



---

# PROYECTO PASANTÍA

Llevado a cabo en el área de Gobierno de Datos de Walmart Chile

---

## Implementación de un repositorio de datos para las operaciones realizadas en el área de la cadena de suministros de Walmart Chile

---

*Romina Gheza Schimunek*

Proyecto para optar al título de Ingeniería Civil Industrial con mención en Business Analytics de la Facultad de Ingeniería y Ciencias de la Universidad Adolfo Ibáñez

Profesor guía:

*Raimundo Sánchez*

Santiago, Chile

2023

---

## Resumen ejecutivo

En el marco de la evolución tecnológica y optimización de procesos, el presente trabajo se enfoca en como la digitalización permite mejorar la eficiencia operativa de las empresas.

El proyecto ingenieril fue realizado en la empresa Walmart Chile. Para esta empresa, el nivel de satisfacción de sus clientes es sumamente importante, por lo que hay diversos procesos que generan que este indicador tenga valores positivos. Dentro de estos procesos, se encuentran todas las áreas, que constantemente están en la búsqueda por mayor eficiencia en sus procesos para cumplir con el eslogan *“Precios bajos todos los días”*. En particular, el proyecto, cuyo objetivo es gobernar y disponibilizar la data necesaria para sustentar las operaciones relacionadas con el área de la cadena de suministros de Walmart Chile, se centró en la implementación de un repositorio de datos propio utilizando la herramienta de Google Cloud Platform.

Producto de la gran cantidad de datos que maneja la empresa, no siempre es conveniente considerar la totalidad de usuarios /consultas para migrar en primera instancia. Con ello, se establece un plan de migración cuyo propósito es priorizar los casos de negocio que generan más impacto en las operaciones de la empresa.

Teniendo en cuenta el sistema migrado, se realizan una serie de pasos que permiten documentar el tipo de datos asociados a las tablas provenientes de distintas fuentes externas no gobernadas, para generar un modelo de datos y habilitar un entorno para la ingesta de estos en un repositorio de datos propio de la empresa.

Como resultado de la implementación del proyecto, se permitió reducir la cantidad de horas que el sistema tenía intermitencias gracias al conexón de un 14% de las consultas realizadas por distintos usuarios, que consumen una data cuya calidad es del 98%.

*Palabras claves:* reabastecimiento, eficiencia, conexión, calidad, gobierno de datos, entorno, repositorio.

---

## Abstract

In the context of technological evolution and process optimization, this work focuses on how digitization enables the improvement of operational efficiency in companies.

The engineering project was carried out at Walmart Chile. For this company, customer satisfaction is paramount, leading to various processes that contribute to positive satisfaction indicators. Among these processes are all areas that are constantly seeking greater efficiency in order to fulfill the slogan "Everyday Low Prices." Specifically, the project, which aims to govern and make available the necessary data to support operations related to Walmart Chile's supply chain, focused on implementing its data repository using the Google Cloud Platform tool.

Due to the vast amount of data handled by the company, it is not always practical to consider the entire set of users/queries for migration initially. Therefore, a migration plan is established with the purpose of prioritizing business cases that have a greater impact on the company's operations.

Considering the migrated system, a series of steps are taken to document the type of data associated with tables from various ungoverned external sources. This process generates a data model and sets up an environment for ingesting this data into the company's data repository.

As a result of the project's implementation, it was possible to reduce the system's intermittent hours by connecting 14% of queries performed by different users, consuming data with a 98% quality.

*Keywords:* replenishment, efficiency, connection, quality, data governance, environment, repository.

---

## Índice

|   |    |
|---|----|
| Resumen ejecutivo .....                 | 2  |
| Abstract .....                          | 3  |
| 1. Introducción .....                   | 5  |
| 1.1 Contexto de la empresa .....        | 5  |
| 1.2 Contexto del problema .....         | 6  |
| 1.3 Identificación de oportunidad ..... | 7  |
| 2. Objetivos .....                      | 8  |
| 2.1 Objetivo general .....              | 8  |
| 2.2 Objetivos específicos .....         | 8  |
| 2.3 Medidas de desempeño .....          | 10 |
| 3. Estado del arte .....                | 11 |
| 4. Metodología .....                    | 16 |
| 5. Implementación .....                 | 22 |
| 6. Análisis de riesgo .....             | 24 |
| 7. Evaluación económica .....           | 26 |
| 8. Resultados .....                     | 28 |
| 9. Conclusiones .....                   | 29 |
| 10. Referencias .....                   | 30 |
| 11. Anexo .....                         | 32 |

## 1. Introducción

### 1.1 Contexto de la empresa

Walmart es una reconocida compañía multinacional estadounidense fundada en 1969 en la ciudad de Bentonville, Arkansas, Estados Unidos. Gracias a su constante proceso de mejora continua, Walmart ha expandido su presencia alrededor del mundo, llevándolo a que hoy en día sea una de las empresas líder a nivel mundial en la industria minorista. En la actualidad, esta empresa abastece a sus distintos clientes ofreciéndoles una amplia variedad de productos bajo la estrategia Every Day Low Price (EDLP) o Precios Bajos Todos Los Días. Su oferta no solo considera productos comestibles, sino que también proporciona productos deportivos, de entretenimiento y artículos para el hogar.

Walmart Internacional tiene más de 5.100 unidades minoristas y cuenta con aproximadamente 550.000 colaboradores distribuidos en todo el mundo. Dentro de los mercados reconocidos se encuentran África, Canadá, Centroamérica, Chile, China, India y México. Adicionalmente, se pueden nombrar Brasil, Japón y Reino Unido, mercados donde Walmart mantiene inversiones de capital.

En particular, Walmart Chile es uno de los principales protagonistas en la industria de las ventas al por menor, cuya actividad comercial consiste en la venta de alimentos y mercadería en los distintos supermercados que dispone por medio de sus más de 370 locales distribuidos a lo largo de Chile, donde se pueden encontrar formatos tales como Lider, Lider.cl, Express Lider, SuperBodega aCuenta, aCuenta.cl y Central Mayorista. A través de estos distintos canales, Walmart Chile ofrece a sus clientes un extenso y atractivo portafolio de productos cuya procedencia recae en 12 marcas propias asociadas a distintas categorías, donde se pueden mencionar limpieza, alimentos, perfumería, entre otras.

Para lograr llevar a cabo las distintas operaciones de la empresa, hay un gran equipo que utiliza distintas herramientas para gestionar y optimizar los distintos canales de distribución. Estas herramientas son empleadas en distintas áreas y en este caso en particular, es importante mencionar que estas permiten

generar valor en los datos para apoyar aquellas decisiones que deben tomarse durante el transcurso de los días.

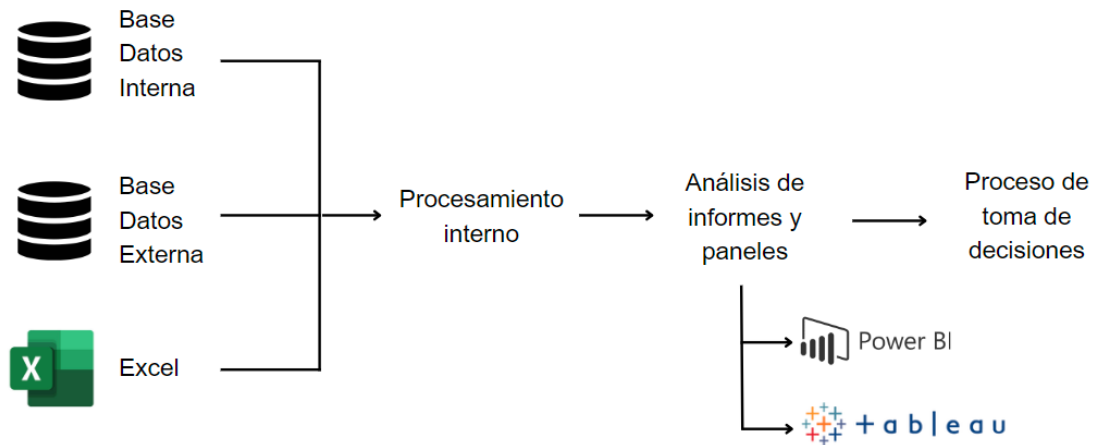
## 1.2 Contexto del problema

Hoy en día, resulta impredecible utilizar distintas fuentes de datos no solo en la industria como tal, sino también a nivel social. El uso de estas bases de datos en las empresas resulta fundamental para la ejecución de sus operaciones y para generar impacto en diversos ámbitos, tales como la eficiencia y la toma de decisiones. Walmart no está exento de esta situación ya que, gracias al gran volumen de sus operaciones, es sumamente necesario contar con estas fuentes para lograr tener conocimiento asociado a los centros de distribución e información relacionada a sus proveedores.

