Diseño de un proceso ETL para la transformación de datos transaccionales de una pasarela de pagos, integrados en Oracle y herramientas de Business Intelligence (BI)

Autores
Daniela Hoyos Arango
Dalvis Hernando Serrato Rico
Maria Nohemi Correa Salinas

Director
Eliasib Naher Rivera Aya

Co-Director
Mauricio Garcés Restrepo

Universidad de Bogotá Jorge Tadeo Lozano
Facultad de Ciencias Naturales e Ingeniería
Especialización en Desarrollo de Bases de Datos

Bogotá - Colombia, Noviembre de 2025

Índice

Índice de figuras

Índice de tablas

Resumen

Las pasarelas de pago procesan millones de transacciones diarias, generando volúmenes críticos de datos que permanecen dispersos en sistemas transaccionales sin capacidad analítica. Este trabajo propone el diseño de una arquitectura ETL especializada para la pasarela de pagos Ipay, transformando datos transaccionales en información estratégica que soporte la estrategia de mercado y la toma de decisiones. La solución integra la extracción automatizada de archivos CSV mediante Python y Pandas; la transformación y validación según estándares ISO 8583 y PCI DSS; el almacenamiento en un modelo dimensional (esquema estrella) en Oracle Database; y la visualización mediante business intelligence. El modelo propuesto garantiza la disponibilidad, integridad y trazabilidad de la información financiera, habilitando el análisis del comportamiento transaccional que acelera la expansión de la pasarela en el mercado colombiano.

Abstract

Payment gateways process millions of transactions daily, generating critical volumesof data that remain scattered across transactional systems without analytical capabilities. Thiswork proposes the design of a specialized ETL architecture for the Ipay payment gateway, transforming transactional data into strategic information that supports market strategyand decision-making. The solution integrates: automated extraction of CSV filesusing Python and Pandas; transformation and validation according to ISO 8583 andPCI DSS standards; storage in a dimensional model (star schema) in Oracle Database; andvisualization using business intelligence. The proposed model guarantees the availability,integrity, and traceability of financial information, enabling the analysis of transactional behaviorthat accelerates the expansion of the gateway in the Colombian market.

Glosario

ACID Hace referencia al conjunto de cuatro propiedades claves que definen una transacción: Atomicidad, Consistencia, Aislamiento y Durabilidad. 8

atomicidad Cada instrucción en una transacción (para leer, escribir, actualizar o eliminar datos) se trata como una sola unidad. 7

botón de pago Es una integración en forma de botón que se realiza desde la página web o aplicación del comercio para la ejecución de sus pagos, aceptando todo tipo de tarjeta. 1

business intelligence Es un conjunto de procesos, tecnologías y herramientas utilizadas para ingerir datos y presentarlos en vistas fáciles de usar e interpretar, ayudando al análisis y a la toma de decisiones. v

control de acceso granular Hace referencia a la práctica de otorgar diferentes niveles de acceso a un recurso específico a usuarios específicos. 8

data warehousing Sistema de gestión de datos diseñado para habilitar y dar soporte a las tareas de inteligencia empresarial (BI), especialmente las analíticas. 2

dispersión de fondos Distribución de dinero desde una única transacción a dos cuentas comerciantes diferentes, comúnmente utilizado para las transacción de agencias que ofertan dentro de su portafolio los tiquetes aéreos. 1

esquema de copo de nieve Las tablas de dimensiones se dividen en subdimensiones más pequeñas para mantener los datos más organizados y detallados. 9

esquema estrella Su estructura se compone de una tabla principal llamada tabla de hechos, que contiene datos medibles como las ventas o los ingresos. A su alrededor están lastablas de dimensiones, que añaden detalles como nombres de productos, información sobre clientes o fechas. 9

estrategia de mercado Plan de acción de una empresa para llegar a los posibles consumidores y convertirlos de clientes potenciales a clientes reales de sus productos o servicios. v

ETL (extraer, transformar, cargar) es un proceso de integración de datos que combina, limpia y organiza los datos de varias fuentes en un conjunto de datos único y coherente para almacenarlos en un almacén de datos. 2

ley 1581 Sistema de gestión de datos diseñado para habilitar y dar soporte a las tareas de inteligencia empresarial (BI), especialmente las analíticas. 4

link de pagos Generación de un enlace único y seguro que permite autogestión del cliente final para el proceso de su pago online. 1

pasarela de pagos Sistema o tecnología que permite a las empresas aceptar pagos con diferentes tipos de tarjeta o medios de pagos. v

PCI DSS Sistema de gestión de datos diseñado para habilitar y dar soporte a las tareas de inteligencia empresarial (BI), especialmente las analíticas. 4

recurrencia de pagos Sistema de cobro automático en el que un cliente autoriza a una empresa a realizar cargos de forma periódica (mensual, anual, etc.) en su cuenta. Estos cobros son programados por los comercios mediante archivos planos y en lotes que pueden contener desde una tarjeta hasta veinte mil tarjetas. 1

rollback Es el proceso mediante el cual, se revierten los cambios realizados en un sistema o software para restaurarlo a un estado previo. 14

1. Introducción

Actualmente, el mercado financiero colombiano presenta un crecimiento exponencial en soluciones de pagos, lo que permite que la pasarela de pagos Ipay ofrezca sus soluciones financieras con múltiples funcionalidades sobre una misma conexión de servicio o proveedor a los comercios de grandes y pequeñas superficies.

Las funcionalidades con las que esta pasarela ha incursionado más en el mercado son: la recurrencia de pagos, la ejecución de pagos mediante dispersión de fondos (para agencias y aerolíneas), link de pagos, botón de pago o venta telefónica. Cada funcionalidad cubre necesidades de nichos o sectores de mercado diversos, lo que le brinda la oportunidad de expansión y, adicionalmente, contribuye al crecimiento del comercio electrónico en Colombia.

2. Descripción del Problema

En la actualidad, las empresas o entidades financieras que cuentan con una implementación de pasarela de pagos para el procesamiento o recaudo de sus operaciones financieras ejecutan diariamente millones de transacciones, generando con ello grandes volúmenes de datos transaccionales considerados como críticos para la operación de sus negocios. Además, estos datos muchas veces se encuentran almacenados en sistemas enfocados y optimizados para el procesamiento transaccional en tiempo real, mas no para el tratamiento orientado al análisis y la toma de decisiones estratégicas.

La causa raíz de esta problemática radica en la limitante que presentan los sistemas transaccionales de las pasarelas de pagos en cuanto a la necesidad de extraer valor analítico de la información registrada. Esto se debe a que los datos se encuentran dispersos en múltiples tablas relacionales con estructuras normalizadas que dificultan las consultas complejas, lo que implica que no se cuente con facilidad ni inmediatez en la información histórica necesaria para identificar patrones de comportamiento, tendencias de crecimiento y métricas de rendimiento operacional. Esta situación genera diversas consecuencias operacionales, debido a que los equipos de trabajo carecen de visibilidad inmediata sobre métricas clave como el comportamiento transaccional, tasas de transacciones exitosas, causales definidas de rechazo, tiempos óptimos de respuesta transaccional, patrones de uso y tendencias temporales fundamentales para la toma de decisiones.

La falta de un proceso ETL estructurado impide la consolidación eficiente de datos provenientes de diferentes fuentes del ecosistema de pagos, incluyendo logs de transacciones financieras, datos de autenticación y registros de eventos del sistema. La problemática es más evidente cuando se considera que las empresas o entidades financieras que cuentan con una pasarela de pagos operan en un entorno regulado, donde la trazabilidad, la auditabilidad y la precisión de la información financiera son requisitos no negociables. Los procesos ETL deben garantizar no solo la disponibilidad de información para el análisis de negocio, sino también el cumplimiento de los estándares de calidad, seguridad de datos y gobernanza.

Por lo expuesto anteriormente, existe una necesidad apremiante de proponer un desarrollo que contemple una solución integral donde se aborden estas limitaciones mediante procesos ETL específicamente optimizados para datos transaccionales de pasarelas de pagos, integrados con sistemas de data warehousing (DW) robustos y herramientas de Business Intelligence (BI) que proporcionen capacidades analíticas en tiempo cuasi real. La presente investigación se orienta hacia el desarrollo de una metodología y arquitectura técnica que permita transformar los datos transaccionales en información estratégica, accesible, confiable y oportuna.

3. Objetivos

3.1. Objetivo General

Diseñar un modelo ETL que facilite el proceso de transformación, análisis y almacenamiento de datos transaccionales provenientes de múltiples fuentes, con el fin de proporcionar data o análisis base relevante que ayude a la toma de decisiones estratégicas y al análisis del comportamiento transaccional.

