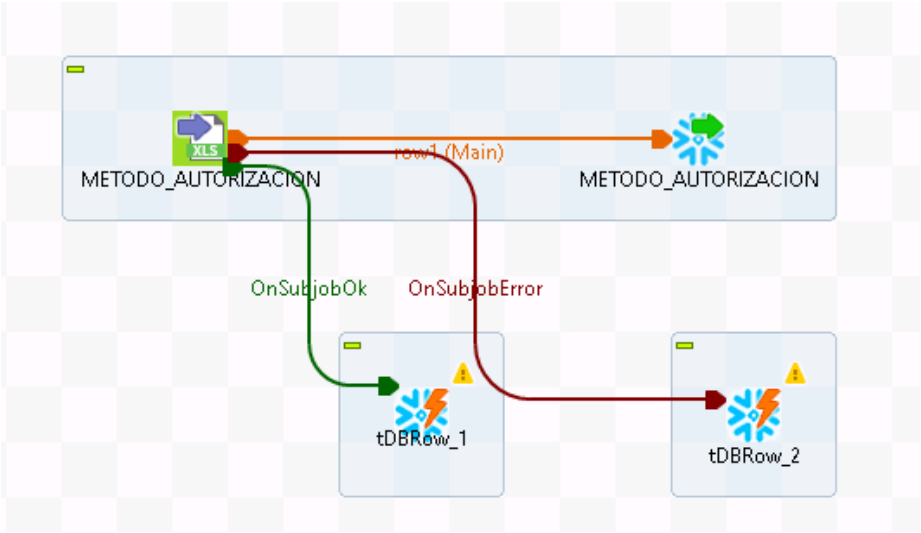


Figura A.7: Job METODO_AUTORIZACIÓN en Talend Studio

TIEMPO

Se utiliza el componente "tFileInputExcel" de Excel para el origen, el cual se encargará de pasar todo lo que se encuentra en el Excel "TIEMPO" hacia el componente "tMap", que se encarga de seleccionar las filas que se deseen, por ende, los únicos datos que se pasarán son los seleccionados, el ID aparece como Numeric.sequence("s1",1,1) ya que se cargará un ID incremental, y por último, el único que no está relacionado es el TIEMPO que queda en blanco ya que en Snowflake se calcula con procedimientos almacenados, esto quiere decir que se carga al ejecutar el SP_UPDATETIEMPO. Para el destino se utiliza el componente "tSnowflakeOutput". El último componente se encarga de llamar al procedimiento almacenado mencionado anteriormente. La figura A.8 muestra la construcción del Job en Talend Studio.

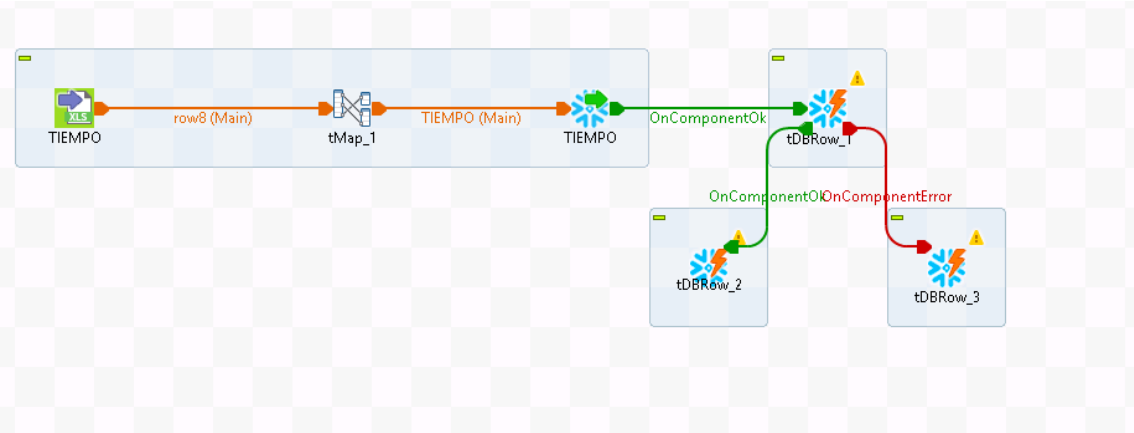


Figura A.8: Job TIEMPO en Talend Studio

TIPO_PROCESAMIENTO

Se utiliza el componente “tFileInputExcel” de Excel para el origen, el cual se encargará de pasar todo lo que se encuentra en el Excel “TIPO_PROCESAMIENTO” hacia Snowflake, para el destino se utiliza el componente “tSnowflakeOutput”. Los tDBRow finales son los encargados de ejecutar los Log. La figura A.9 muestra la construcción del Job en Talend Studio.