Ahora bien, es muy importante comprender que el uso de datos no tiene sentido si no se agrega valor a estos. Por esta razón, llevar a cabo lo que hoy en día se conoce como Business Intelligence (BI) o Inteligencia de negocios, se torna esencial para la empresa, un proceso en donde se utilizan distintas herramientas y prácticas para garantizar que el procedimiento de toma de decisiones cuente con un respaldo cuantificable. Luego, para poder realizar este proceso, se deben recopilar, procesar y analizar los datos para generar una vista o dashboard, dando paso a la última etapa relacionada con la toma de decisiones.

En particular, en el entorno de Walmart Chile, los datos se encuentran almacenados en tres tipos de fuentes. En primer lugar, se tienen aquellas fuentes que resultan ser bases de datos internas generadas y gobernadas por Walmart Chile. En segundo lugar, se tiene información proveniente de fuentes externas, en donde no existe un dominio por parte de la empresa y, por último, se tienen datos provenientes de distintos Excel.

Para ayudar a la comprensión y resumir la información del contexto que hoy en día vive la empresa, se entrega el siguiente esquema:



Esquema 1: “Proceso de Inteligencia Empresarial”

### 1.3 Identificación de oportunidad

Para comprender la oportunidad identificada, es importante tener en cuenta que aquellas fuentes cuyos dominios no son parte de la propiedad de Walmart, generan una dependencia de parte de la empresa hacia el proveedor del servicio. Esta dependencia afecta a dos entidades.

- Por un lado, se tiene a los usuarios, que están expuestos a experimentar cambios inesperados en cuanto a las políticas de acceso, mientras que también pueden presenciar interrupciones del sistema y problemas técnicos como consecuencia de un sobre consumo de las fuentes, afectando el proceso de toma de decisiones.
- Desde la perspectiva de la empresa, dado que los datos son externos, estos no siempre se adaptan a las necesidades que busca satisfacer la empresa y no se tiene control en cuanto a la calidad, privacidad y seguridad de los datos. Adicionalmente, dependiendo del tipo de fuente, la empresa debe disponer grandes cantidades de dinero para acceder a estas fuentes y mantener un equipo que sea capaz de brindar soporte, no solo a los usuarios, sino que también a las mismas bases de datos para que tengan un funcionamiento relativamente estable.

Teniendo lo anterior en cuenta, hay distintos procedimientos o alternativas que pueden brindar una solución a corto plazo, que si bien, permitirán a la empresa ahorrar algunos costos y evitar vulnerabilidades, a futuro estos problemas pueden surgir nuevamente. Por esta razón, la oportunidad que se tiene en estos momentos es centralizar los datos en un repositorio cuyo dominio es 100% parte de la propiedad de Walmart. Implementando un repositorio se logrará generar un estándar de calidad de datos, información centralizada, control de accesos, soporte centralizado, generación y disponibilidad de documentación y uno de los puntos más relevantes, la creación y acumulación de historia.

## 2. Objetivos

Lo que busca este proyecto es: “Gobernar y disponibilizar la data necesaria para sustentar las operaciones relacionadas con el área de la cadena de suministros de Walmart Chile, generando así un beneficio no solo con relación al dinero ahorrado asociado al soporte, sino que también en la calidad y confiabilidad de los datos.”

### 2.1 Objetivo general

Asegurar calidad y disponibilidad de los datos existentes para las operaciones del área de reabastecimiento a través de la democratización y uso efectivo de data gobernada en el repositorio de Walmart.

### 2.2 Objetivos específicos

Para cumplir con el objetivo general, es necesario tomar en cuenta tres objetivos específicos. En primer lugar, incrementar en un 40% el volumen de datos validados en el transcurso de los próximos cuatro meses. En segundo lugar, desarrollar un plan de migración de usuarios con el objetivo de incrementar en un 20% la proporción de usuarios que acceden y utilizan la información almacenada en el repositorio. Finalmente, elevar en un 20% la cantidad de consultas vinculadas al Data Lake en el transcurso de los próximos cuatro meses.



A continuación, se presenta un modelo de optimización que busca maximizar la cantidad de usuarios que están conectados a un sistema considerando algunos aspectos definidos en los objetivos del proyecto.

$$\text{Maximize } w = \sum_{i=1} \sum_{s=1} S_{i,s} \cdot U_{i,s} \quad (2.1)$$

$$\text{Subject to } \sum_{i=1}^m U_{i,s} \geq 1, \forall s \in R \quad (2.2)$$

$$\sum_{j=1}^{ts} G_s \leq CC \quad (2.3)$$

$$0,95 \leq \frac{\sum_{t=1}^{ts} PC_t}{TT} \leq 1 \quad (2.4)$$

$$\frac{\sum_{t=1}^{ts} E_{t,f}}{TT} = 1 \quad (2.5)$$

$$0,95 \leq C_s \cdot 0.5 + E_s \cdot 0.5 \leq 1 \quad (2.6)$$

$$S_{i,s} \in \{0,1\}, i = 1, \dots, m; s = 1, \dots, ts$$

Utilizando los siguientes parámetros:

$G_s$ : Cantidad de gigabyte que requiere el sistema s.

$CC$ : Capacidad Cloud.

$DM_{t,f}$ : Cantidad de datos de muestra de la tabla t que son iguales a la fuente f.

$TDM_f$ : Total de datos de muestra de la fuente f.

$E_s$ : Exactitud de datos del sistema s.

Y las siguientes variables de decisión:

$S_{i,s}$ : { 1 Si el sistema s cuenta con al menos un usuario i ; 0 en otro caso }

$U_{i,s}$ : Cantidad de usuarios i que consumen el sistema s.

$C_s$ : Completitud de datos del sistema s.

$PC_t$ : Prueba de completitud asociada a la tabla t.

$TT$ : Total de tablas sometidas a pruebas de completitud.

Luego, para que el modelo se utilice de forma correcta, es necesario considerar las siguientes restricciones:

- (2.2): Asegura que para cada sistema exista al menos un usuario migrado.
- (2.3): Asegura que la cantidad de gigabytes que utiliza el sistema sea menor que la capacidad disponible de la nube.
- (2.4): Establece un límite inferior y superior para aceptar una prueba de completitud de los datos, en donde se analiza la cantidad de datos no nulos versus la cantidad total de datos existentes en la tabla, tomando en cuenta la totalidad de tablas que se analizan.
- (2.5): Asegura que una muestra de datos de una tabla sea exactamente igual a una muestra de datos de la fuente original, tomando en cuenta la totalidad de tablas que se analizan.
- (2.6): Establece un límite inferior y superior para aceptar la calidad de los datos

### 2.3 Medidas de desempeño

Con la finalidad de cuantificar el desempeño asociado a la implementación del proyecto, se establecen tres medidas que buscan representar el impacto generado por cada uno de los objetivos.