3.2. Objetivos Específicos

- Diseñar la arquitectura del proceso ETL considerando las características específicas de los datos.
- Plantear mecanismos de gobernanza que permitan dar cumplimiento a la seguridad de los datos, fortaleciendo la auditabilidad, el cumplimiento normativo y la confiabilidad de la información financiera en el ecosistema de pagos.
- Definir los flujos de datos de las fuentes transaccionales a trabajar.
- Estructurar procesos de captura de datos o batch según los requerimientos del negocio.
- Plantear procesos de limpieza, validación y estandarización de los datos transaccionales.

4. Requerimientos

El planteamiento de un proceso de ETL y la integración con Oracle con herramientas de BI requiere una definición de requerimientos a nivel de negocio, funcionales y de calidad. Con la definición de los requerimientos se garantiza que se dé cumplimiento a los objetivos planteados anteriormente. A continuación, se describen los requerimientos:

- 4.1. Requerimientos de negocio.
  - Formular una herramienta de análisis de datos confiable para la toma de decisiones estratégicas.
  - Garantizar la disponibilidad de la información.
  - Dar cumplimiento a las regulaciones y estándares del sector financiero, como lo son: ley 1581, la ISO 8583 y PCI DSS.
  - Mejorar la eficiencia de los procesos internos, atribuyendo a la reducción de tiempos en la consolidación y análisis de información.
  - Aumentar la confianza en clientes y aliados con análisis precisos y consistentes.
- 4.2. Requerimientos funcionales.
  - El dashboard debe permitir la extracción automatizada de datos transaccionales desde las diferentes fuentes de la pasarela de pagos.
  - El sistema debe transformar los datos aplicando procesos de limpieza, validación y estandarización para garantizar su integridad.
  - El sistema debe almacenar los datos transformados en un repositorio de tipo data warehouse optimizado para consultas analíticas.
  - El dashboard debe generar tableros de control interactivos (BI) que representen los análisis de información clave para la organización.
- 4.3. Requerimientos de calidad.
  - La presentación de los datos debe garantizar claridad y precisión para una fácil interpretación.
  - Se debe garantizar un uso correcto de la información recopilada para dar cumplimiento a las normativas vigentes.
  - La propuesta debe contemplar la intención de optimizar la estructura del modelo ETL para lograr tiempos de respuesta eficientes, especialmente en consultas.
  - El sistema debe garantizar el acceso a la información con un rango de tiempo no mayor a un día de caída.
  - La solución debe permitir ajustes y mejoras en los procesos sin afectar la operación.

5. Estado del Arte

El presente trabajo aborda el diseño de un proceso ETL para la transformación de datos transaccionales de una pasarela de pagos, utilizando Oracle como plataforma de almacenamiento y herramientas de Business Intelligence para facilitar el análisis y visualización del comportamiento transaccional.

Saldarriaga (2024) [1], desarrolló un proceso ETL (Extract, Transform, Load) orientado a la optimización del almacenamiento y al análisis de datos históricos. El proyecto consistió en la implementación de un módulo desarrollado en Python y SQL, siguiendo la metodología ágil Scrumban, con el propósito de mejorar la gestión de los datos y apoyar los procesos de toma de decisiones en el área de analítica. Para ello, se realizó un diagnóstico del sistema ETL existente, el cual presentaba ineficiencias operativas y utilizaba tecnologías obsoletas, lo que afectaba directamente el rendimiento del procesamiento de datos y, por ende, la calidad del análisis de información.

La calidad de los datos en los procesos ETL es fundamental para garantizar la precisión y relevancia de los sistemas de Business Intelligence y Analytics (BIyA). Souibgui (2019) [2], examina las características de calidad del proceso ETL y presenta una visión general de los enfoques existentes para abordar problemas de calidad de datos, clasificándolos en dos perspectivas principales: enfoques centrados en el proceso y enfoques centrados en los datos. A través de un análisis comparativo de herramientas ETL comerciales.

El Payment Gateway (pasarela de pago) se ha convertido en un componente fundamental para el comercio electrónico, funcionando como un sistema que autoriza y procesa pagos entre compradores y vendedores de manera segura y eficiente. Supriyati (2019) [3], analiza la efectividad de las pasarelas de pago en el e-commerce, explicando su mecanismo de trabajo que incluye la conexión entre sitios web de ventas y entidades bancarias a través de un servidor seguro que verifica y procesa las transacciones en tiempo real.

Lapura (2018) [4], presenta el desarrollo e implementación de un data warehouse financiero universitario diseñado para transformar grandes volúmenes de datos transaccionales acumulados en información útil para la toma de decisiones informadas en la Universidad MSU-IIT en Filipinas. Siguiendo el enfoque de modelado de Kimball con una arquitectura bottom-up, los investigadores desarrollaron un data warehouse multidimensional con esquema estrella que integra dimensiones de tiempo, unidad financiera, cuenta y fondos, el cual es actualizado periódicamente mediante un proceso ETL (Extract-Transform-Load) utilizando Pentaho Data Integration desde la base de datos transaccional del sistema de gestión financiera (FMIS) de la universidad.

Short (2025) [5], presenta un análisis exhaustivo del rol de los frameworks ETL en la optimización de pipelines de datos dentro de sistemas de Business Intelligence, examinando cómo estos marcos han evolucionado para abordar los desafíos de eficiencia, escalabilidad y confiabilidad en el procesamiento de datos. A través de una revisión sistemática de literatura pre-2023, el estudio analiza la transición desde sistemas ETL tradicionales basados en procesamiento por lotes hacia frameworks modernos que incorporan procesamiento en tiempo real, arquitecturas basadas en la nube y automatización impulsada por inteligencia artificial y machine learning.

Encalada (2025) [6], presenta el diseño y validación de un marco de trabajo optimizado para la implementación de procesos ETL orientado a mejorar la eficiencia en el manejo de grandes volúmenes de datos en sistemas de Business Intelligence. A través de tecnologías modernas como Python (con librerías pandas, NumPy y pyodbc) para automatizar la extrac-

ción, transformación y carga de datos, junto con Power BI para la visualización interactiva de información estratégica, incorporando conceptos avanzados como integración semántica, procesamiento distribuido y arquitecturas parametrizables que incrementan la flexibilidad y adaptabilidad del sistema a diversos contextos empresariales.

Liu (2014) [7], realiza una optimización para flujos de datos ETL que aborda la complejidad y el alto costo en tiempo y recursos computacionales de estos procesos mediante técnicas de caché compartido y paralelización, con el objetivo de minimizar el tiempo de procesamiento y la huella de memoria en sistemas de data warehousing. El marco propuesto clasifica los componentes ETL en tres categorías basadas en sus propiedades de procesamiento de datos —componentes row-synchronized (procesamiento fila por fila), semi-block (requieren acumulación parcial), y block (requieren todos los datos antes de procesar)— y utiliza esta clasificación para particionar el flujo de datos (dataflow) en árboles de ejecución mediante un algoritmo de búsqueda en profundidad (DFS) que identifica puntos de partición en componentes block y semi-block.

Henao (2015) [8], presenta una revisión sistemática de literatura sobre Business Intelligence (BI) que examina artículos científicos publicados entre 1981 y 2014 en las revistas más influyentes de la disciplina según su factor de impacto y referencias cruzadas, con el objetivo de identificar los aportes más relevantes de la literatura académica sobre cómo la inteligencia de negocios impacta la gestión y toma de decisiones organizacionales en múltiples contextos empresariales.

Zapata (2017) [9], aborda la problemática crítica de que la administración de bodegas de datos (data warehouses) requiere procedimientos robustos para garantizar veracidad, integridad y centralización de información proveniente de fuentes heterogéneas, contexto en el cual los aplicativos especializados de ETL (Extract-Transform-Load) existentes tanto comerciales como open-source presentan limitaciones significativas incluyendo dificultades en parametrización para casos específicos, carencia de filtros de corrección adaptables a características particulares de diferentes dominios de datos, y costos de implementación prohibitivos especialmente para organizaciones medianas y pequeñas que limitan su adopción masiva.

Zapata (2019) [10], aborda la problemática de que PL/SQL (Procedural Language extension to Structured Query Language), lenguaje híbrido propietario desarrollado por Oracle Corporation que combina capacidades declarativas de lenguajes de consulta SQL con paradigma procedimental imperativo para desarrollo de aplicaciones empresariales complejas, tradicionalmente requiere codificación manual intensiva, propensa a errores de sintaxis y lógica, difícil de mantener y evolucionar, y con desconexión frecuente entre modelos conceptuales de diseño y código implementado que genera inconsistencias cuando requisitos cambian, situación agravada porque aunque existen propuestas académicas de generación automática de código SQL desde modelos visuales como diagramas entidad-relación y esquemas preconceptuales, así como trabajos de ingeniería inversa que generan productos como grafos de flujo de datos y modelos de arquitectura desde código SQL/PLSQL.