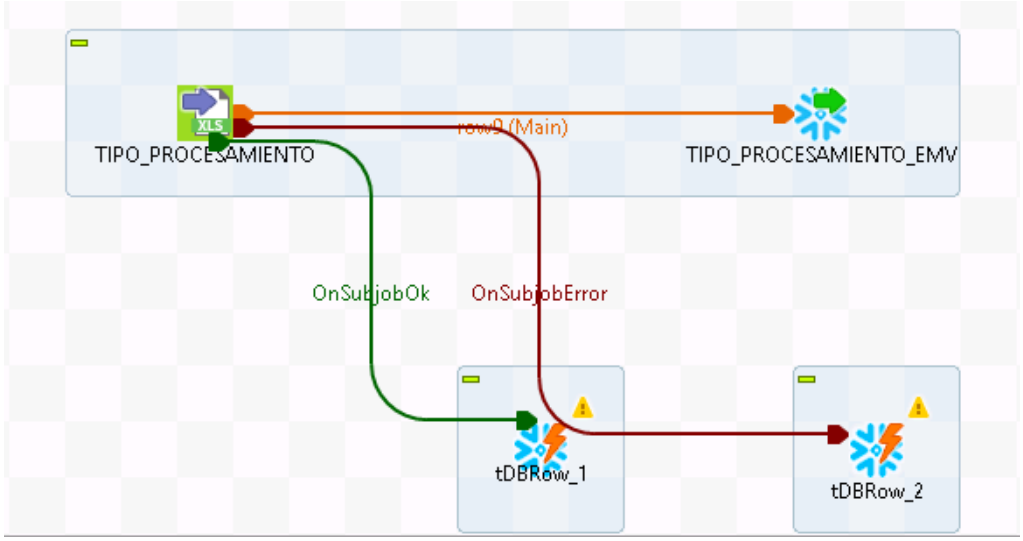


Figura A.9: Job TIPO_PROCESAMIENTO en Talend Studio

TIPO_TRANSACCION

Se utiliza el componente “tFileInputExcel” de Excel para el origen, el cual se encargará de pasar todo lo que se encuentra en el Excel “TIPO_TRANSACCION” hacia Snowflake, para el destino se utiliza el componente “tSnowflakeOutput”. Los tDBRow finales son los encargados de ejecutar los Log. La figura A.10 muestra la construcción del Job en Talend Studio.

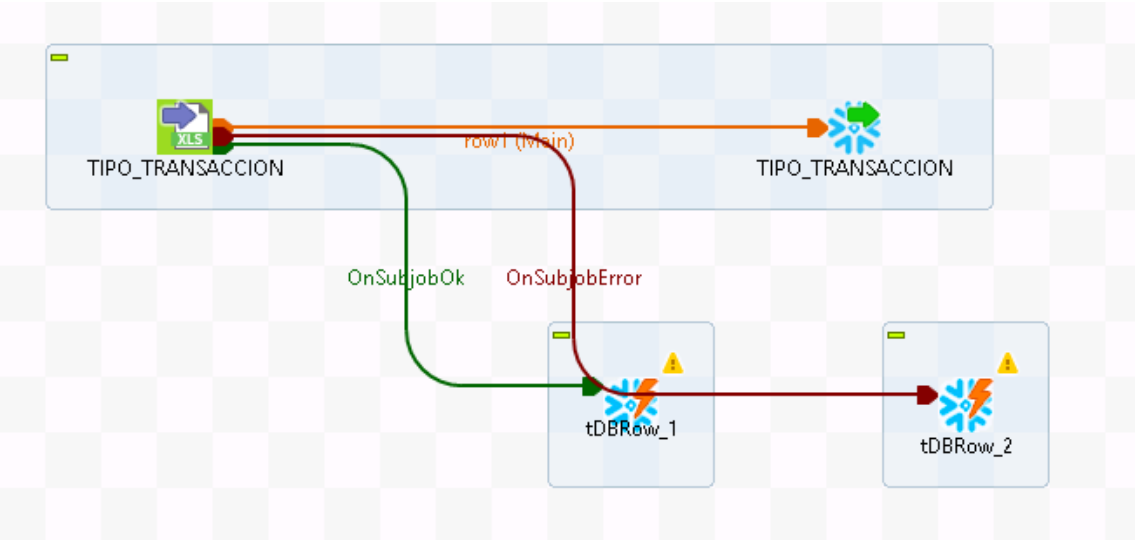


Figura A.10: Job TIPO_TRANSACCIÓN en Talend Studio

A.1.2. Vistas

DWH

Primero se consulta la fecha máxima y versión de la vista y se guarda en una variable, después se borran en la tabla destino los registros con esa fecha. Luego se consulta la vista con el filtro de la máxima fecha y versión y se cargan las filas a la tabla en Staging. Los dos últimos tDBRow hacen referencia a los Log. La figura A.11 muestra la construcción del Job en Talend Studio.

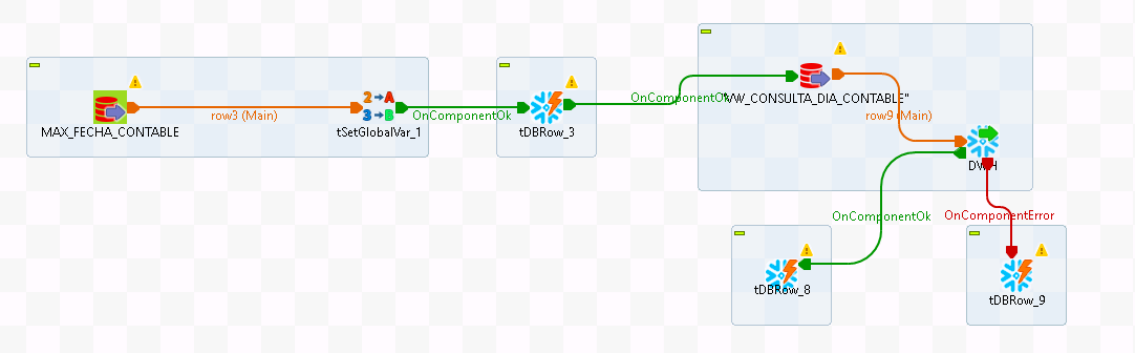


Figura A.11: Job DWH en Talend Studio

SP_TRANSACCIONES

Este Job contiene cada uno de los SP que se ejecutan para cargar los datos del DWH a TRANSACCIONES. La figura A.12 ilustra la construcción del Job en Talend Studio.

