

Creación de un almacén de datos para la empresa CITMATEL

**Trabajo de práctica profesional de 4er año de la carrera de ingeniería informática**

**Autor/es:**

Marcos Alejandro Fernández Piñero

**Tutores:**

La Habana, Cuba

Diciembre, 2024

**Índice**

[**Resumen** 3](#_Toc180491505)

[**Abstract** 4](#_Toc180491506)

[**Introducción** 5](#_Toc180491507)

[**Desarrollo** 9](#_Toc180491508)

[**Capítulo 1: Análisis del marco teórico del problema** 9](#_Toc180491509)

[**Capítulo 2: Análisis del negocio y Diseño de la solución** 10](#_Toc180491510)

[**2.1: Análisis y modelado del negocio** 10](#_Toc180491511)

[**2.2 Diseño de la solución** 11](#_Toc180491512)

[**Capítulo 3: Diseño y ejecución de pruebas** 12](#_Toc180491513)

[**Conclusiones** 13](#_Toc180491514)

[**Referencias bibliográficas** 14](#_Toc180491515)

# **Resumen**

# **Abstract**

# **Introducción**

En la era digital actual, las empresas enfrentan un crecimiento exponencial en el volumen de datos que provienen de diversas fuentes. Este fenómeno es impulsado por la proliferación de dispositivos conectados, plataformas digitales y tecnologías emergentes que generan grandes cantidades de información en tiempo real. Según un estudio, herramientas tecnológicas avanzadas han incrementado la eficiencia y precisión en la recopilación de datos en un 85%, lo que ha permitido a empresas como IBM y FedEx mejorar significativamente sus procesos. Sin embargo, este aumento en la cantidad de datos también presenta desafíos significativos para las organizaciones [1].

Este fenómeno, del flujo constante y variado de información ha obligado a las empresas a adoptar nuevas estrategias para su gestión y análisis. La diversidad en los formatos y las fuentes de estos datos complica su integración dentro de los sistemas empresariales existentes y para la toma de decisiones asociada al análisis de estos datos [2]. Además, el manejo inadecuado o la falta de integración eficiente pueden llevar a la creación de silos de datos, donde la información queda aislada dentro de diferentes departamentos o sistemas [3].

Todo esto desembocado en que la capacidad para procesos de Inteligencia de Negocios o Minería de dato se vean afectados. En primer lugar, la calidad y consistencia de los datos son esenciales para obtener resultados precisos. Sin embargo, muchas organizaciones enfrentan problemas con la calidad de los datos debido a su dispersión y falta de estandarización [4]. Además, las herramientas tradicionales a menudo no están equipadas para manejar el volumen y velocidad actuales del flujo de datos, lo que limita su utilidad en la toma de decisiones estratégicas [5].

Para superar estos desafíos, las empresas están adoptando soluciones que centralizan la gestión y análisis de datos. La inteligencia de negocios (BI) se presenta como una solución efectiva al integrar todos los datos distribuidos dentro de un único almacén centralizado. Esta centralización permite una visión más completa del desempeño empresarial y facilita la toma rápida y precisa de decisiones [2]. El uso del proceso ETL (extracción, transformación y carga) es crucial para consolidar los datos provenientes de diversas fuentes en un repositorio centralizado, mejorando así su integridad y accesibilidad [4].

En el contexto cubano actual, empresas como CITMATEL se han visto inmersas en esta situación. CITMATEL, una empresa cubana con presencia en el ámbito digital, enfrenta al desafiante obstáculo de la dispersión de sus datos en múltiples bases de datos correspondientes a sus diversas tiendas digitales. Esta fragmentación dificulta la realización de análisis complejos necesarios para procesos críticos como el márquetin, las finanzas y las estadísticas, se hace necesario en acceso frecuente y variante a estos datos.

La falta de un sistema centralizado y los mecanismos adecuados para el acceso, gestión y visualización de los datos impide a la empresa aprovechar plenamente su potencial de datos, lo que limita su capacidad para responder ágilmente a las demandas del mercado y optimizar sus operaciones internas. Actualmente, Los departamentos solicitan al equipo de desarrollo los datos a través de correo electrónico, los cuales deben hacer consultas SQL directamente en las bases de datos en producción para obtener la información requerida, exigiendo la inversión de tiempo y esfuerzo en tareas que no deberían ser su responsabilidad y limitando la capacidad de gestión y procesamiento de la empresa.