En primer lugar, para lograr un estándar de calidad en los datos, se debe cuantificar la cantidad de pruebas de aceptación realizadas. Para eso, se considera el porcentaje de calidad como la relación entre las pruebas realizadas satisfactoriamente versus el total de pruebas que se deben realizar. En donde las pruebas completadas corresponden a una ponderación del nivel de completitud y exactitud de los datos. El indicador de calidad corresponde a:

$$\% \text{ Calidad} = \frac{\text{Pruebas Completadas}}{\text{Total de pruebas}} \cdot 100$$

En segundo lugar, dado que se desea medir el progreso del plan de migración, se considera una relación entre la cantidad de consultas que se realizan al sistema versus la cantidad de consultas que se estima que deben estar conectadas a este. Para cuantificar esta situación se considera la siguiente relación:

$$\% \text{ Conexión} = \frac{\text{Cantidad de consultas conectadas}}{\text{Total de consultas}} \cdot 100$$

Por último, se considera una medida de desempeño que busca reflejar que tan bien se ha cumplido el proceso completo, desde la data hasta la migración del usuario. Por esta razón, es necesario aclarar que un usuario migrado hace referencia a la disponibilidad de datos de calidad y una vista o consulta que permite a este extraer información. De esta forma, se tiene la relación entre la cantidad de usuarios migrados con éxito y la cantidad de usuarios que se busca migrar.

$$\% \text{ Migración} = \frac{\text{Usuarios migrados}}{\text{Total de usuarios}} \cdot 100$$

### 3. Estado del arte

La adopción de técnicas efectivas en el área de manejo de datos, utilizando distintas herramientas tecnológicas, han dado paso a un nuevo mundo al momento de desarrollar proyectos, pero al mismo tiempo, ha obligado a ciertas instituciones o entidades a desarrollar otro tipo de estrategias o herramientas para lograr sostener estos modelos de gestión eficiente de grandes volúmenes de información.

Por el lado de migración de datos, distintos usuarios implementan estrategias que implican almacenar datos en la nube, gracias a las distintas posibilidades que entrega, tales como la flexibilidad y escalabilidad de los datos almacenados. Con estas características, estas plataformas ofrecen una solución que permite a las empresas mover grandes cantidades de datos de manera eficiente para poder

realizar sus operaciones. Luego, el proceso de extracción, transformación y carga (ETL), se ha posicionado como una de las técnicas fundamentales para la migración y aumento en la calidad de los datos.

La nube se describe como una red que facilita la distribución de diversas capacidades asociadas al procesamiento, a aplicaciones y almacenamiento entre diversos sistemas que se encuentran ubicados de forma remota en todo el mundo. En el contexto de la computación en la nube, los recursos de tecnología de la información se utilizan y liberan según las necesidades mediante el empleo de Internet. Optar por servicios (IaaS, PaaS, SaaS, DaaS, entre otros) ofrecidos por proveedores de servicios en la nube, con un modelo de pago en base al uso, se presenta como una opción bastante favorable tanto para organizaciones como también para usuarios común y corrientes. Dado que la oferta de proveedores en el mercado actual es abundante, se tomaron en cuenta Amazon, Microsoft y Google. Ahora bien, dos protagonistas destacados en este panorama son Amazon Web Services (AWS) y Google Cloud Platform (GCP). AWS, lanzado en 2006 por Amazon, se destaca como uno de los proveedores de servicios en la nube más grande y diversificado, ofreciendo todo tipo de productos en donde se puede mencionar el almacenamiento inteligencia artificial y el de análisis de datos. Dentro de los productos que ofrece, se incluyen servicios como EC2 para la creación de máquinas virtuales y S3 para el almacenamiento de objetos. Por otro lado, GCP, lanzado en 2011 por Google, ha tomado mayor protagonismo gracias a su enfoque en la innovación y la inteligencia artificial. Sus servicios incluyen Compute Engine y App Engine para la infraestructura, destacando el uso de BigQuery para llevar a cabo análisis de datos a gran escala y TensorFlow para realizar aprendizaje automático. AWS destaca por su madurez y gran presencia a nivel global, mientras que GCP se posiciona con énfasis en la innovación y la integración de tecnologías como TensorFlow y Kubernetes, además de la auto escalabilidad que ofrece para los sistemas que utilizan sus clientes.

En cuanto a los servicios, a lo largo del tiempo han surgido diversas empresas que proporcionan una amplia variedad de servicios, brindando a los usuarios la posibilidad de elegir productos que se adapten mejor a sus necesidades

individuales. Entre estos servicios, se destacan el Software as a Service (SaaS), el Platform as a Service (PaaS) y el Infrastructure as a Service (IaaS). El primero ofrece aplicaciones de terceros, el segundo proporciona herramientas para el desarrollo, y el tercero suministra una infraestructura similar a la de un hardware tradicional. Estos avances han permitido una mayor flexibilidad y personalización, facilitando a los usuarios la elección de soluciones más alineadas con sus requerimientos específicos.

Luego, en cualquier proyecto de migración y gobierno de data, es sumamente relevante tomar en cuenta la seguridad de los datos y el cumplimiento normativo, ya que, de esta forma, los distintos datos no están en manos de cualquier usuario.

En el complejo ecosistema de la gestión de datos, los almacenes de datos desempeñan un papel esencial al actuar como cimientos sólidos para la construcción y optimización de estrategias empresariales basadas en la información. A medida que las organizaciones se sumergen en la era actual, donde la transformación digital es el principal foco, la diversidad de almacenes de datos se presenta como un abanico de posibilidades, cada uno diseñado para abordar necesidades específicas en relación con la recolección, almacenamiento y análisis de datos. Tomando en cuenta lo anterior, dentro del universo de almacenes de datos se tiene el Data Warehouse, Data Lake y Data Mart. A continuación, se entregan más características asociadas a cada uno de estos repositorios.

- a. Data Warehouse consiste en un repositorio centralizado de datos procesados y estructurados, que está diseñado principalmente para el análisis y toma de decisiones relacionadas con un tema en particular, por ejemplo, manejo de inventario. Los datos que permite almacenar este repositorio deben someterse a un proceso de ETL antes de ser pasar por el proceso de ingesta en el almacén de datos, razón por la cual, requiere un diseño en particular en cuanto a los esquemas. Si bien, Data Warehouse, permite procesar grandes cantidades de consultas analíticas gracias a su tiempo reducido de

respuesta, el proceso de carga de datos es bastante prolongado, provocando que cualquier cambio o adaptación sea más lento. Además, gracias a su sistema de restricción de acceso que busca segmentar según el tópico de la información, ofrece un alto nivel de seguridad garantizando confidencialidad e integridad de la información presente en el almacén.

- b. Data Lake es un repositorio centralizado que se destaca por su capacidad para almacenar datos puros (no procesados) en distintos formatos, ya sean estructurados, semiestructurados o no estructurados. Gracias a su diseño, no requiere que los datos sean procesados previamente antes de su ingesta, lo que permite proporcionar flexibilidad para almacenar y realizar distintas transformaciones con los distintos datos provenientes de diferentes fuentes. En relación con la administración de un data lake, se tiene una facilidad de gestión y provisión de metadatos exhaustivos en donde cada dato cuenta con información relevante como su fuente, formato y permisos, permitiendo de esta forma brindar un mayor contexto en cuanto al origen del dato. En relación con la seguridad y confidencialidad de los datos, los data lakes presentan un sistema de autenticación y control de permisos, generando que la arquitectura de los datos sea segura y fiable. Por otro lado, un data lake es altamente escalable, por lo tanto, su rendimiento no se ve afectado a medida que aumenta la cantidad de datos gracias a su capacidad de adaptación y respuesta.
- c. Data Mart es una subcategoría que está presente en un Data Warehouse, capaz de contener datos procesados y organizados en función de un tema o unidad del negocio en particular. Comparte múltiples características de los data Warehouse, por ejemplo, los niveles de seguridad y diseño de esquemas para garantizar que la carga de datos sea coherente y esté organizada correctamente. Por otro lado, los usuarios que consumen la data presente en un data

mart, solo tienen la posibilidad de visualizar los datos, provocando un impedimento en caso de que se desee transformar un tipo de datos.