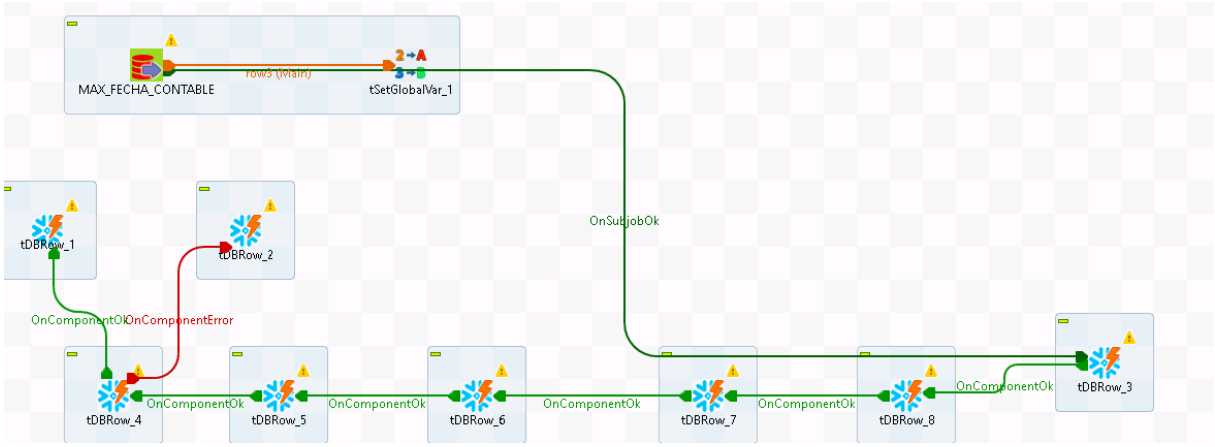


Figura A.12: Job SP_TRANSACCIONES en Talend Studio

Los dos ultimos tDBRow, que se conectan a tDBRow_4, son los que hacen referencia a los Log.

PROVIEW-ATM

El primer componente lo que hace es llamar a la vista DM_PROVIEWATM, para pasarlo a lo que está en Staging, a la tabla BASE_PROVIEW. Luego los 3 últimos componentes es correr 3 Stored procedures que ocupan esta tabla, como por ejemplo cargar a la tabla ATM los datos que vienen de la vista, luego el segundo cargar a la tabla ATM_ACTUAL y finalmente llama al SP para cargar a la tabla LOCALIDAD. Los tDBRow que conectan al componente DM_PROVIEWATM son los que hacen referencia a los Log. La figura A.13 muestra el Job en Talend.

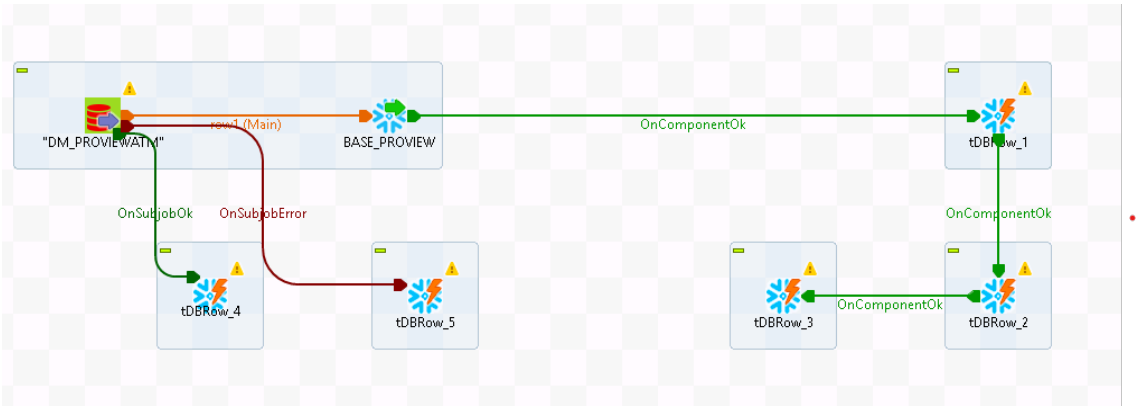


Figura A.13: Job PROVIEW-ATM en Talend Studio

PROVIEW-EVENTOS

Primero se consulta la fecha máxima de la vista (IB24) y se guarda en una variable, después se borran en la tabla destino los registros con esa fecha. Luego se consulta la vista con el filtro de la máxima fecha y se cargan las filas a la tabla en Staging, para finalmente correr los distintos

SP, evaluado en la máxima fecha. Los dos tDBRow que se conectan a tDBRow_2 hacen referencia a los Log. La figura A.14 muestra el Job en Talend.