6. Marco Teórico

Fundamentos Técnicos para la Implementación de Procesos ETL en Sistemas de Pasarelas de Pago: Integración Oracle, Python y Business Intelligence

En el contexto actual de los sistemas financieros digitales, la gestión eficiente de datos transaccionales constituye un pilar fundamental para la toma de decisiones estratégicas y el cumplimiento de objetivos operacionales. La integración de pasarelas de pago con sistemas de almacenamiento estructurado y herramientas de Business Intelligence representa un desafío técnico que requiere procesos de extracción, transformación y carga de datos (ETL), garantizando no solo la centralización y disponibilidad de la información, sino también su integridad, trazabilidad y conformidad regulatoria.

El presente proyecto aborda el diseño de un proceso ETL para la transformación y análisis de datos transaccionales provenientes de una pasarela de pagos, como Ipay, integrados en Oracle SQL y herramientas de Business Intelligence. Este proceso tiene como objetivo centralizar información crítica del negocio, asegurar la calidad de los datos, facilitar el análisis mediante dashboards interactivos y cumplir con los estándares normativos del sector financiero colombiano.

6.1. Conceptos Fundamentales

6.1.1. ETL (Extract, Transform, Load)

El término ETL hace referencia a las tres fases esenciales para la migración y el tratamiento de datos en entornos empresariales:

Extracción (Extract): Se refiere a la recuperación de datos desde diversas fuentes operacionales. En el contexto de pasarelas de pago, esto incluye la captura de información transaccional desde bases de datos operacionales, archivos de log, sistemas de mensajería ISO 8583 y archivos CSV generados por procesos batch.

Transformación (Transform): Implica la limpieza, validación, enriquecimiento y normalización de los datos para adaptarlos al modelo de destino. En datos transaccionales financieros, esto incluye:

- Normalización de formatos de fecha y hora considerando zonas horarias.
- Validación de integridad de datos como transacciones, comercios y métodos de pago.
- Cálculo de métricas derivadas (tasas de aprobación y tiempos de respuesta promedio).
- Categorización de transacciones según el tipo de operación y el estado.
- Encriptación de datos.

Carga (Load): Es el proceso de inserción de datos transformados en el sistema de destino, en este caso Oracle SQL. La carga debe implementarse considerando:

- Estrategias de carga incremental vs. carga completa.
- Gestión de transacciones para garantizar la atomicidad.
- Mantenimiento de metadatos de ejecución para auditoría.

6.1.2. Pasarelas de Pago y el Estándar ISO 8583

Las pasarelas de pago son sistemas intermediarios que facilitan las transacciones electrónicas entre comercios, adquirentes, procesadores y emisores de tarjetas. Ipay, como pasarela de pagos, procesa transacciones bajo el estándar ISO 8583, que define la estructura de mensajes para el intercambio de información en transacciones con tarjetas.

6.1.3. Oracle SQL como Base de Datos

Oracle Database es un sistema de gestión de bases de datos relacional (RDBMS) ampliamente utilizado en entornos empresariales por su robustez, escalabilidad y características avanzadas. En el contexto de este proyecto, Oracle ofrece:

- Transaccionalidad ACID, garantizando la consistencia de los datos financieros.
- Stored Procedures y Packages, que permiten encapsular la lógica de negocio en el servidor de base de datos.
- Particionamiento de tablas, facilitando el manejo eficiente de grandes volúmenes de datos históricos.
- Vistas materializadas, que optimizan consultas agregadas frecuentes en dashboards BI.
- Seguridad avanzada, con encriptación de datos, auditoría detallada y control de acceso granular.

6.1.4. Python como Lenguaje para ETL

Python se ha consolidado como uno de los lenguajes más utilizados para implementar procesos ETL debido a su versatilidad, su amplio ecosistema de bibliotecas y su facilidad de mantenimiento. Las librerías clave para este proyecto incluyen:

- Pandas: Manipulación y análisis de datos estructurados mediante DataFrames.
- cx_Oracle/oracledb: Conectividad nativa con Oracle Database.
- csv: Lectura y escritura eficiente de archivos CSV.
- Logging: Generación de registros detallados de ejecución.

La capacidad de Python para procesar archivos CSV de gran tamaño de forma eficiente, combinada con su integración con Oracle, lo convierte en una opción ideal para este proyecto.

6.1.5. Modelado Dimensional: Tablas de Hechos y Dimensiones

El modelado dimensional organiza la información en dos tipos principales de estructuras: tablas de hechos y tablas de dimensiones, optimizando la ejecución de consultas analíticas y facilitando la comprensión de la información por parte de los usuarios.

Las tablas de hechos almacenan las métricas cuantitativas del negocio, representando eventos o transacciones medibles. En el contexto de una pasarela de pagos, la tabla de hechos

principal contiene las transacciones financieras con medidas como monto, tiempo de respuesta y contadores de transacciones.

Las tablas de dimensiones proporcionan el contexto descriptivo que permite analizar los hechos desde diferentes perspectivas. Incluyen atributos textuales y categóricos que responden a quién, qué, dónde, cuándo y cómo ocurrieron los eventos registrados en las tablas de hechos.

6.2. Datos Transaccionales en Pasarelas de Pago: Características y Desafíos

6.2.1. Volumen y Velocidad

Las pasarelas de pago pueden procesar miles de transacciones por minuto, generando grandes volúmenes de datos que deben ser almacenados, procesados y analizados. Este volumen creciente presenta desafíos en cuanto a:

- Escalabilidad de la solución ETL para manejar picos de transacciones.
- Optimización de tiempos de procesamiento para cumplir con la promesa de venta.
- Diseño de esquemas eficientes (esquema estrella o esquema de copo de nieve).
- Implementación de estrategias de archivado de datos históricos.

6.2.2. Variedad de Datos

Los datos transaccionales provienen de múltiples fuentes y tienen diferentes estructuras, como:

- Transacciones aprobadas, rechazadas, reversadas y anuladas.
- Datos de comercios, terminales, métodos de pago, fecha y hora.
- Logs de sistema y eventos de auditoría.
- Datos de autenticación.

El proceso ETL debe integrar estas fuentes heterogéneas manteniendo la trazabilidad y consistencia de la información.

6.2.3. Calidad y Consistencia

La calidad de los datos transaccionales es crítica para la operación del negocio y el cumplimiento regulatorio. Los principales desafíos incluyen:

- Duplicados: Transacciones registradas múltiples veces por reintentos o fallas de red.
- Inconsistencias entre sistemas de origen.
- Datos faltantes en campos obligatorios.
- Formatos incorrectos (fechas, códigos o montos).

6.3. Procesamiento de Archivos CSV en Python

6.3.1. Ventajas de CSV como Formato de Intercambio

Los archivos CSV son ampliamente utilizados para exportación de datos transaccionales debido a:

- Simplicidad.
- Universalidad.
- Eficiencia.
- Portabilidad.

6.3.2. Bibliotecas de Python para Procesamiento de CSV

Python incluye un módulo `csv` nativo adecuado para casos simples, pero limitado para procesamiento complejo. Por ello, la librería `pandas` resulta fundamental, gracias a su estructura DataFrame y sus capacidades avanzadas para manipular datos.

6.3.3. Optimización del Rendimiento

Es fundamental aplicar técnicas que mejoren la eficiencia, como lectura selectiva de columnas y la asignación explícita de tipos de datos.

6.4. Integración Python–Oracle para ETL

6.4.1. Conectividad con Oracle desde Python

Oracle proporciona la interfaz `python-oracledb` para la conexión de aplicaciones Python con Oracle. Esta interfaz permite ejecutar sentencias SQL o PL/SQL y trabajar con diversos tipos de datos, incluido JSON.

6.4.2. Estrategias de Carga de Datos

Carga Completa (Full Load): Reemplazo total de los datos en la tabla destino. Útil para dimensiones pequeñas o reprocesos completos.

Carga Incremental (Incremental Load): Solo se cargan registros nuevos o modificados. Requiere:

- Un campo de control.
- Lógica de identificación de cambios.
- Manejo de actualizaciones y eliminaciones.

Carga por Lotes (Batch Insert): Agrupación de múltiples inserts en una sola transacción para optimizar el rendimiento.

6.5. Apache Spark y PySpark para Procesamiento Distribuido

6.5.1. Introducción a Apache Spark

Apache Spark es un motor de procesamiento distribuido que ofrece velocidades de 10 a 100 veces superiores a Hadoop MapReduce. Sus características incluyen procesamiento en memoria, tolerancia a fallos, escalabilidad horizontal y APIs en múltiples lenguajes.

6.5.2. PySpark para ETL

PySpark combina la facilidad de Python con la potencia del procesamiento distribuido. En datos de pasarelas de pago permite:

- Procesamiento paralelo.
- Optimización Catalyst.
- Streaming en tiempo real.

6.5.3. Cuándo usar Spark vs. Python

Usar PySpark cuando:

- Volumen >50GB.
- Se requiere cluster.
- Procesamiento en tiempo real.

Usar Python/pandas cuando:

- Datos <20GB.
- Se requiere desarrollo rápido.
- Se implementan transformaciones altamente personalizadas.