Con este proyecto se busca diseñar e implementar una solución que permita centralizar los datos en un único repositorio accesible, facilitando así el acceso rápido y seguro a información confiable en tiempo real por personal no relacionado a la informática. Esta centralización no solo mejorará la integridad y seguridad de los datos, sino que también fomentará una mayor colaboración entre los departamentos al proporcionar una única fuente de verdad.

Con esta iniciativa, se espera transformar la manera en que la Entidad gestiona sus datos, permitiendo una toma de decisiones más ágil y fundamentada. El proyecto incluirá tanto el diseño e implementación técnica como la documentación exhaustiva del proceso, asegurando que todos los aspectos del sistema sean comprensibles y replicables en el futuro.

De ahí que se establezca la siguiente **situación problemática**: las consultas a los datos de debe hacer de forma manual directamente a las bases de datos consumiendo demasiado tiempo y esfuerzo de parte del personal de desarrollo de software que es el único capacitado para llevar a cabo esta tarea.

Esto plantea una contradicción entre los consumidores de estos datos que requieren un sistema capaz de proporcionárselos de forma clara y directa para su análisis y las estructuras de almacenaje y gestión de los mismos. Planteando la siguiente **pregunta científica**: Cómo se pueden centralizar todas las fuentes de datos en un mismo lugar que permita el acceso a los datos de forma óptima en función de las necesidades de los consumidores y como permitirles a estos acceder a los datos de forma clara y entendible para ellos.

Para poder resolver este problema se ha establecido como **objetivo de la investigación** Implementar un sistema de almacenamiento y gestión de todos los datos de la empresa que satisfaga la demanda de los datos por parte de los consumidores actuales y futuros.

Para la culminación de estas prácticas laborales se acoto la solución hasta la implementación de la Base de Datos centralizada y el sistema de ETL.

Para esto se plantearon como objetivos específicos:

1. Documentación de las bases de datos existentes para la comprensión de las fuentes de datos.
2. Estudio de las diferentes formas de centralizar los datos precedentes de diferentes almacenes de datos en un repositorio.
3. Estudio de las diferentes tecnologías con el fin de la detección de la más adecuada para la resolución del problema
4. Diseño de una solución general que se adapte a las necesidades específicas de la empresa.
5. Desnormalización de las bases de datos existentes y la renormalización de los datos en función de la demanda de los departamentos que requieren los datos.
6. Implementación de la solución propuesta.
7. Implementación de pruebas que validen la efectividad del sistema propuesto.
8. Validación de la capacidad del sistema para adaptase a las necesidades actuales y futuras de la empresa
9. Despliegue general de la solución
10. Documentación de capacitación para el uso del sistema por los usuarios finales

# **Desarrollo**

## **Capítulo 1: Análisis del marco teórico del problema**

El presente capítulo establece el marco teórico esencial para el desarrollo de un Sistema de Soporte a la Toma de decisiones (DSS en inglés) donde se unifiquen y pueden gestionar todos los datos de la entidad **CITMATEL** que gestiona cuatro tiendas en línea, cada una con una base de datos estructuralmente distinta. Este marco teórico proporciona la base conceptual y analítica necesaria para abordar los desafíos inherentes a la integración de múltiples sistemas de bases de datos.

Inicialmente, se presentarán los antecedentes del estudio, donde se revisarán investigaciones previas y estudios relacionados que han abordado problemáticas similares. Se identificarán las brechas existentes en el conocimiento actual que este proyecto busca llenar, destacando la necesidad de soluciones innovadoras para la integración de datos heterogéneos.

A continuación, se procederá con la definición de conceptos clave, donde se clarificarán los términos y conceptos fundamentales que serán recurrentes a lo largo del proyecto. Estas definiciones estarán respaldadas por fuentes académicas pertinentes, asegurando un entendimiento común y preciso.