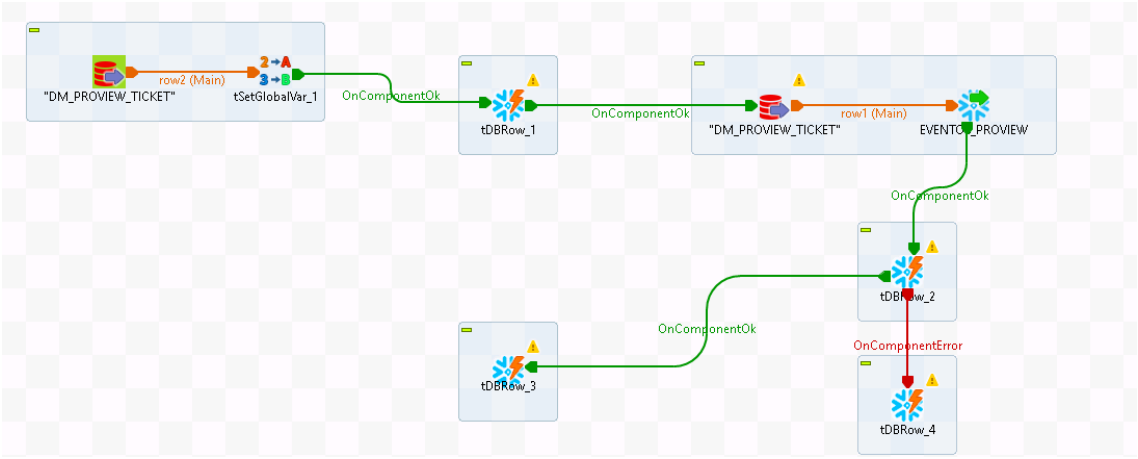


Figura A.14: Job PROVIEW-EVENTOS en Talend Studio

A.2. Stored Procedures

SP_CARGARATM

Carga todos los atributos que vienen de la vista BASE_PROVIEW a la tabla ATM. Para formar el Identificador único y que un cajero sea exclusivo, se utiliza la fecha de partida (FECHAS_PARTIDA) que es la fecha en la cual empieza a funcionar el cajero.

Además, se modificó la hora debido a que vienen como una cadena de números, los cuales indican la hora de inicio y fin del funcionamiento del cajero. La función TIME_FROM_PARTS substrae la hora, minuto y segundos, pero en este caso solo se requiere la hora y minutos, por ende, en los segundos se destina el 0 y finalmente la hora queda separada por hora de inicio y fin. Un ejemplo de esto es: HORARIO_OPERACION_LUNES_INICIO y HORARIO_OPERACION_LUNES_FIN.

La figura A.15 presenta el diagrama de procesos correspondiente a este Stored Procedure.

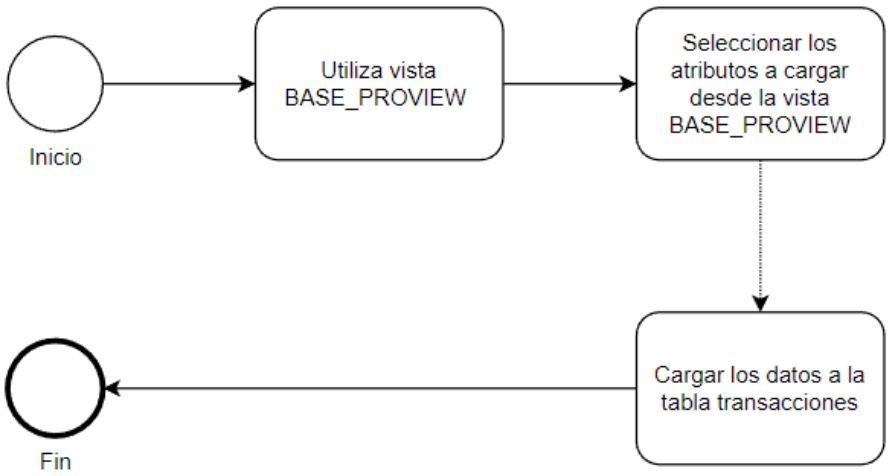


Figura A.15: Diagrama funcionamiento SP_CARGARATM

SP_CARGARLOCALIDAD

Para formar la tabla LOCALIDAD, se emplea un ID incremental (ID), país que se utiliza CHL (Chile) ya que no hay ningún otro país y los demás atributos se extraen de BASE_PROVIEW sin realizar modificación alguna.

La figura A.16 presenta el diagrama de procesos correspondiente a este Stored Procedure.

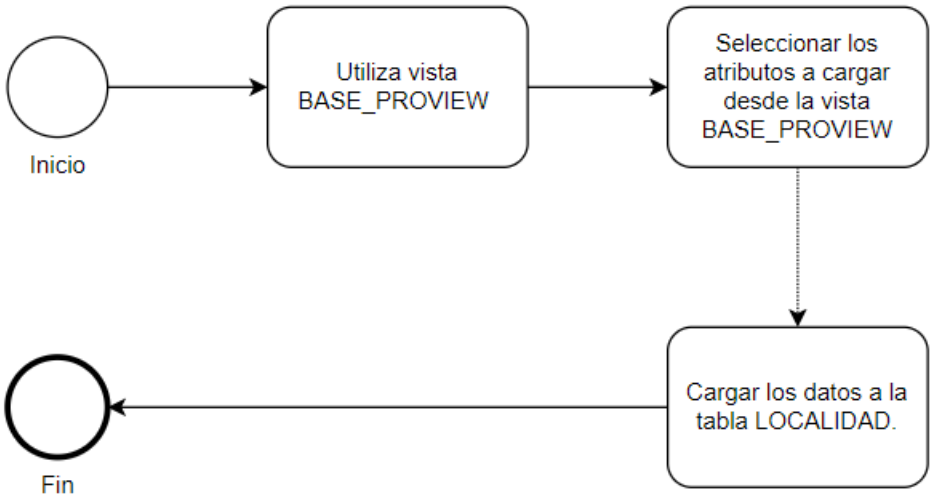


Figura A.16: Diagrama funcionamiento SP_CARGARLOCALIDAD

SP_CARGARCANAL

Este procedimiento carga los datos que están en una tabla temporal que contiene todos los datos que se introdujeron desde el maestro Canal, para luego pasarlos a la tabla CANAL del modelo analítico. La figura A.17 presenta el diagrama de procesos correspondiente a este Stored Procedure.