6.6. Visualización de Datos

Los dashboards BI transforman datos en información accionable mediante visualizaciones interactivas.

Métricas Operacionales:

- Volumen de transacciones.
- Tasa de aprobación/rechazo.
- Tiempo promedio de respuesta.
- Disponibilidad del servicio.

Métricas Financieras:

- Valor total transaccionado.
- Top comercios.
- Distribución por método de pago.

Métricas de Calidad:

- Transacciones con error.
- Intentos de fraude.
- Reversas y contracargos.
- Conciliación.

Herramientas BI:

- Power BI.
- Tableau.
- QlikView/Qlik Sense.
- Oracle Analytics Cloud.

6.7. Evaluación de Herramientas de Business Intelligence Según Gartner

6.7.1. El Cuadrante Mágico de Gartner

El Cuadrante Mágico de Gartner (figura 1) es una metodología de investigación que analiza la posición competitiva de los proveedores de tecnología en mercados específicos.

Figura 1: Cuadrante Mágico de Gartner 2025

Gartner evalúa las plataformas BI en dos dimensiones: capacidad de ejecución y exhaustividad de la visión, clasificando a los proveedores en líderes, retadores, visionarios y actores de nicho.

6.7.2. Líderes del Cuadrante Mágico 2025

Los líderes reconocidos en 2025 son: Microsoft Power BI, Tableau (Salesforce), Qlik, Oracle, Google (Looker) y ThoughtSpot [11]. Estos proveedores se destacan por sus capacidades analíticas avanzadas, facilidad de uso y escalabilidad.

6.7.3. Criterios de Evaluación para Pasarelas de Pago

Para la selección de herramientas BI aplicadas a datos transaccionales, destacan los siguientes criterios:

1. Conectividad con Oracle.
2. Análisis en tiempo real.
3. Seguridad y cumplimiento normativo.
4. Escalabilidad.
5. Capacidades de análisis temporal.
6. Flexibilidad de despliegue.
7. Costo total de propiedad.

6.8. Conclusión del Marco Teórico

Este marco teórico establece los fundamentos conceptuales y técnicos para el diseño de un proceso ETL robusto, escalable y conforme a regulaciones para datos transaccionales de pasarelas de pago.

Considerando nuestro análisis de requerimientos, se toman las siguientes decisiones:

- **PYTHON + PANDAS:** Procesa 200M de transacciones/mes (<50GB). Permite desarrollo rápido y eficiente.
- **ORACLE SQL:** Cumple con PCI DSS y proporciona seguridad avanzada, control de acceso y vistas materializadas para dashboards.
- **PYSPARK (a futuro):** Se activará si el volumen crece 10x (>50GB).
- **ESQUEMA ESTRELLA:** Permite disponibilidad en un día y acelera consultas analíticas.

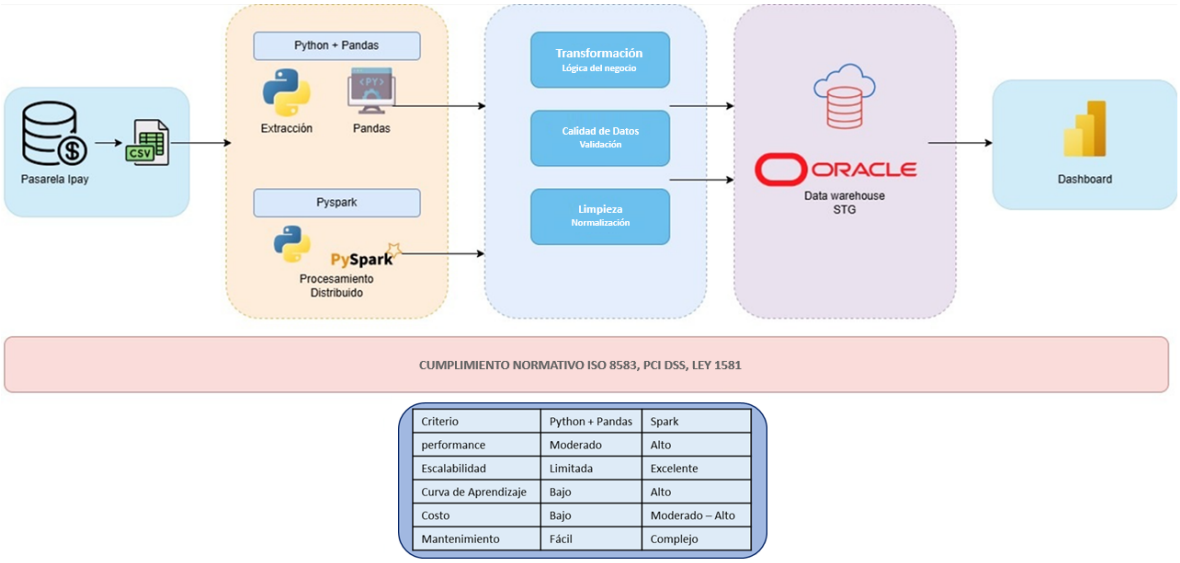
Estas decisiones se validan en el estado del arte (Sección 5), donde [1], [4] y [6] utilizan combinaciones similares (Python + Oracle y esquema estrella) para proyectos comparables.

7. Solución propuesta

7.1. Descripción general de la solución

La solución consiste en un sistema ETL automatizado que extrae datos transaccionales desde la pasarela de pagos Ipay, los transforma aplicando reglas de limpieza y validación, y los carga en una base de datos Oracle SQL para su posterior análisis mediante herramientas de Business Intelligence.

El sistema se estructura en cuatro componentes principales representados en la figura 2



Criterio	Python + Pandas	Spark
performance	Moderado	Alto
Escalabilidad	Limitada	Excelente
Curva de Aprendizaje	Bajo	Alto
Costo	Bajo	Moderado – Alto
Mantenimiento	Fácil	Complejo

Figura 2: Diagrama arquitectura ETL

Extracción de Datos: La pasarela de pagos (Ipay) genera archivos CSV con información transaccional que son procesados mediante Python con la biblioteca Pandas para volúmenes moderados, o PySpark para procesamiento distribuido cuando se manejan grandes volúmenes de datos. Los datos se extraen de forma estructurada considerando la naturaleza crítica de la información financiera.

Transformación de Datos: Aplica tres capas de procesamiento sobre los datos extraídos. Primero, una capa de transformación que estandariza formatos y estructuras. Segundo, una capa de calidad de datos que valida completitud, exactitud y consistencia. Tercero, una capa de limpieza que elimina duplicados, corrige valores nulos y normaliza tipos de datos. Este proceso asegura que solo información validada llegue al repositorio final.

Carga a Base de Datos: Los datos transformados se cargan en un Data Warehouse implementado en Oracle SQL con área de staging (STG) que permite validación previa antes de la carga definitiva. Se implementa carga incremental que evita duplicados mediante comparación de identificadores únicos, garantizando integridad transaccional mediante manejo de excepciones y rollback automático ante fallos.

Presentación y Análisis: Los datos almacenados se conectan con herramientas de Business Intelligence para generar dashboards interactivos que proporcionan visibilidad sobre métricas operacionales, tendencias transaccionales y análisis de comportamiento para la toma

de decisiones estratégicas.

El sistema opera bajo estricto cumplimiento normativo del estándar ISO 8583 para mensajería transaccional financiera, PCI DSS para seguridad de datos de tarjetas de pago y Ley 1581 de protección de datos personales, esto en cuanto a regulación colombiana, garantizando trazabilidad, confidencialidad e integridad de la información procesada.

La arquitectura permite escalar el procesamiento mediante PySpark cuando el volumen de transacciones aumenta, manteniendo la flexibilidad de usar Python + Pandas para operaciones de menor escala, logrando un balance óptimo entre performance, escalabilidad, costo y complejidad de mantenimiento según lo muestra la tabla comparativa de criterios técnicos incluida en el diseño.

7.2. Modelo conceptual

El modelo conceptual representa la estructura general del sistema ETL diseñado para la integración y análisis de datos transaccionales de la pasarela de pagos Ipay. Este modelo describe los principales componentes del sistema, los procesos que intervienen en la extracción, transformación y carga de los datos, y la forma en que se relacionan dentro del ecosistema analítico.

7.2.1 Componentes del Sistema

Fuente de Datos: La pasarela de pagos Ipay genera información transaccional que se exporta en archivos CSV estructurados. La información generada por la pasarela de pagos cumple con el estándar ISO 8583 para mensajería transaccional del sector financiero, garantizando interoperabilidad y trazabilidad de las operaciones.

Capa de Extracción: Esta capa se implementa mediante dos alternativas tecnológicas que se seleccionan según el volumen de datos a procesar:

Python + Pandas: Utilizado para el procesamiento de volúmenes moderados de datos. Ofrece desarrollo rápido, sintaxis simple, amplio ecosistema de bibliotecas y es adecuado para prototipado y operaciones donde los datos caben en memoria.