En el ámbito de los proyectos de gestión de datos, diversas arquitecturas han emergido para abordar la complejidad inherente a las operaciones contemporáneas. Un estudio pertinente, identificado en el marco de la investigación conducida por (A. A. Munshi and Y. A. -R. I. Mohamed, in IEEE Access, vol. 6, pp. 40463-40471, 2018) destaca la importancia de la arquitectura lambda en tales contextos. La arquitectura lambda constituye un modelo de implementación que las organizaciones adoptan para llevar a cabo un proceso sistemático y automatizado de movimiento eficiente y confiable, transformación y gestión de datos, permitiendo así un acceso expedito a la información. Este enfoque ha adquirido una posición destacada en las herramientas de tecnología de la información y en los desarrollos a medida de las organizaciones. Ahora bien, la arquitectura Lambda, utiliza como fundamento el cálculo de funciones arbitrarias sobre conjuntos de datos que están distribuidos en tiempo real, buscando generar una combinación de capacidades de procesamiento por lote y en tiempo real para buscar un equilibrio en el rendimiento, latencia y una tolerancia a los posibles fallos que se pueden experimentar con los datos. A pesar de que existen múltiples herramientas que permiten realizar este tipo de tareas, esta arquitectura brinda una solución bajo el concepto de una estructura de tres capas: una capa de lotes, una capa de velocidad y una capa de servicio. Este enfoque permite abordar el desafío de calcular funciones arbitrarias de manera paralela en datos distribuidos en tiempo real. La convergencia de estos procesos sistemáticos y automatizados representa una solución versátil y eficiente para satisfacer las exigencias de la gestión de datos en tiempo real y por lotes, en un escenario caracterizado por su dinamismo y cambio constante.

---

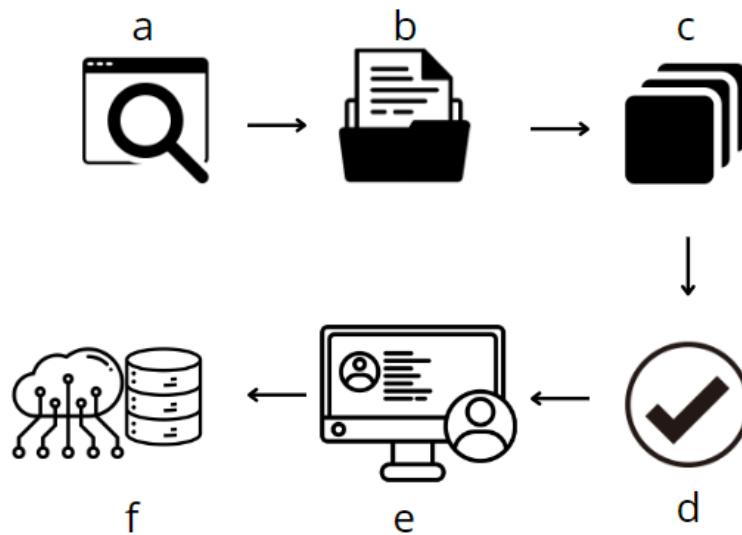
## 4. Metodología

Luego de las investigaciones realizadas en base a las distintas tecnologías y enfoques que se están utilizando en la industria, se decide que adoptar el sistema Google Cloud Platform (GCP) para implementar un almacén de datos. Esta decisión, tiene su fundamento en la destacada conveniencia que se puede llegar a experimentar con el servicio que ofrece Google, en donde se proporciona a sus clientes herramientas con múltiples funcionalidades que permiten gestionar de forma eficiente los datos. Luego, dado que data lake es un repositorio de almacenamiento que permite una gran flexibilidad en cuanto a los datos, se considera que se adapta de mejor forma a las necesidades no solo del proyecto, sino que también de la empresa.

Para lograr correctamente la implementación del Data Lake, es necesario contar con el acceso a los datos del área de reabastecimiento y utilizar Bigquery, que corresponde a un servicio de nube de Google Cloud Platform.

Para que la data se pueda consumir directamente del repositorio, se requiere dividir el proceso en dos grandes etapas, los sistemas y las vistas de usuarios. La primera etapa, hace referencia a los pasos que deben ser realizados para traspasar la data disponible desde las distintas fuentes no gobernadas hacia el repositorio de datos y, la segunda, está relacionada al proceso que se debe llevar a cabo con los usuarios pertenecientes al negocio para que puedan consumir la data disponible en el repositorio.





Esquema 2: “Diagrama de flujo de la metodología para desarrollar el proyecto”. (Elaboración propia)

- a. Valorización de iniciativa e identificación de datos a migrar.
- b. Documentación de tópicos.
- c. Modelamiento de datos.
- d. Ingesta, revisión de eficiencia y pruebas de calidad.
- e. Comprensión y adaptación de consultas.
- f. Generación de vistas personalizadas.

La etapa (a) consiste en definir un acuerdo con el negocio en donde se establecen los principales requerimientos para migrar el sistema que utiliza. Para ello, se debe valorizar la iniciativa cuya finalidad consiste en dar visibilidad de la importancia de la migración que se quiere realizar. A continuación, se entregan la información que se necesita para valorizar la iniciativa.

| Tema                              | Tipo de información que se espera   |
|-----------------------------------|---|
| Macro iniciativa                  | Nombre general de la iniciativa.  |
| Dominio                           | Área de la empresa involucrada.   |
| Subdominio                        | Subárea de la empresa involucrada.  |
| Sponsor/ Patrocinador             | Persona a cargo.  |
| Iniciativa                        | Nombre de la iniciativa.  |
| Área de datos                     | Departamento que gestiona los datos.  |
| Prioridad                         | Categorización de la prioridad de la iniciativa.                            |
| Caso de Negocio                   | A que área impacta la iniciativa.   |
| Descripción                       | Entregar más detalles asociados a la iniciativa.                            |
| Justificación del caso de negocio | Proporcionar detalles que expliquen por qué la iniciativa es relevante.     |
| Alcance                           | Especificar la cantidad de tablas que se buscan migrar desde el servidor.   |
| Impacto estratégico               | Impacto que generará en la organización la implementación de la iniciativa. |
| Valorización Cuantitativa         | Cuantificar las áreas u operaciones que reciben impacto.                    |
| Valorización Cualitativa          | Describir aspectos no cuantificables  |

*Tabla 1: “Aspectos a considerar en una valoración de iniciativa”. (Elaboración propia)*

Una vez definida la valorización de la iniciativa, se debe definir las tablas que serán migradas para que el negocio pueda consumir la data directamente desde el Data Lake, en donde se busca comprender de forma detallada la estructura, el tipo, el contenido y las relaciones entre los datos.