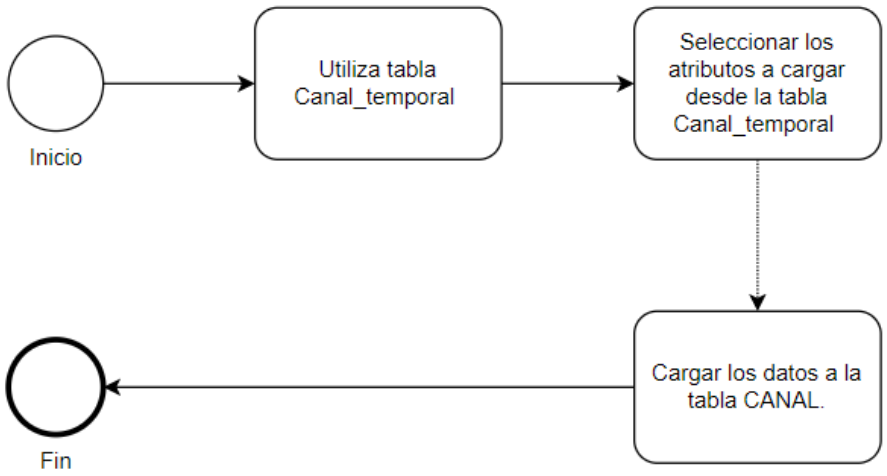


Figura A.17: Diagrama funcionamiento SP_CARGARCANAL

SP_CARGARATMACANAL

Este procedimiento inserta el identificador único de la tabla ATM a la tabla CANAL que son de tipo ATM. La figura A.18 presenta el diagrama de procesos correspondiente a este Stored Procedure.

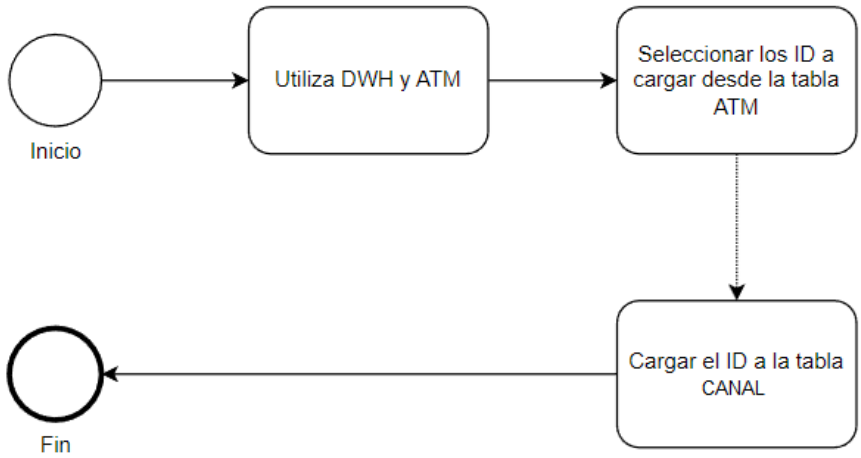


Figura A.18: Diagrama funcionamiento SP_CARGARATMACANAL

SP_CARGARLOCALIDADACANAL

Este procedimiento inserta el identificador único de la tabla LOCALIDAD a la tabla CANAL que son de tipo ATM. La figura A.19 presenta el diagrama de procesos correspondiente a este Stored Procedure.

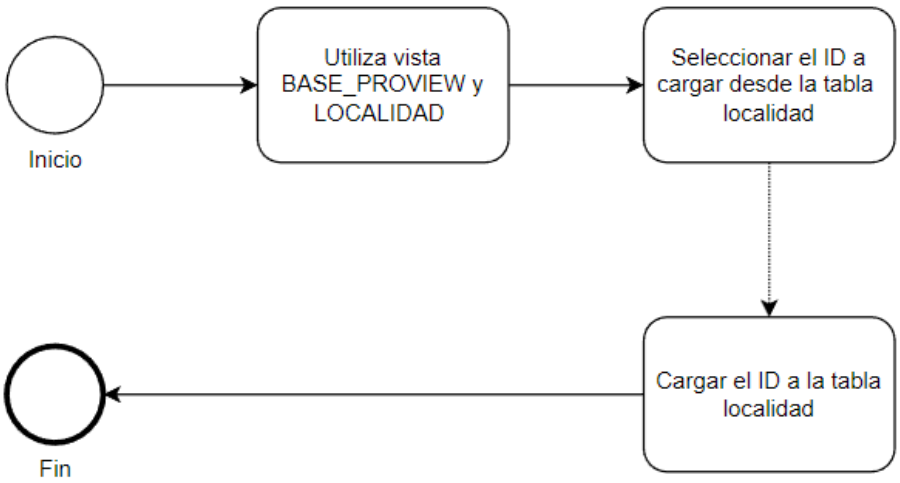


Figura A.19: Diagrama funcionamiento SP_CARGARLOCALIDADACANAL

SP_CARGARTRANSACCIONES

Este procedimiento carga cada uno de los atributos que provienen de la vista DWH a la tabla del Modelo Analítico Transacciones, los cuales se insertan por día contable. La figura A.20 presenta el diagrama de procesos correspondiente a este Stored Procedures.