PySpark: Empleado para procesamiento distribuido de grandes volúmenes. Permite procesamiento en paralelo, escalabilidad horizontal y tolerancia a fallos mediante RDDs (Resilient Distributed Datasets).

Capa de Transformación:

- **Transformación:** Aplica conversiones de formato y estructura para estandarizar los datos según el modelo destino. Incluye normalización de fechas y horas, conversión de tipos de datos, extracción de información desde estructuras JSON anidadas y renombrado de columnas según convenciones establecidas.

- **Calidad de Datos:** Define las validaciones para garantizar que la información cumple con estándares de calidad. Verifica completitud de campos obligatorios, valida rangos permitidos para valores numéricos, asegura integridad referencial entre entidades relacionadas y detecta inconsistencias entre registros vinculados.

- **Limpieza:** Elimina o corrige datos problemáticos identificados durante las validaciones. Remueve registros duplicados mediante comparación de identificadores únicos, completa valores nulos con valores por defecto establecidos o los elimina según criticidad o reglas de negocio, corrige formatos inconsistentes y normaliza representaciones de datos categóricos.

Base de datos Oracle: Repositorio centralizado implementado en Oracle SQL que almacena los datos procesados. El Data Warehouse incluye un área de staging (STG) donde los datos transformados son validados antes de su incorporación definitiva. Esta arquitectura por capas permite aplicar controles adicionales de calidad, facilita la identificación de problemas antes de afectar datos productivos y proporciona capacidad de *rollback* ante detección de anomalías.

Los datos validados en el área de staging se cargan en un modelo dimensional optimizado para análisis, estructurado mediante un esquema estrella que centraliza la información analítica.

7.2.2 Modelo dimensional El modelo dimensional diseñado sigue el esquema estrella (star schema), el cual optimiza el rendimiento de consultas analíticas y facilita la comprensión del negocio. A continuación, se muestra el modelo lógico y relacional mapeado para el Data Warehouse, el cúal refleja los lineamientos del proyecto propuesto.

Modelo lógico En la figura 3 se ilustra la estructura conceptual del Data Warehouse, mostrando la tabla de hechos central y sus dimensiones asociadas. Este diseño permite análisis multidimensional eficiente desde diferentes perspectivas de negocio.

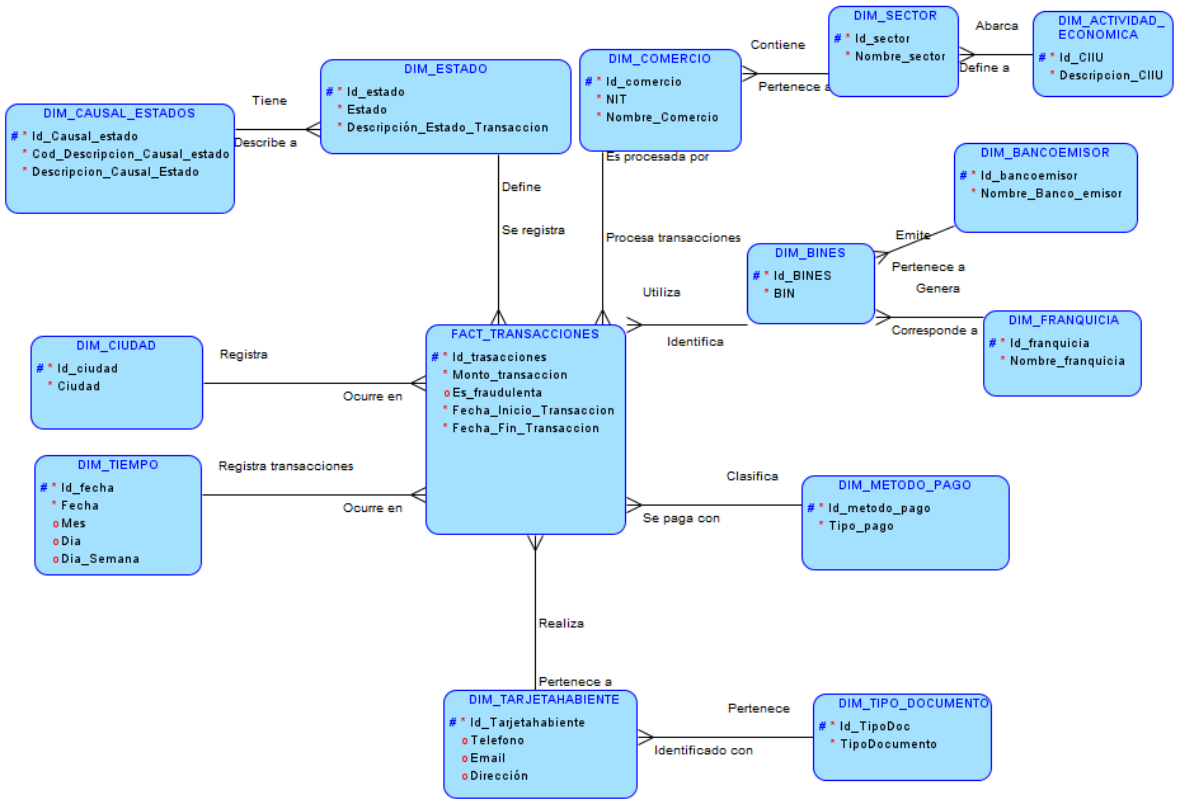


Figura 3: Modelo Lógico - Esquema Estrella del Data Warehouse

El modelo lógico presenta los siguientes componentes:

Tabla de Hechos (FACT_TRANSACCIONES): Almacena cada transacción procesada por la pasarela con sus métricas cuantificables. Esta contiene:

- Id_fecha (referencia temporal).
- Id_comercio (identificación del negocio afiliado).
- Id_metodo_pago (método de pago utilizado).
- Id_estado (estado transaccional final).
- Id_ciudad (ubicación geográfica nacional de la transacción).
- Id_sector (sector económico del comercio).
- Id_CIIU (Descripción actividad económica).
- Id_franquicia (franquicia de la tarjeta si aplica).
- Id_bancoemisor (institución financiera).
- Id_tiempo
- Id_tarjetahabiente
- Id_bines
- Monto_transaccion (valor monetario procesado en la transacción).
- Es_fraudulenta (indicador binario de fraude detectado).

La granularidad se establece a nivel de transacción individual, permitiendo agregaciones flexibles según necesidades analíticas.

Tablas de Dimensiones: Proporcionan el contexto descriptivo que permite analizar los hechos desde diferentes perspectivas:

- DIM_COMERCIO: Información completa de comercios afiliados a la pasarela.
- DIM_METODOPAGO: Caracteriza los métodos de pago aceptados y sus franquicias.
- DIM_ESTADO: Mapea todos los códigos de respuesta transaccional posibles (creada, aprobada, rechazada, pendiente de pago, anulada).
- DIM_CAUSAL_ESTADOS: Información completa sobre el estado de la transacción.
- DIM_TIEMPO: Jerarquía temporal completa para análisis de series de tiempo.
- DIM_CIUDAD: Estructura geográfica.
- DIM_FRANQUICIA: Información de franquicias de pago (Visa, Mastercard, American Express, Diners Club, etc.).
- DIM_BANCOEMISOR: Información de la institución financiera.
- DIM_BINES: Información sobre identifiación del codigo de las tarjetas.
- DIM_TARJETAHABIENTE: Información sobre el cliente o usuario.
- DIM_TIPO_DOCUMENTO: Mapea todos los tipos de documento de identidad.

Modelo relacional En la figura 4 se detalla la implementación física del diseño en Oracle Database, especificando tipos de datos, restricciones de integridad y relaciones entre tablas.