Luego, para la etapa (b), se procederá a generar un documento que abarca metadatos específicos y detalles relacionados con la normalización de los datos. Este paso no solo facilitará una mayor comprensión de los datos, sino que también, entregará la posibilidad de generar un diccionario de datos detallado que funcionará como una guía técnica y conceptual para los usuarios que desean consumir los datos asociados a las tablas existentes dentro de este diccionario. El diccionario considera lo siguiente:

| Tema                                 | Información por completar  |
|--------------------------------------|--|
| Nombre de la tabla en el origen      | Indicar nombre de la tabla en la fuente de origen.   |
| Nombre lógico de la tabla            | Indicar nombre lógico de la tabla, debe estar en inglés, sin abreviaciones, permite espacios, debe ser entendido por el dominio de negocio.  |
| Descripción de la tabla              | Describir la tabla en inglés.  |
| Nombre del capo en el origen         | Indicar el nombre del campo en el origen.  |
| Nombre lógico del campo en el origen | Indicar nombre lógico del campo, debe estar en inglés, sin abreviaciones, permite espacios, único por tabla, y se escribe en singular.   |
| Descripción del campo                | Descripción del campo en inglés.   |
| Tipo de dato en el origen            | Indicar el tipo de dato en el origen, por ejemplo: int, decimal, date, varchar, etc.   |
| Tipo de dato en el destino           | Sugerir el tipo de dato en BigQuery.   |
| Nulo                                 | Indicar SI, cuando el campo puede tomar valores nulos, NO cuando no puede tomar valores nulos.   |
| Primary Key                          | Indicar SI cuando el campo es Primary Key, si no lo es, marcar NO.   |
| Foreign Key                          | Indicar SI cuando el campo es Foreign Key, si no lo es, marcar NO.<br>Se indicará NO DEFINIDO, cuando se reconozca al campo como Foreign Key, pero se desconozca el nombre de la tabla y campo FK. |
| Nombre de la tabla Foreign Key       | Si el campo es Foreign Key, indicar el nombre de la tabla relacionada.   |
| Nombre del campo Foreign Key         | Si el campo es Foreign Key, indicar el nombre del campo en la tabla relacionada.   |
| Clasificación de los datos           | Indicar sensibilidad del dato: HS, SE, NS.   |
| Comentarios adicionales              | Indicar comentarios relevantes para el proceso de modelado.  |

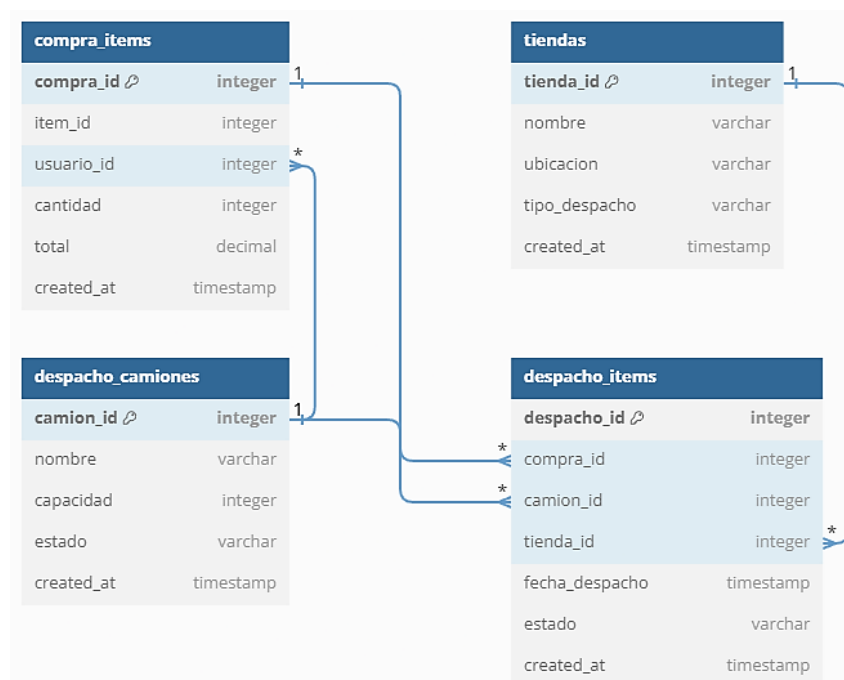
Tabla 2: “Esquema diccionario de datos”. (Elaboración propia)

Además de elaborar el diccionario de datos, es necesario crear un archivo de texto que contenga el *Lenguaje de Definición de Datos o Data Definition Language (DDL)*, permitiendo especificar la forma en que se organizarán y almacenarán los datos, considerando también las relaciones existentes con otros

elementos de la base de datos. A continuación, se entrega un ejemplo de cómo sería un DDL:

```
CREATE TABLE 'nombre_proyecto.nombre_bbdd_nombre_tabla' (  
  
    nombre_campo1 tipo_dato NULL,  
  
    nombre_campo2 tipo_dato NULL  
  
);
```

Contando con esta información, se inicia la etapa (c), en donde junto con un modelador e ingeniero de datos, se elabora un modelo que busca emplear técnicas avanzadas tales como, la indexación o particionamiento, para optimizar la velocidad de recuperación de los datos. Luego, utilizando distintas herramientas de modelamiento, se genera un diagrama relacional tal como se puede apreciar en la *Figura 1*.



*Figura 1: "Ejemplo de modelo de datos" (Elaboración propia)*

Posteriormente, para la etapa (d), el ingeniero de datos asignado al proyecto debe hacerse cargo de la ingesta y habilitación de ambiente para poder realizar las pruebas de completitud y exactitud. Estas pruebas, buscan medir la calidad de los datos en base a una serie de muestras que permiten comparar los datos provenientes del proceso de ingesta versus los datos provenientes de la fuente original. Luego, la persona a cargo de proceso de testeo, utilizando un archivo asociado a una fecha en particular extraído desde la fuente de origen, se llevan a cabo las pruebas de aceptación, en donde se verifica que el esquema de los datos sea correcto y que los datos presentes en PRU sean consistentes con los datos provenientes de la fuente de origen. Este proceso es sumamente importante ya que, en caso de que la data sea inconsistente con el archivo, se debe realizar una revisión y levantamiento de información asociada a los problemas presentes para que sean corregidos. Luego, cuando la data presente en PRU es consistente con la data presente en el archivo, se dice que la data cumple con los estándares de calidad de Walmart. Una vez que la data cumple con los estándares de calidad de Walmart, son llevados a PRO, repositorio en donde se encuentran los datos gobernados.

Por otro lado, para la etapa (e) de integración de las consultas, resulta esencial comprender y adaptar de forma correcta las queries que utilizan los distintos usuarios que consumen la data del sistema. Luego, con la finalidad de agilizar el proceso de transformación de las consultas, se generará un código en Python que utilizará como referencia el diccionario de datos definido en la etapa (b), permitiendo de esta forma adaptar el campo proveniente de la fuente original al campo nuevo que se define bajo los estándares para los datos del Data Lake.

Finalmente, dado que, al migrar un sistema, se busca deprecia la fuente original, todo reporte generado con la información proveniente de esta fuente, desaparece. Por esta razón, la etapa (f) es sumamente relevante en el proceso ya que las vistas permiten generar trazabilidad operacional, pero para esto, se necesita generar una documentación especial por cada vista creada. Para ello, la

nueva documentación especial para la vista en cuestión debe considerar los siguientes puntos:

| Tema              | Información                       |
|-------------------|-----------------------------------|
| Proyecto BigQuery | Indicar nombre del proyecto en BQ |
| Data set BigQuery | Indicar nombre del Data set en BQ |
| Tabla BigQuery    | Indicar nombre de la tabla en BQ  |
| Campo BigQuery    | Indicar nombre del campo en BQ    |

Tabla 3: “Esquema de diccionario de datos para input de la vista”. (Elaboración propia)

| Tema        | Información por completar  |
|-------------|--|
| Name field  | Indicar nombre del campo en BQ. Si es un campo nuevo, indicar en comentarios el cálculo de cómo se obtiene |
| Description | Indicar la descripción del campo   |
| Sensitivity | Indicar la sensibilidad del campo: NS, SE, o HS  |
| Data type   | Indicar el tipo de dato en BigQuery  |

Tabla 4: “Esquema de diccionario de datos para output de la vista”. (Elaboración propia)

Definiendo los esquemas de entrada y salida de la vista, esta se puede habilitar para el uso de los usuarios que necesitan acceder a la data en cuestión.