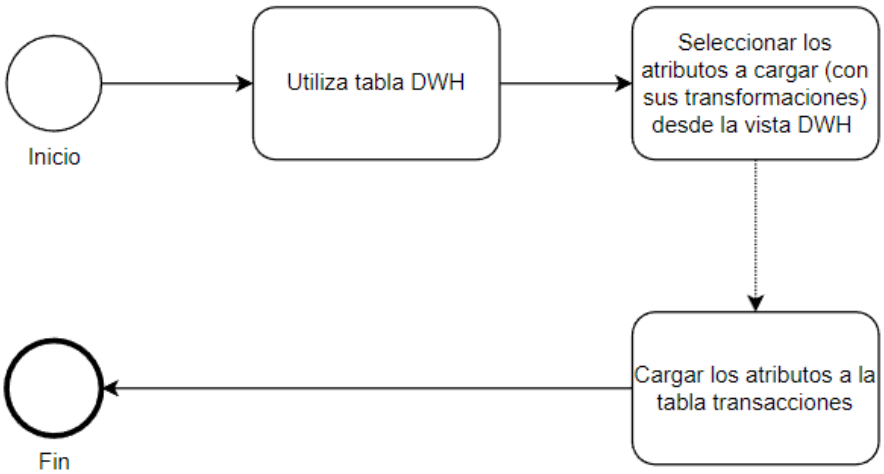


Figura A.20: Diagrama funcionamiento SP_CARGARTRANSACCIONES

SP_CARGARBINTRX

Este procedimiento inserta el identificador único de la tabla BIN a la tabla TRANSACCIÓN, por lo que une la vista DWH, el maestro y la tabla transacción. La figura A.21 presenta el diagrama de procesos correspondiente a este Stored Procedure.

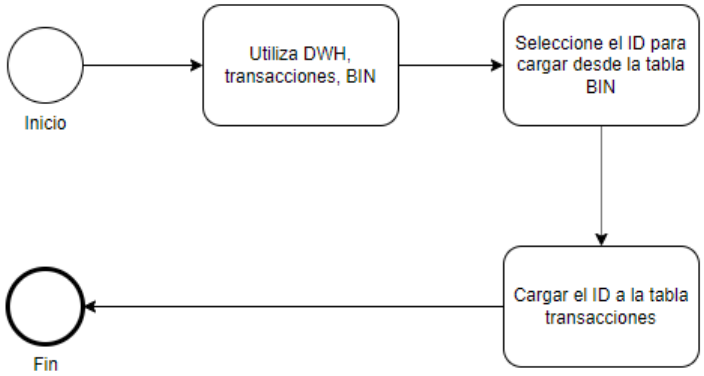


Figura A.21: Diagrama funcionamiento SP_CARGARBINTRX

SP_CARGARMETODOAUTORIZACIONTRX

Este procedimiento inserta el identificador único de MÉTODO AUTORIZACIÓN a la tabla TRANSACCIÓN, desde maestro denominado METODO AUTORIZACION. Este procedimiento se realiza debido a que el maestro METODO AUTORIZACION tiene un identificador numérico y no utiliza el id que proviene de la empresa que es el AUTH METHOD, por lo que hace el Join entre la vista DWH, el maestro y la tabla transacción.

La figura A.22 presenta el diagrama de procesos correspondiente a este Stored Procedure.

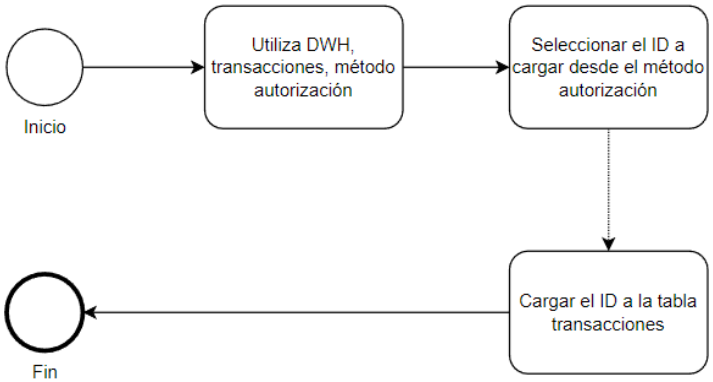


Figura A.22: Diagrama funcionamiento SP_CARGARMETODOAUTORIZACIONTRX

SP_CARGARCANALTRX

Este procedimiento carga el identificador único del maestro CANAL a la tabla transacciones. Aclarar, que el identificador único del maestro CANAL e incremental con el fin que sea numérico, dado que el identificador único que proviene de la empresa cliente que es la concatenación entre TERMINAL NUMBER, TRML TYPE CODE y TLML SUBTYPE CODE lo guarda como varchar dado que contiene algunas letras. Es por eso, que se realiza una vista

donde almacena el Join que se realiza entre la vista DWH, el maestro y la tabla transacción, para luego hacer una modificación y cargar el ID INCREMENTAL A la tabla TRANSACCIONES. La figura A.23 presenta el diagrama de procesos correspondiente a este Stored Procedure.

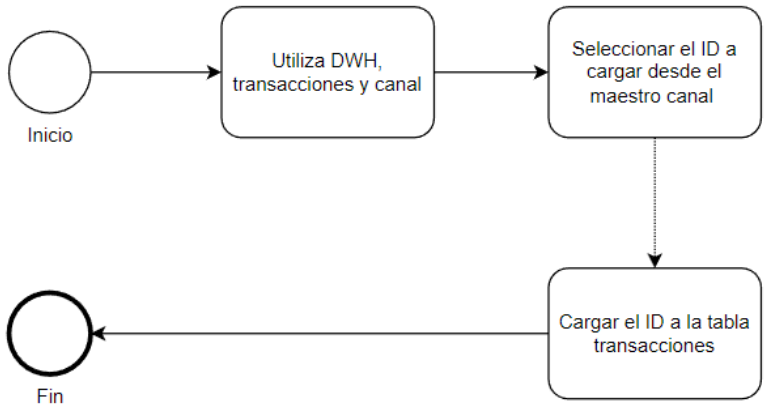


Figura A.23: Diagrama funcionamiento SP_CARGARCANALTRX

SP_CARGARTIPOTRANSACCIONTRX

Este procedimiento carga el identificador único del maestro TIPO_TRANSACCION a la tabla TRANSACCIONES. La figura A.24 presenta el diagrama de procesos correspondiente a este Stored Procedure.