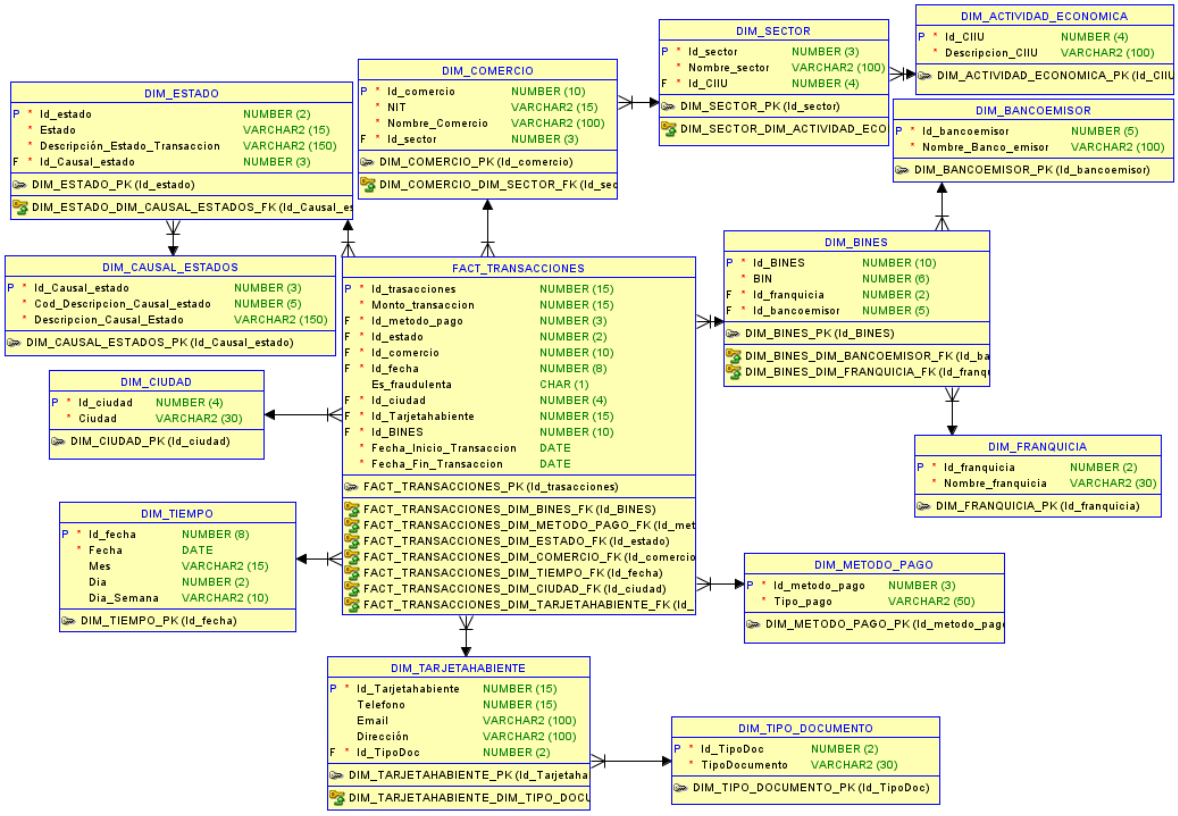


Figura 4: Modelo Relacional - Implementación en Oracle Database

El modelo relacional define:
Tipos de Datos:
Campos de texto: VARCHAR2 con longitudes específicas optimizadas:

- VARCHAR2(15): Campos cortos como estados y NIT.
- VARCHAR2(30): Campos cortos como franquicias, tipo de documento y ciudades.
- VARCHAR2(50): Campos cortos como tipo de pago.
- VARCHAR2(100): Campos como email, dirección, nombre del sector, descripción de CIIU y de estado de la transacción..
- VARCHAR2(150): Descripción de las causales de los estados.

Campos numéricos:

- INTEGER: Para contadores, indicadores y códigos de estado.
- NUMBER: Para montos transaccionales y métricas (tiempos).

Llaves y Relaciones:

- FKFACTESTADO: Vincula FACT_TRANSACCIONES con DIM_ESTADO.

- FKFACTCOMERCIO: Vincula FACT_TRANSACCIONES con DIM_COMERCIO.
- FKFACTTIEMPO: Vincula FACT_TRANSACCIONES con DIM_TIEMPO.
- FKFACTCIUDAD: Vincula FACT_TRANSACCIONES con DIM_CIUDAD.
- FKFACTTARJETAHABIENTE: Vincula FACT_TRANSACCIONES con DIM_TARJETAHABIENTE.
- FKFACTBINES: Vincula FACT_TRANSACCIONES con DIM_BINES.
- FKFACTMETODOPAGO: Vincula FACT_TRANSACCIONES con DIM_METODO_PAGO.

Consideraciones de Implementación:

- Los campos VARCHAR2 utilizan la codificación UTF-8 para soporte de caracteres especiales colombianos.
- Los campos TIMESTAMP incluyen información de zona horaria para correcta interpretación temporal.
- Los indicadores booleanos se implementan como CHAR(1) con restricciones CHECK ('S','N').
- Normalización sin duplicación: La estructura garantiza 3FN (Tercera Forma Normal) sin duplicación de atributos. Por ejemplo, DIM_TARJETAHABIENTE solo contiene Id_TipoDoc como referencia, evitando replicación de datos de tarjetahabiente y posibles inconsistencias cuando cambian atributos de tipos de documentos.

Este modelo dimensional permite realizar análisis eficientes respondiendo preguntas de negocio como:

- ¿Cuál es el volumen transaccional por comercio y período?
- ¿Qué métodos de pago tienen mayor tasa de aprobación?
- ¿Cuáles son los patrones temporales de transacciones (horarios pico, días de mayor actividad)?
- ¿Qué ciudades generan más transacciones?
- ¿Cuál es el tiempo promedio de respuesta por tipo de transacción?

Capa de Presentación: Conecta las herramientas de Business Intelligence con la base de datos Oracle para generar visualizaciones y análisis.

7.3. Estándares de la solución

- **Estándares de Desarrollo:** El código se desarrolla en Python 3.9+ utilizando bibliotecas como Pandas para procesamiento en memoria, PySpark para procesamiento distribuido, cxOracle para conectividad con base de datos y logging para registro de ejecuciones.

- **Estándares de Interoperabilidad:** Los archivos CSV utilizan encoding UTF-8 para soporte de caracteres especiales. Las fechas se manejan en formato ISO 8601 (YYYY-MM-DD HH:MM:SS) en UTC, realizando conversiones a zona horaria local solo en la capa de presentación. La conexión a Oracle se realiza mediante cxOracle con parámetros de conexión almacenados de forma segura en archivos de configuración o variables de entorno.
- **Estándares de Calidad de Datos:** Se implementan controles de validación para garantizar exactitud, completitud y coherencia. Los campos críticos deben estar completos en más del 98 por ciento de los registros. Se detectan y eliminan duplicados mediante comparación de identificadores únicos. Los valores deben pertenecer a dominios válidos según catálogos definidos.
- **Estándares de Seguridad:** El sistema cumple con tres marcos normativos fundamentales para el sector financiero: ISO 8583: Estándar internacional para intercambio de mensajes en transacciones con tarjetas de pago. PCI DSS (Payment Card Industry Data Security Standard): Requisitos de seguridad para proteger información de tarjetas de pago. Ley 1581 de 2012: Ley colombiana de protección de datos personales.
- **Estándares de Documentación:** Se mantiene documentación actualizada del diccionario de datos describiendo la estructura de información procesada. Los flujos de datos están diagramados mostrando el recorrido completo desde fuentes hasta destino con transformaciones aplicadas.

8. Planeación del Trabajo

8.1. Descomposición de actividades WBS

Para la ejecución del proyecto consideramos importante las fases de planificación, diseño, validación y cierre, donde se contempla de manera conceptual definir los diferentes procedimientos que se deben cumplir como guía de una posible implementación de la solución propuesta, a continuación se detalla cada fase: figura 5.

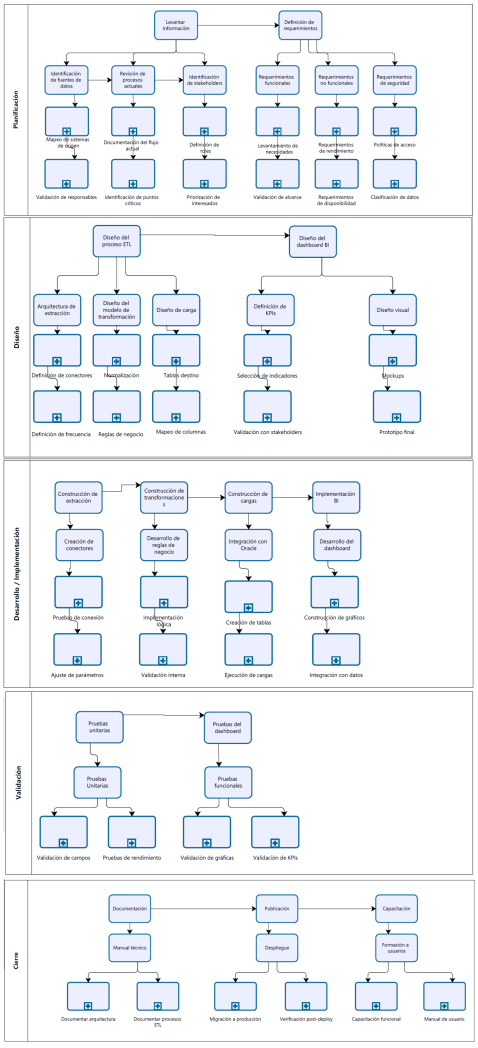


Figura 5: Diagrama WBS

8.2. Diagrama de Gantt

Mediante el siguiente diagrama de gantt, se busca ilustrar las fases de las actividades contenidas dentro del proyecto, donde permite mostrar de manera visual e ilustrativa, las fechas de inicio y fin de cada fase, y los tiempos tomados para su ejecución, los podrá visualizar mediante la tabla 1 que se muestra a continuación y la figura 6.