## 5. Implementación

Tal como se mencionó anteriormente, se requiere dos fases para llevar a cabo este proyecto, en donde se considerará:

- I. Fase recopilación: Levantamiento de caso de negocios junto con mapeo de tablas necesarias con sus respectivas fuentes. (Etapa a)
- II. Fase de modelo: Diccionario y creación de modelo de datos. (Etapa b y c).
- III. Fase de muestreo: Calidad de ingesta. (Etapa d).
- IV. Fase de conexión: Habilitar las consultas del usuario. (Etapa e y f).

La fase de recopilación comenzó con la definición y valorización de la iniciativa de migrar el sistema de reabastecimiento junto con la identificación de las tablas asociadas al sistema. A continuación, se indica la información asociada a fase:

| Tema                              | Información   |
|-----------------------------------|---|
| Macro iniciativa                  | Nombre general de la iniciativa   |
| Dominio                           | Área de la empresa  |
| Subdominio                        | Replenishment   |
| Sponsor                           | Persona a cargo   |
| Iniciativa                        | Nombre de la iniciativa   |
| Área de datos                     | Data Experience & Enablement  |
| Prioridad                         | Categorización de la prioridad de la iniciativa   |
| Caso de Negocio                   | A que área impacta la iniciativa  |
| Descripción                       | Entregar más detalles asociados a la iniciativa.  |
| Justificación del caso de negocio | <ul style="list-style-type: none"> <li>- Para democratizar el acceso a la información.</li> <li>- Habilitar herramientas de gestión.</li> <li>- Entender el nivel de calidad de la data.</li> <li>- Tener soporte en Chile.</li> <li>- Garantizar la disponibilidad y linaje de la data.</li> </ul>                     |
| Alcance                           | Migración de 17 tablas de GRS desde Oracle.   |
| Impacto estratégico               | <ul style="list-style-type: none"> <li>- Growth.</li> <li>- Omni CVP Availability.</li> <li>- Omni CVP Assortment.</li> <li>- Associate Experience.</li> <li>- Customer Growth.</li> <li>- Customer Trust.</li> <li>- Data Venture.</li> <li>- Availability of Critical Data.</li> </ul>                                |
| Valorización Cuantitativa         | <ul style="list-style-type: none"> <li>- Ahorro de horas del soporte actual de equipos y de actualización.</li> <li>- Disponibilidad y calidad.</li> <li>- Cantidad de usuarios y reportes.</li> <li>- Reducción de quiebre de stock.</li> <li>- Aumento en ventas.</li> <li>- Mejoras en fill rate instock.</li> </ul> |
| Valorización Cualitativa          | <ul style="list-style-type: none"> <li>- Habilitación de analítica para el flujo de reabastecimiento.</li> <li>- Mejora en la experiencia de usuario.</li> <li>- Aumento en la productividad.</li> <li>- Gobierno de KPI's</li> </ul>   |

Tabla 5: "Valorización iniciativa Migración GRS". (Elaboración propia)

Luego, en conjunto con el negocio, se identificaron 74 tablas que alimentan al sistema con más de 157 usuarios que consultan la data de este constantemente. Para lograr un mejor uso de los recursos, se enumeró a los distintos usuarios utilizando el criterio de cuán importante era la data que utiliza día a día. De esta forma, se priorizó la migración de 97 usuarios, mientras que los 60 restantes fueron re - clasificados ya que no todos eran usuarios activos. Como consecuencia, la cantidad de conexiones a realizar consideraba un total de 27.380.582 en vez de considerar 51.556.982.

En la fase de documentación, se establecieron múltiples diccionarios que proporcionan una referencia integral a la estructura y terminología utilizada. Definiendo de forma correcta los campos, junto con sus respectivas conexiones con otras tablas, se generó el modelo de datos.

Para la fase de testeo de calidad, se buscó analizar la totalidad de los datos para realizar la prueba de completitud, mientras que, para la prueba de exactitud, se estableció un criterio para que la muestra sea representativa. De esta forma, en general, se tomó una fecha en particular para analizar el nivel de exactitud de un 35% de los datos.

Por último, en la fase de habilitación de consultas, se establecieron distintas estrategias para optimizar el rendimiento del sistema, ya que si el sistema recibe consultas cuyo procesamiento es muy alto, implica un mayor costo para la empresa.

## 6. Análisis de riesgo

Previo a la ejecución del proyecto, se analizó las distintas situaciones que pueden dificultar la correcta implementación de este, por lo que se define una serie de criterios que buscan clasificar los escenarios según probabilidad de ocurrencia e impacto que genera. Estos criterios se encuentran en la sección A del anexo.

Ahora bien, dentro de las distintas situaciones identificadas, es importante mencionar que no solo hay procesos internos, sino que también procesos externos, cuyas medidas de mitigación de riesgo no permiten una interacción mayor.



Luego, tal como se puede observar en la tabla 1, se da a conocer la importancia de gestionar correctamente los distintos procesos asociados a las distintas etapas del proyecto. Teniendo esto en cuenta, se pueden establecer las distintas medidas mencionadas para prevenir o disminuir el riesgo que implica la ocurrencia de cada evento.

| Riesgo | Descripción  | Plan de acción   |
|--------|--|--|
| R2     | No contar con documentación o mapeo de información a considerar en la migración.   | Expresar la necesidad de generar un estándar de despliegue en cuanto a la información de cada negocio, quedando registrado en Confluence.  |
| R3     | Variación en las tablas a migrar.  | Para evitar considerar menos campos de los cuales el negocio quiere migrar, se debe realizar una reunión o una mesa de consultas para dejar constancia de la totalidad de tablas y tópicos considerados. |
| R5     | Inconsistencia en los datos  | Definir el tipo de situación que genera la inconsistencia para destinar recursos con los equipos correspondientes.   |
| R4     | Interrupción del servicio de Google Cloud Platform. Se puede considerar como interrupción una caída o bien un periodo de latencia. | Solicitar soporte avanzado de la plataforma.   |
| R4     | Retraso por parte de internacional para aprobar modelos.   | Solicitar apoyo a cargos superiores para presionar la priorización de la aprobación de los modelos.  |
| R2     | Modificaciones en requisitos comerciales o modelos de negocio.   | Para evitar problemas de ajustes en la arquitectura del repositorio, se debe implementar una arquitectura modular.   |
| R4     | Vulnerabilidades asociadas a la seguridad, control de accesos y ataques cibernéticos.  | Monitorear constantemente las distintas actividades que se realizan.   |

*Tabla 5: “Eventos de riesgo” (Elaboración propia)*

## 7. Evaluación económica

Dado que hoy en día, el sistema de reabastecimiento sufre intermitencias en su funcionamiento, se lleva a cabo la presente evaluación económica, cuyo propósito consiste en analizar la viabilidad del proyecto en cuestión, tomando en cuenta metodologías sólidas que permiten determinar los parámetros correctos para cuantificar los indicadores financieros. A continuación, se explican los parámetros utilizados en el modelo CAPM para estimar la tasa de retorno que permitirá posteriormente calcular el Valor Actual Neto (VAN) asociado al proyecto.

Para estimar la tasa de retorno se considera la tasa libre de riesgo, el coeficiente beta y la rentabilidad esperada del mercado, parámetros que se combinan en la fórmula CAPM presentada a continuación.

$$r_i = r_{lr} + \beta[E(r_m) - r_{lr}]$$

$r_i$ : Tasa de descuento.

$r_{lr}$  : Tasa libre de riesgo.