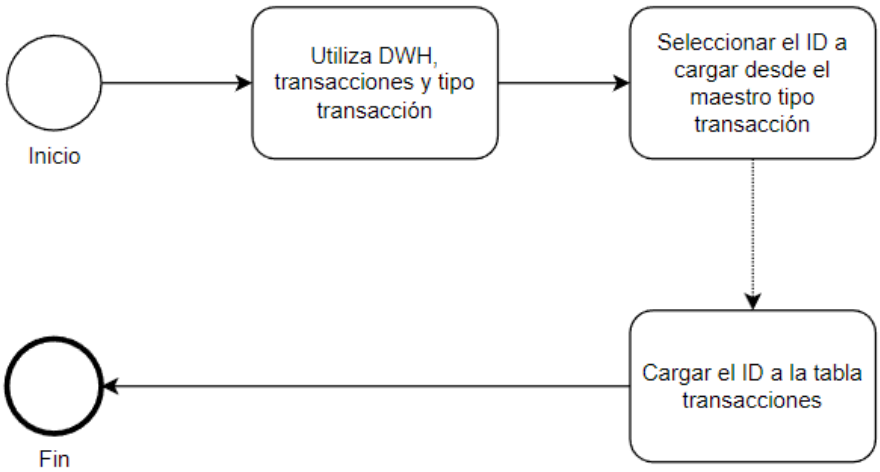


Figura A.24: Diagrama funcionamiento SP_CARGARTIPOTRANSACCIONTRX

SP_CARGARPROCESAMIENTOTRX

Este procedimiento carga los identificadores únicos de la tabla TIPO DE PROCESAMIENTO a la tabla TRANSACCIONES, denominados “ID_TIPO_PROC”. Se crea una vista VIEW_PROCESAMIENTO y se crean los siguientes casos:

Si f.tlf_flex_vgest_id_52 = 0 y f.tlf_flex_vgest_id_57 = 0 y f.tlf_flex_vgest_id_53 = 0 indica

que es banda.

Si f.tlf_flex_vgest_id_52 = 1 y f.tlf_flex_vgest_id_57 = 0 y f.tlf_flex_vgest_id_53 = 0 indica que es con Chip.

Si f.tlf_flex_vgest_id_52 = 1 y f.tlf_flex_vgest_id_57 = 0 y f.tlf_flex_vgest_id_53 = 1 indica que es con Fallback.

Si f.tlf_flex_vgest_id_52 = 1 y f.tlf_flex_vgest_id_57 = 1 y f.tlf_flex_vgest_id_53 = 0 indica que es con Contactless.

En caso de no cumplir con ninguna de las anteriores se asigna el número 9 que corresponde NA. La figura A.25 presenta el diagrama de procesos correspondiente a este Stored Procedure.

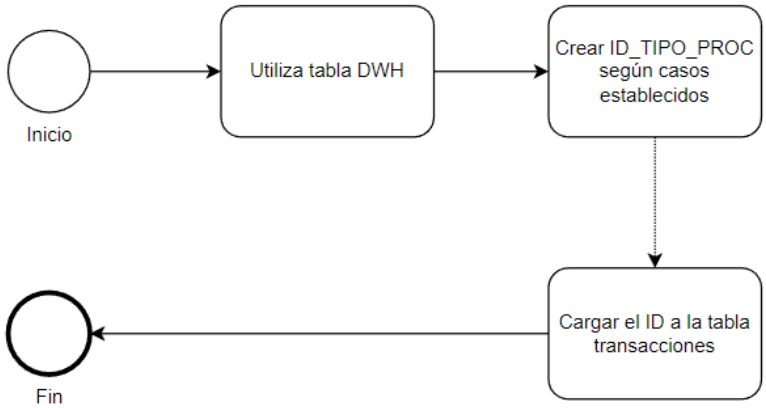


Figura A.25: Diagrama funcionamiento SP_CARGARPROCESAMIENTOTRX

SP_UPDATEIDFECHA

Modifica el ID del maestro de fecha a través de un método UPDATE (actualización). La figura A.26 presenta el diagrama de procesos correspondiente a este Stored Procedure.

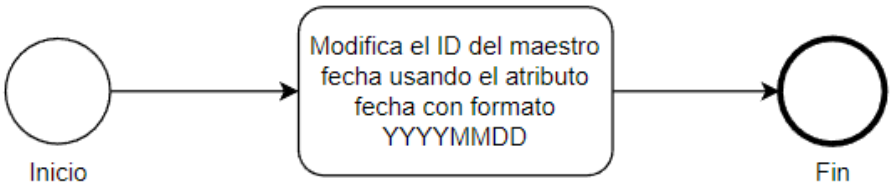


Figura A.26: Diagrama funcionamiento SP_UPDATEIDFECHA

SP_UPDATETIEMPO

Modifica el tiempo con la función TIME_FROM_PARTS que indica la hora, minuto y para los segundos un 0. Solamente consideramos hora, minuto dado que la empresa solo considera estas dos, dejando fuera los segundos. La figura A.27 presenta el diagrama de procesos correspondiente a este Stored Procedure.