Nº	Actividad	Tiempo (días)	Responsable	Riesgo
1	Planteamiento del problema y propuesta de información de valor para Dashboard	114	Grupo D	Tiempos prolongados en la definición del problema, alcance del proyecto y estándar de información.
2	Establecimiento de objetivos	15	Grupo D	Definición de objetivos no claros o alcanzables.
3	Levantamiento de requerimientos	5	Grupo D	Definición de requerimientos no claros.
4	Identificación de fuentes de datos	3	Grupo D	No cumplimiento de algunas normativas.
5	Limpieza de datos	30	Grupo D	Eliminación de data de valor.
6	Conexión y cargue de la data	3	Grupo D	Capacidad suficiente para la ejecución de la actividad.
7	Modelado de la data	8	Grupo D	Estructura o Captura de la data duplicada o incompleta.
8	Pruebas de la data	5	Grupo D	Pruebas no exhaustivas que den lugar a la corrupción de datos.
9	Diseño del dashboard	10	Grupo D	Diseño poco intuitivo o amigable con el usuario final.
10	Pruebas del dashboard	5	Grupo D	Cálculos cuantitativos sobre datos no relevantes.
11	Publicación	1	Grupo D	Exposición de información a personal no autorizado.
12	Entrega operativa	3	Grupo D	Baja participación del personal objetivo.

Tabla 1: Diagrama de actividades

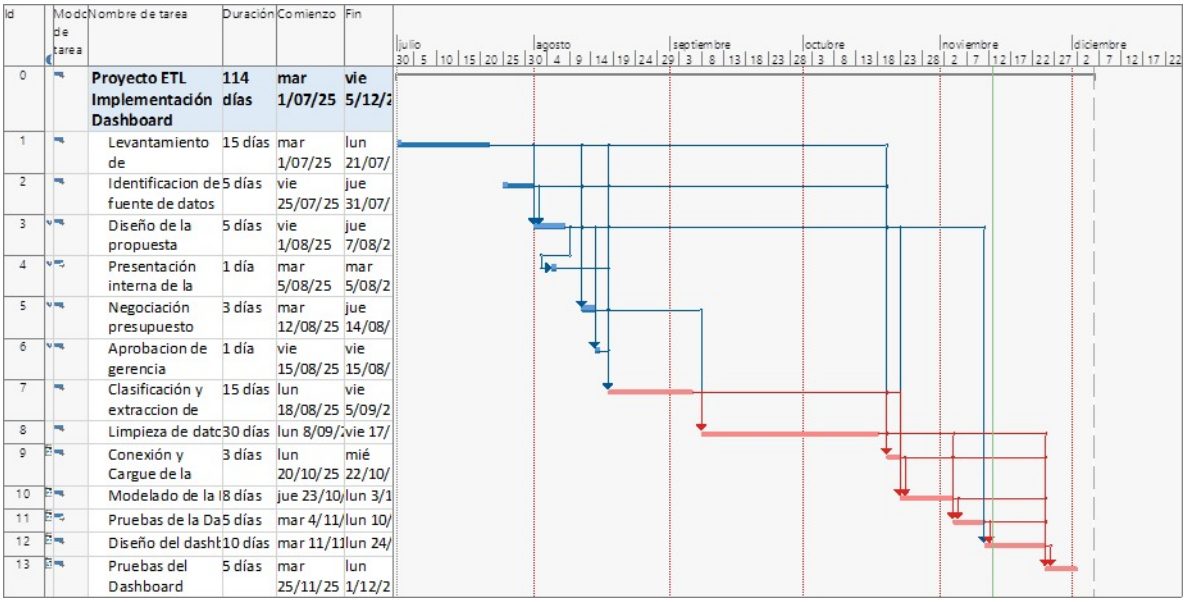


Figura 6: Diagrama de Gantt

9. Presupuesto

Presupuesto del planteamiento del proyecto

En la tabla 2 se presenta el presupuesto propuesto para este proyecto, correspondiente a los costos a considerar en el diseño de un proceso ETL, teniendo en cuenta las herramientas, componentes y herramientas a utilizar. El valor planteado es de: $4,450,000($COP$).

COSTOS DE HERRAMIENTAS - SOLUCIÓN ETL - BI - PASARELA				
Nº	**CONCEPTO**	**CANTIDAD**	**PRECIO (COP)**	**TOTAL (COP)**
1	Oracle Database Academy	3	$ -	$ -
2	Bizagi Modeler	3	$ -	$ -
3	Oracle SQL Developer Modeler	3	$ -	$ -
4	Microsoft Project	3	$ -	$ -
5	Equipos de cómputo	5	$ 800.000	$ 4.000.000
6	Conectividad	3	$ 150.000	$ 450.000
TOTAL				**$4.450.000**

Tabla 2: Presupuesto herramientas y/o componentes

A continuación se estima presupuesto de los costos asociados al equipo de trabajo involucrado para el diseño del proyecto propuesto, se estima un valor de $11,518,880($COP$), se establece la discriminación del concepto, cantidad, valor hora y el total general, teniendo en cuenta la legislación laboral vigente de 42 horas laborales. Ver tabla 3

Nº	**CONCEPTO**	**CANTIDAD HORAS**	**PRECIO HORA (COP)**	**TOTAL (COP)**
1	Horas hombre Co-Director	120	$ 27.490	$ 3.298.800
2	Horas hombre Director	08	$ 27.490	$ 219.920
3	Horas hombre estudiante 1	160	$ 16.667	$ 2.666.720
4	Horas hombre estudiante 2	160	$ 16.667	$ 2.666.720
5	Horas hombre estudiante 3	160	$ 16.667	$ 2.666.720
TOTAL				**$ 11.518.880**

Tabla 3: Presupuesto Recurso Humano

Finalmente, en la tabla 4 se determina presupuesto final del proyecto, donde se contempla el consolidado de los costos relacionados anteriormente, para un monto total de $15,968,880($COP$):

Nº	**CONCEPTO**	**CANTIDAD**	**PRECIO (COP)**	**TOTAL (COP)**
1	Horas hombre Co-Director	120	$ 27.490	$ 3.298.800
2	Horas hombre Director	08	$ 27.490	$ 219.920
3	Horas hombre estudiante 1	160	$ 16.667	$ 2.666.720
4	Horas hombre estudiante 2	160	$ 16.667	$ 2.666.720
5	Horas hombre estudiante 3	160	$ 16.667	$ 2.666.720
1	Oracle Database Academy	3	$ -	$ -
2	Bizagi Modeler	3	$ -	$ -
3	Oracle SQL Developer Modeler	3	$ -	$ -
4	Microsoft Project	3	$ -	$ -
5	Equipos de cómputo	5	$ 800.000	$ 4.000.000
6	Conectividad	3	$ 150.000	$ 450.000
TOTAL GENERAL DEL PROYECTO				**$ 15.968.880**

Tabla 4: Costo total del proyecto

Presupuesto proyectado del proyecto (proyección teórica) a seis meses:

En la tabla 5 se presenta el presupuesto del proyecto con una proyección a seis meses. Esta estimación corresponde a una simulación realizada únicamente con fines académicos, basada en el planteamiento del trabajo, el costo proyectado es de \$154,300,000(COP).

COSTOS DE HERRAMIENTAS - SOLUCIÓN ETL - BI - PASARELA				
HERRAMIENTA	**TIPO**	**CANTIDAD**	**COSTO UNITARIO (COP)**	**TOTAL 6M (COP)**
Apache Spark	ETL Engine (Open Source)	1	\$ -	\$ -
Python + Pandas	Procesamiento (Open Source)	1	\$ -	\$ -
Oracle Database Enterprise	Base de Datos	1	\$ 45.000.000	\$45.000.000
Power BI Pro	BI - Visualización	10	\$ 270.000	\$ 2.700.000
Tableau Creator	BI - Análisis Avanzado	3	\$ 1.800.000	\$ 5.400.000
Servidor ETL Cloud	Infraestructura	1	\$ 21.000.000	\$ 21.000.000
Servidor BD Cloud	Infraestructura	1	\$ 36.000.000	\$ 36.000.000
Almacenamiento (5TB)	Infraestructura	3	\$ 4.800.000	\$4.800.000
Backup y DR	Seguridad	1	\$ 150.000	\$ 450.000
Monitoreo y Logging	Operación	1	\$ 2.400.000	\$ 2.400.000
Firewall y Seguridad	Seguridad	1	\$ 150.000	\$ 450.000
Certificación PCI DSS	Cumplimiento	1	\$ 25.000.000	\$ 25.000.000
GitHub Enterprise	Control de Versiones	1	\$ 1.200.000	\$ 1.200.000
TOTAL				**\$ 154.300.000**

Tabla 5: Presupuesto herramientas y/o componentes proyectado

- Apache Spark y Python son herramientas open source (sin costo de licencia).
- Los costos de infraestructura cloud son estimados mensuales x 6 meses.
- Power BI Pro: 45,000COP/usuario/mes x 10 usuarios x 6 meses.
- Tableau Creator: 300,000COP/usuario/mes x 3 usuarios x 6 meses.
- Certificación PCI DSS es un costo único durante el proyecto.
- El Total NO incluye costos de personal ni capacitación.