$\beta$ : Coeficiente beta.

$E(r_m)$  : Rentabilidad esperada de mercado.

Los valores establecidos para estos parámetros se han determinado utilizando las siguientes consideraciones:

- La tasa libre de riesgo se ha determinado a través del rendimiento esperado a 5 años de los bonos del Banco Central.
- Dado que el coeficiente beta refleja el riesgo sistemático de la inversión, se debe considerar el valor asociado al riesgo de la industria minorista en Chile.
- En particular, para establecer la rentabilidad esperada del mercado de retail en Chile, se considera como proxy la rentabilidad del IPSA, que corresponde al Índice de Precios Selectivo de Acciones.

Considerando lo anterior, la tasa libre de riesgo es del 5,96%, el coeficiente beta toma un valor de 2,84 y la rentabilidad esperada del mercado corresponde a un 7,1%. De esta forma, el valor de la tasa de descuento para este proyecto es de 9,8%.

Luego, para combinar la tasa de retorno con los costos y beneficios de las distintas operaciones de la empresa, se debe realizar un flujo de caja, permitiendo de esta forma, conocer el VAN asociado al proyecto.

En términos monetarios, dado que el sistema sufre caídas, se estima que, en un mes, no opera durante 2 horas, generando un costo anual de \$9.600.000 monedas. Por otro lado, para lograr que el sistema este operativo nuevamente, es necesario contar con un equipo que permita entregar soporte, lo que se traduce en un costo anual de \$1.317.877 monedas.

Luego, al implementar el proyecto, se debe asumir un costo anual de \$8.660.000 monedas para sostener la digitalización y \$3.000.000 monedas para contar con el almacenamiento necesario disponible en la nube.

Teniendo en cuenta los distintos valores entregados, al implementar el proyecto, se debe asumir un costo de \$11.660.000 monedas, permitiendo de esta forma ahorrar \$10.917.877 monedas, dando la posibilidad de agregar nuevos proyectos al portafolio que se estima que pueden significar un beneficio de \$20.000.000 monedas al año.

Considerando estos valores y una proyección de 5 años del proyecto que se puede encontrar en la sección B del anexo, se obtiene un VAN positivo de \$2.160.325.428 monedas, lo que significa que el proyecto es rentable y es recomendable implementarlo.

## 8. Resultados

Los resultados obtenidos luego de desarrollar el proyecto reflejan el cumplimiento exitoso de uno de los tres objetivos específicos. Durante el desarrollo del proyecto se logró documentar más de 4.500 campos y realizar el 100% de las pruebas de calidad asociadas a las distintas tablas relacionadas con el sistema de reabastecimiento.

Por otro lado, para el caso de la migración de usuarios, si bien la transición no fue del todo exitosa, se logró migrar un porcentaje significativo de usuarios, permitiendo obtener algunos resultados asociados a la conexión de consultas. Considerando lo anterior, gracias a la migración exitosa de 12 usuarios, se logró conectar de manera exitosa 3.829.547 de consultas al Data Lake, que permitieron proporcionar respuestas más rápidas y eficientes para las distintas operaciones que se deben realizar en el reabastecimiento de la cadena de suministros. A pesar de no lograr el objetivo de conectar las consultas en su totalidad, el sistema de reabastecimiento redujo la cantidad de horas sufriendo intermitencias, cuyo último episodio registro una duración de 47 minutos aproximadamente, lo que significa una reducción de 1 hora y 12 minutos.

A continuación, se presenta un cuadro que resume las métricas analizadas considerando los valores iniciales, los porcentajes definidos en el proyecto, junto con los valores esperados y el valor final alcanzado.

| Métrica                | Valor inicial         | Porcentaje definido | Valor esperado | Valor alcanzado |
|------------------------|-----------------------|---------------------|----------------|-----------------|
| Calidad                | 60 pruebas            | 40%                 | 24             | 24              |
| Usuarios migrados      | 97 usuarios           | 20%                 | 19             | 12              |
| Cantidad de conexiones | 27.380.582 conexiones | 20%                 | 5.476.116      | 3.829.547       |

*Tabla 6: “Resumen de resultados” (Elaboración propia)*

---

## 9. Conclusiones

Luego de obtener los resultados del proyecto, se evidencia que la migración de un sistema no solo puede ser evaluada desde una perspectiva técnica, sino que también debe considerar una serie de estrategias para gestionar el cambio, particularmente para minimizar la resistencia por parte del usuario. Esto es importante mencionarlo ya que con los resultados obtenidos se puede evidenciar una mejora en cuanto a la continuidad operacional, por lo tanto, si la resistencia de los usuarios está dada por la preferencia que tienen estos por utilizar una plataforma y un sistema conocido, no hay impedimentos para que el plan de migración busque ser lo más cercano a lo que los usuarios conocen. De esta forma, si los usuarios presentan un nivel más alto de colaboración para implementar la migración, se producirá una disminución significativa en relación con los tiempos asociados a la primera etapa del proyecto, en donde se levanta el caso de negocio y se hace un mapeo de las tablas que son necesarias.

Por otro lado, a pesar de que la completitud de los datos se mantuvo dentro de los límites establecidos por el modelo, en más de una ocasión se debió dejar registros con un nivel de completitud cercano al límite inferior para que, posteriormente en un futuro, se analizara de manera más detallada por qué no se realizó de forma correcta la ingesta de datos. Esta situación genera algunas preocupaciones de cara al futuro ya que próximas migraciones podrían generar problemas en la experiencia de los usuarios que consultan la data. Por esta misma razón, es importante que, para futuros sistemas a migrar, se realice un reajuste del parámetro asociado a la completitud de los datos para evitar cualquier tipo de brecha que pueda impactar de manera negativa la calidad de la información existente en el repositorio de datos.

Finalmente, si bien no se logró el cumplimiento de todos los objetivos específicos, cabe destacar que se observó una reducción significativa del tiempo asociado a las intermitencias presentadas por el sistema, lo que permite establecer como supuesto que, a futuro, cuando se conecte la totalidad de usuarios y consultas, el sistema dejara de tener intermitencias, brindando un entorno confiable y estable para el consumo de datos.

## 10. Referencias

1. *About Walmart*. (s. f.). About Walmart. <https://corporate.walmart.com/about>  
*¿Qué es business intelligence o inteligencia de negocios?* (s. f.). Tableau.  
<https://www.tableau.com/es-es/learn/articles/business-intelligence>
2. WalmartChile. (2022, 10 febrero). *Nosotros – WalmartChile*. WalmartChile – Walmart Chile es uno de los principales actores en la industria de las ventas al por menor del país (retail). <https://www.walmartchile.cl/contenidos/nosotros/>
3. WalmartChile. (2023, 15 mayo). *Nuestro negocio – WalmartChile*. WalmartChile – Walmart Chile es uno de los principales actores en la industria de las ventas al por menor del país (retail). <https://www.walmartchile.cl/contenidos/nosotros/nuestronegocio/>
4. Cantalapiedra, C. G., & Soler, E. V. (2017). Evolución de los entornos Big Data y los retos para el Arquitecto de Datos. *Economía industrial*, (405), 21-31.
5. Devi, T., & Ganesan, R. (2015). Platform-as-a-Service (PaaS): model and security issues. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 15(1), 151-161.
6. Kumar, S., Kumar, S., Bajwa, M. S., & Sharma, H. (2019). A Comparative Study of Various Cloud Service Platforms. *Journal of Advanced Database Management and Systems*, 6(1), 7-12.
7. Jain, P., Kraft, P., Power, C., Das, T., Stoica, I., & Zaharia, M. (2023). *Analyzing and Comparing Lakehouse Storage Systems*. CIDR.
8. Munshi, A. A., & Mohamed, Y. A. -R. I. (2018). *Data Lake Lambda Architecture for Smart Grids Big Data Analytics*. *IEEE Access*, 6, 40463-40471. doi: 10.1109/ACCESS.2018.2858256.
9. Wong, B. I. Á. (2017). Los repositorios digitales para la conservación. *Un acercamiento a la preservación digital a largo plazo*. *Ciencias de la Información*, 48(2), 15-22.
10. Duperet Cabrera, E., Pérez Martínez, D. G., Cedeño Rodríguez, M. Y., Ramírez Mustelier, A., & Montoya Acosta, L. A. (2015). *Importancia de los repositorios para preservar y recuperar la información*. *Medisan*, 19(10), 1283-1290.