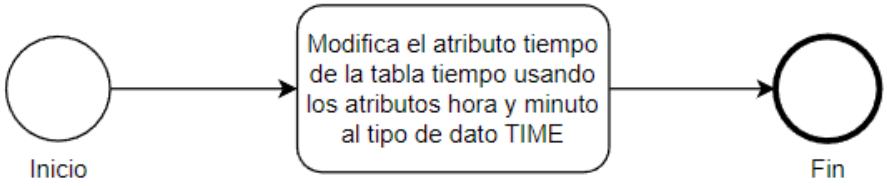


Figura A.27: Diagrama funcionamiento SP_UPDATETIEMPO

SP_CARGARTIEMPOTRX

Se carga el Identificador único (ID) del maestro TIEMPO a la tabla TRANSACCIONES. Se crea una vista view_tiempo y luego hace un UPDATE (actualización) para modificar e insertar el ID a TRANSACCIONES. La figura A.28 presenta el diagrama de procesos correspondiente a este Stored Procedure.

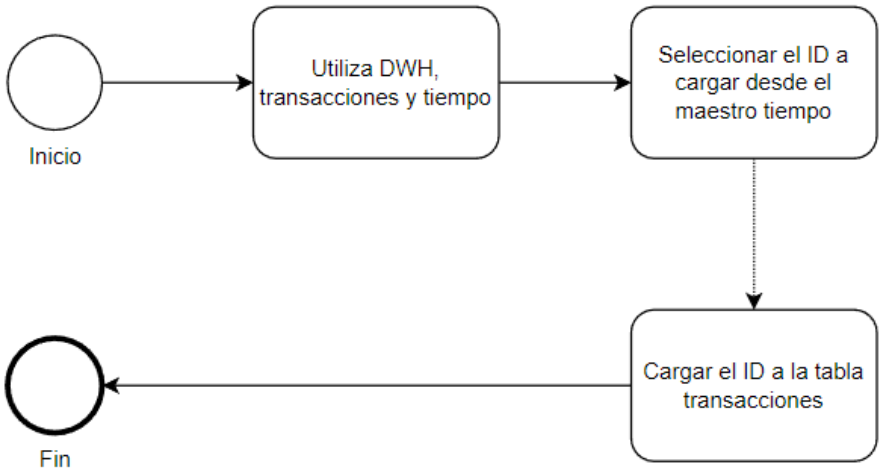


Figura A.28: Diagrama funcionamiento SP_CARGARTIEMPOTRX

SP_CARGARATM_ACTUAL

Se inserta el ID y ATM_ID con la fecha máxima de partida. Sirve para guardar todos los ATM más actuales sin perder información.

La figura A.29 presenta el diagrama de procesos correspondiente a este Stored Procedure.

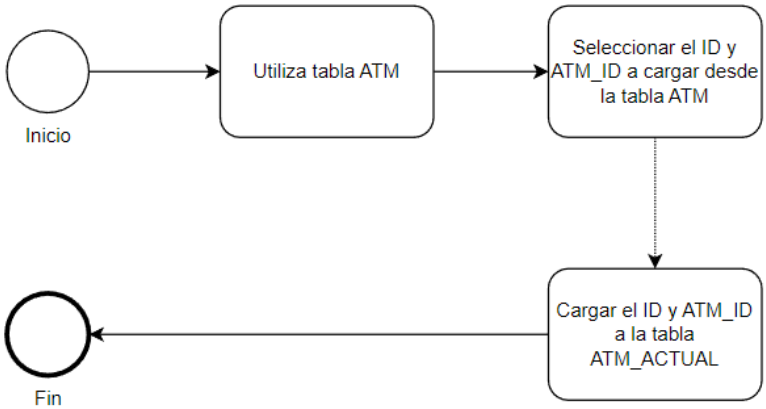


Figura A.29: Diagrama funcionamiento SP_CARGARATM_ACTUAL

SP_CARGARINCIDENTES

Este procedimiento guarda todos los incidentes producidos en el cajero. Vienen de la tabla EVENTOS_PROVIEW. El identificador único de esta tabla incidentes se crea a través de un identificador incremental, se llama al campo ATM ID de la tabla ATM, el campo TIEMPO ID y FECHA ID se forma a través de la fecha de la columna IB24, y el DOWNTIME es el tiempo en segundo que indica cuanto estuvo el cajero fuera de servicio, que se obtiene con la resta entre el campo IB24 y CB24. Si el DOWNTIME tiene un -1 es porque el CB24 está blanco y eso indica que el atm aún no está operativo.

Para diferenciar que el cajero operativo o pendiente, se agregó una columna ESTADO donde indica "FINALIZADO" si el cajero se operativo y si es "PENDIENTE'' porque sigue sin funcionar.

La figura A.30 presenta el diagrama de procesos correspondiente a este Stored Procedure.

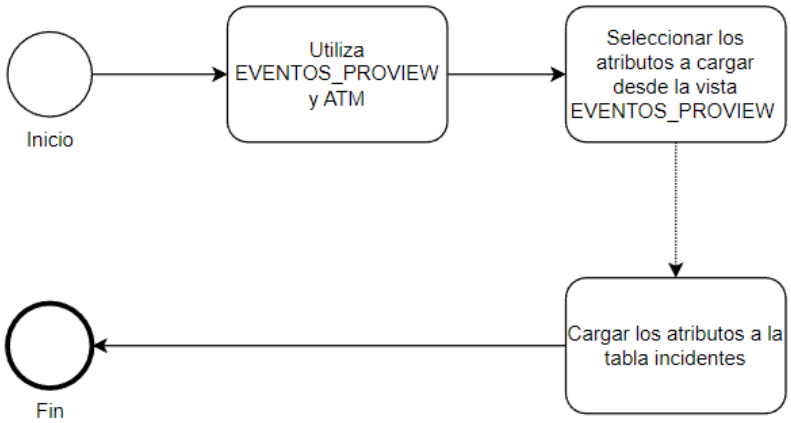


Figura A.30: Diagrama funcionamiento SP_CARGARINCIDENTES