En la tabla 6 nes académicos, basada en el planteamiento del trabajo, el costo proyectado es de \$629,600,000(COP).

PRESUPUESTO DE IMPLEMENTACIÓN			
ITEM	**CANTIDAD**	**PRECIO UNITARIO (COP)**	**TOTAL (COP)**
INFRAESTRUCTURA Y SOFTWARE			
Licencia Oracle Database Enterprise	1	\$190.000.000	\$190.000.000
Soporte Oracle (anual)	1	\$ 41.800.000	\$ 41.800.000
Servidor Base de Datos (16 cores, 128GB RAM)	1	\$ 60.000.000	\$ 60.000.000
Servidor Procesamiento ETL	1	\$ 48.000.000	\$ 48.000.000
Python + PySpark (open source)	1	\$ -	\$ -
Licencias Power BI Pro (10 usuarios)	10	\$ 480.000	\$ 4.800.000
Almacenamiento SSD 5TB	1	\$ 10.000.000	\$ 10.000.000
Sistema de Backup	1	\$ 12.000.000	\$12.000.000
RECURSO HUMANO DESARROLLO (6 meses)			
Arquitecto de Datos Senior (3 meses)	3	\$10.000.000	\$30.000.000
Desarrollador ETL Senior (6 meses)	6	\$7.500.000	\$45.000.000
Desarrollador ETL Junior (6 meses)	6	\$4.000.000	\$24.000.000
DBA Oracle (4 meses)	4	\$ 8.000.000	\$32.000.000
Ingeniero DevOps (3 meses)	3	\$ 7.000.000	\$21.000.000
Analista de Datos/BI (4 meses)	4	\$ 6.000.000	\$24.000.000
QA/Tester (3 meses)	3	\$ 5.000.000	\$15.000.000
Gerente de Proyecto (6 meses)	6	\$ 12.000.000	\$72.000.000
TOTAL			**\$ 629.600.000**

Tabla 6: Presupuesto de implementación proyectado

10. Conclusiones

En el planteamiento del trabajo se evidencia una necesidad de implementar procesos ETL especializados para el tratamiento de datos transacciones en pasarelas de pago, dado que puede generar un gran impacto en el análisis, presentación y disponibilidad de los datos.

Atendiendo a dicha necesidad, se genera una propuesta de diseño de arquitectura ETL que permite integrar, transformar, almacenar de manera eficiente grandes volúmenes de información transaccional. A su vez, garantiza la calidad, trazabilidad y seguridad de los datos, generando una base sólida para el monitoreo y análisis del negocio, ayudando a tomar decisiones rápidas y estratégicas.

Durante el desarrollo conceptual del trabajo se definieron flujos de datos y mecanismos de extracción como tipo batch, ya que se consideró los requerimientos del negocio. De igual manera, la limpieza, validación y estandarización de los datos asegurando la consistencia y precisión de la información almacenada como procesos de gobernanza y seguridad para dar cumplimiento normativo.

- **Logros alcanzados:** Se diseñó una arquitectura ETL que permite de forma integral, consolidar y transformar de manera eficiente datos transaccionales, además de generar un proceso automatizado que da valor a la hora de monitorear y generar alertas, que permite prever fallas y garantizar la operatividad del sistema.
- **Mejoras futuras:** La incorporación de tecnologías avanzadas que generen una mayor capacidad de respuesta de acuerdo al aumento del tráfico de datos, así como modelos de entrenamientos neuronales complejos que permita predecir comportamientos incorrectos.
- **Usos alternativos:** La arquitectura propuesta se podría ajustar a la banca digital, a la captación y fidelización de clientes, así como a otros tipos de negocio que manejen grandes volúmenes de información y requieran análisis datos para tomar decisiones estratégicas.
- **Cursos de acción:** Se recomienda iniciar con la implementación del modelo en un entorno de pruebas en ambiente UAT para hacer las pruebas necesarias y determinar que es seguro dar el paso a producción. En paralelo, se sugiere capacitar al área de negocio y al área técnica, así como también buscar promover a través de correo interno y mensajes corporativos.
- **Lecciones aprendidas:** El planteamiento del trabajo permitió identificar la importancia de generar una definición de requisitos de manera clara y coherente para conocer las limitantes qué se pueden presentar, a nivel técnico. Por otra parte, los diferentes factores, como lo económico, las tecnologías y lo relacionado con el negocio, juegan un papel crucial para poder impulsar una propuesta que tenga afinidad no solo con la necesidad del cliente, sino con lo que busca el mercado actualmente.

Referencias

[1] K. S. Garcia, «Desarrollo de una ETL para la Optimización del Almacenamiento y Análisis de Datos Históricos en Sistecredito,» Trabajo de grado, POLITÉCNICO COLOMBIANO JAIME ISAZA CADAVID, Medellín, Colombia, 2024.

[2] M. Souibgui, F. Atigui, S. Zammali, S. Cherfi y S. Ben Yahia, «Data quality in ETL process: A preliminary study,» en *Procedia Computer Science*, 23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, vol. 159, Elsevier B.V., 2019, págs. 676-687. DOI: 10.1016/j.procs.2019.09.223. dirección: https://www.sciencedirect.com/science/article/pii/S1877050919314097.

[3] S. Supriyati y E. Nurfiqo, «Effectiveness of Payment Gateway in E-Commerce,» en *Proceedings of the International Conference on E-Commerce*, Conference Paper, EAI, jun. de 2019. DOI: 10.4108/eai.18-7-2019.2287932. dirección: https://www.researchgate.net/publication/336417831.

[4] E. V. F. Lapura, J. K. J. Fernandez, M. J. K. Pagatpat y D. D. Dinawanao, «Development of a University Financial Data Warehouse and its Visualization Tool,» en *Procedia Computer Science*, 3rd International Conference on Computer Science and Computational Intelligence 2018, vol. 135, Elsevier Ltd., 2018, págs. 587-595. DOI: 10.1016/j.procs.2018.08.229. dirección: https://www.sciencedirect.com/science/article/pii/S1877050918315254.

[5] E. Short, «A Comprehensive Analysis of Extract Transform Load Frameworks for Enhancing Data Pipeline Efficiency in Business Intelligence Systems,» *QIT Press - International Journal of Business Intelligence*, vol. 5, n.º 1, págs. 1-5, ene. de 2025, Journal ID: QITP0249. dirección: https://www.qitpress.com/articles/QITP-IJBI_05_01_001.

[6] D. A. Encalada Garcia, «Diseño de un marco de trabajo para la implementación de procesos ETL,» Carrera de Computación, Trabajo de titulación, Universidad Politécnica Salesiana, Guayaquil, Ecuador, ene. de 2025.

[7] X. Liu, «Optimizing ETL Dataflow Using Shared Caching and Parallelization Methods,» *arXiv*, sep. de 2014. arXiv: 1409.1639 [cs.DB]. dirección: https://arxiv.org/abs/1409.1639.

[8] C. Henao Villa y G. Cardona Montoya, «Revisión bibliográfica: Business Intelligence en la toma de decisiones para la competitividad,» Facultad de Ingeniería, Trabajo de grado de Maestría, Universidad de Antioquia, Medellín, Colombia, 2015. dirección: https://bibliotecadigital.udea.edu.co/entities/publication/58ef060f-1eb0-4008-8ed5-0a8050dee95c.

[9] C. M. Zapata Jaramillo, A. F. Uribe, J. F. Bedoya y V. Grisales, «Modelo para el proceso de extracción, transformación y carga en bodegas de datos. Una aplicación con datos ambientales,» *Entre Ciencia e Ingeniería*, vol. 11, n.º 22, págs. 27-35, 2017, Universidad Nacional de Colombia, Sede Manizales. DOI: 10.31908/19098367.3339. dirección: https://www.redalyc.org/journal/911/91146925006/html/.

[10] J. S. Zapata Tamayo, «Generación semiautomática de código PL/SQL a partir de representaciones de eventos basadas en esquemas preconceptuales,» Tesis de Maestría en Ingeniería - Ingeniería de Sistemas, Universidad Nacional de Colombia, Medellín, Colombia, jun. de 2019. dirección: https://repositorio.unal.edu.co/handle/unal/76882.

[11] A. Ganeshan, E. Macari, J. O'Brien, K. Schlegel y C. Long, «Magic Quadrant for Analytics and Business Intelligence Platforms,» Gartner, Inc., Stamford, CT, USA, Reporte de Investigación G00806576, jun. de 2025, ID de Documento: 6576602. dirección: https://www.gartner.com/en/documents/6576602.

A. Anexos

A.1. Anexo A: Diagrama WBS.

A.2. Anexo B: Modelo lógico.

A.3. Anexo C: Modelo relacional.

A.4. Anexo D: Diccionario de datos.