11. Gutiérrez Retamal, G. (2013). *Propuesta de implementación de un Data Warehouse para el área de soporte de información, Rabie SA.*
12. Gupta, B., Mittal, P., & Mufti, T. (2021, March). A review on amazon web service (aws), microsoft azure & google cloud platform (gcp) services. In Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development, ICIDSSD 2020, 27-28 February 2020, Jamia Hamdard, New Delhi, India.
13. Saraswat, M., & Tripathi, R. C. (2020, December). Cloud computing: Comparison and analysis of cloud service providers-AWs, Microsoft and Google. In *2020 9th international conference system modeling and advancement in research trends (SMART)* (pp. 281-285). IEEE.
14. Romero-Chuquital, A., & Melendres-Velasco, J. J. (2023). *Uso de data Warehouse para la toma de decisiones empresariales: una revisión literaria. Revista Científica de Sistemas e Informática*, 3(2), e543-e543.
15. Rodríguez-Aguilar, V., Magdaleno, S. L. C., Andrade, E. L. M., & Clemente, R. G. (2022). *Repositorio de Software Educativo: Una aproximación de desarrollo conceptual. EDMETIC*, 11(1), 7-7.
16. Hazelcast. (s.f.). What Is Lambda Architecture?  
<https://hazelcast.com/glossary/lambda-architecture/>
17. Morales Vargas, A., & Codina, L. (2019). Atributos de calidad web para repositorios de datos de investigación en universidades.
18. Ávila Barrientos, E. (2023). *La gobernanza de los datos de investigación en el contexto de su organización y sistematización.*

## 11. Anexo

### a. Anexo riesgo.

A continuación, se entregan los criterios establecidos para generar la matriz de riesgo. En primer lugar, se definen los niveles de aceptación de la situación cuyos valores se representan en la siguiente tabla.

| Rango   | Nivel       | Criterio   |
|---------|-------------|--|
| 1 – 4   | Aceptable   | No es necesario aplicar acciones, pero se recomienda mantener un control de estatus. |
| 5 – 9   | Adecuado    | Se debe tener en cuenta que acciones se pueden tomar para disminuir el riesgo.       |
| 10 – 16 | Tolerable   | Se debe considerar ejecutar acciones que permitan mejoras.                           |
| 17 – 25 | Intolerable | Se debe implementar un plan de contingencia.   |

*Tabla 2: “Nivel de aceptación”*

|                   | Insignificante | Menor      | Significativo | Mayor       | Severo      |
|-------------------|----------------|------------|---------------|-------------|-------------|
| Bastante probable | Medio 5        | Alto 10    | Muy alto 15   | Extremo 20  | Extremo 25  |
| Muy probable      | Medio 4        | Medio 8    | Alto 12       | Muy alto 16 | Extremo 20  |
| Probable          | Bajo 3         | Medio 6    | Medio 9       | Alto 12     | Muy alto 15 |
| Poco probable     | Muy bajo 3     | Bajo 4     | Medio 6       | Medio 8     | Alto 10     |
| Muy poco probable | Muy bajo 1     | Muy bajo 2 | Bajo 3        | Medio 4     | Medio 5     |

*Tabla 3: “Matriz de riesgo”*



| Probabilidad      | Impacto        | Riesgo      | Clasificación |
|-------------------|----------------|-------------|---------------|
| Muy poco probable | Insignificante | Muy bajo 1  | R1            |
| Muy poco probable | Menor          | Muy bajo 2  |               |
| Poco probable     | Insignificante | Muy bajo 3  |               |
| Probable          | Insignificante | Bajo 3      | R2            |
| Poco probable     | Menor          | Bajo 4      |               |
| Muy poco probable | Significativo  | Bajo 3      |               |
| Bastante probable | Insignificante | Medio 5     | R3            |
| Muy probable      | Insignificante | Medio 4     |               |
| Muy probable      | Menor          | Medio 8     |               |
| Probable          | Menor          | Medio 6     |               |
| Probable          | Significativo  | Medio 9     |               |
| Poco probable     | Significativo  | Medio 6     |               |
| Poco probable     | Mayor          | Medio 8     |               |
| Muy poco probable | Mayor          | Medio 4     |               |
| Muy poco probable | Severo         | Medio 5     |               |
| Bastante probable | Menor          | Alto 10     | R4            |
| Muy probable      | Significativo  | Alto 12     |               |
| Probable          | Mayor          | Alto 12     |               |
| Poco probable     | Severo         | Alto 10     |               |
| Bastante probable | Significativo  | Muy alto 15 | R5            |
| Muy probable      | Mayor          | Muy alto 16 |               |
| Probable          | Severo         | Muy alto 15 |               |
| Bastante probable | Mayor          | Extremo 20  | R6            |
| Bastante probable | Severo         | Extremo 25  |               |
| Muy probable      | Severo         | Extremo 20  |               |

*Tabla 4: “Clasificación de riesgo según probabilidad e impacto”*

## b. Anexo evaluación económica.

Con la finalidad de cuantificar de forma más certera si el proyecto conviene o no, se realiza un flujo de caja simple, considerando un horizonte de evaluación de 5 años, en donde se estimó los ingresos y costos asociados a la digitalización bajo ciertos coeficientes.

|                                       | Año 0 |             | Año 1 |             | Año 2 |             | Año 3 |             | Año 4 |             | Año 5 |             |
|---------------------------------------|-------|-------------|-------|-------------|-------|-------------|-------|-------------|-------|-------------|-------|-------------|
| (+) Ingresos por año                  | \$    | -           | \$    | 600.000.000 | \$    | 606.000.000 | \$    | 612.060.000 | \$    | 618.180.600 | \$    | 624.362.406 |
| (-) Costos de venta                   | \$    | -           | \$    | 9.600.000   | \$    | -           | \$    | -           | \$    | -           | \$    | -           |
| (-) Gastos de administración y ventas | \$    | -           | \$    | 31.801.970  | \$    | 34.861.362  | \$    | 38.347.498  | \$    | 42.182.247  | \$    | 46.400.472  |
|                                       | \$    | -           | \$    | -           | \$    | -           | \$    | -           | \$    | -           | \$    | -           |
| <b>Resultado operacional (A)</b>      | \$    | -           | \$    | 558.598.030 | \$    | 571.138.638 | \$    | 573.712.502 | \$    | 575.998.353 | \$    | 577.961.934 |
|                                       |       |             |       |             |       |             |       |             |       |             |       |             |
| (-) Inversión fija                    | \$    | 11.660.000  | \$    | -           | \$    | -           | \$    | -           | \$    | -           | \$    | -           |
|                                       |       |             |       |             |       |             |       |             |       |             |       |             |
| <b>Flujo de capitales</b>             | \$    | -11.660.000 | \$    | -           | \$    | -           | \$    | -           | \$    | -           | \$    | -           |
| Flujo de caja privado                 | \$    | -11.660.000 | \$    | 558.598.030 | \$    | 571.138.638 | \$    | 573.712.502 | \$    | 575.998.353 | \$    | 577.961.934 |

Figura 2: “Flujo de caja del proyecto” (Elaboración propia)