Posteriormente, se expondrán las teorías relevantes que sustentan el trabajo. Se explicará cómo estas teorías se aplican específicamente al contexto del proyecto, proporcionando un marco conceptual robusto para guiar el desarrollo e implementación del sistema.

El estado del arte ofrecerá una visión general de los desarrollos actuales en el ámbito de la integración de bases de datos y gestión de datos distribuidos. Se incluirán innovaciones recientes y tendencias futuras que podrían influir en el diseño y ejecución del proyecto, asegurando que las soluciones propuestas sean contemporáneas y efectivas.

## **Antecedentes del Estudio**

Para realizar un análisis del estado del arte se precisa primero realizar un levantamiento exhaustivo de investigaciones previas sobre problemas similares y las soluciones aplicadas, es esencial considerar diversos estudios y enfoques que han abordado cuestiones análogas en el campo de interés. Para ello, se realizó una búsqueda en bases de datos académicas sobre términos como “*Business intelligence*”, “Repositorios de datos”, “Almacenes de datos” y “Sistemas de Apoyo a la toma de Decisiones”. De esta investigación se obtuvo una serie de documentos que serán analizados en esta sección.

En el contexto actual, donde las empresas enfrentan un aumento exponencial en el volumen y variedad de datos, se han identificado varios enfoques y soluciones que abordan problemas similares al que enfrenta CITMATEL.

El uso de Data Warehouses (DW) se ha consolidado como una herramienta esencial para integrar y analizar grandes volúmenes de datos provenientes de diversas fuentes para sistemas de soporte a la toma de decisiones (DSS). Estos sistemas permiten transformar datos brutos en información útil y accesible, facilitando así el análisis por parte de los usuarios finales sin necesidad de conocimientos técnicos avanzados [6]. La metodología Hefesto, por ejemplo, ha sido ampliamente utilizada para guiar el diseño e implementación de DW, destacándose por su enfoque en la recolección de requerimientos y procesos ETL (Extracción, Transformación y Carga) eficientes [7][6].

En investigaciones previas, se ha demostrado que la implementación de un DW puede mejorar significativamente la eficiencia operativa y la capacidad de respuesta empresarial. Por ejemplo, en el ámbito bibliotecario, se ha utilizado un sistema de soporte a las decisiones basado en DW para integrar datos heterogéneos y mejorar la gestión bibliotecaria [7]. Este enfoque ha permitido a los gestores generar informes personalizados y optimizar procesos transaccionales diarios.

Asimismo, en sistemas de gestión de proyectos, los repositorios de datos han facilitado la investigación al proporcionar una base estructurada para el análisis y la toma de decisiones. Estos repositorios permiten compartir y reutilizar datos, lo cual es crucial para el descubrimiento de nuevo conocimiento dentro de las organizaciones [8]. En particular, se han desarrollado sistemas que utilizan técnicas avanzadas como la minería de datos para asegurar ingresos y evaluar proyectos bajo condiciones inciertas [8].

Por otro lado, el concepto de data lakes ha sido mencionado como una solución prometedora para manejar grandes volúmenes de datos estructurados y no estructurados por otros estudios como una alternativa más flexible. Este enfoque permite a las empresas almacenar datos en su formato nativo, ofreciendo flexibilidad para el análisis y experimentación por parte de científicos de datos y analistas [9]. Los data lakes pueden servir como áreas de preparación para los DW, facilitando la integración de diversas fuentes de datos sin necesidad de transformaciones iniciales. Sin embargo, presentan desafíos como la gobernanza de datos y la calidad de los mismos, aspectos que requieren atención cuidadosa durante su implementación [10].

De esta exhaustiva se definen como conceptos de interés temas como “Data WareHouse”, “Data lake”, “Data Marks”, “DSS”, “Business intelligence and analytics” y “ETL”.

## **1.2 Análisis de la teoría asociada**

## **Capítulo 2: Análisis del negocio y Diseño de la solución**

### **2.1: Análisis y modelado del negocio**

### **2.2 Diseño de la solución**

## **Capítulo 3: Diseño y ejecución de pruebas**

# **Conclusiones**

# **Referencias bibliográficas**