

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:**

Dra. C. Lisandra Bravo Ilisástigui

La Habana, Cuba

Diciembre, 2024

**Índice**

[**Resumen** 3](#_Toc181053378)

[**Abstract** 4](#_Toc181053379)

[**Introducción** 5](#_Toc181053380)

[**Desarrollo** 9](#_Toc181053381)

[**Capítulo 1: Análisis del marco teórico del problema** 9](#_Toc181053382)

[**1.1 Antecedentes del Estudio** 9](#_Toc181053383)

[**1.2 Análisis de la teoría asociada** 11](#_Toc181053384)

[**Capítulo 2: Análisis del negocio y Diseño de la solución** 12](#_Toc181053385)

[**2.1: Análisis y modelado del negocio** 12](#_Toc181053386)

[**2.2 Diseño de la solución** 13](#_Toc181053387)

[**Capítulo 3: Diseño y ejecución de pruebas** 14](#_Toc181053388)

[**Conclusiones** 15](#_Toc181053389)

[**Referencias bibliográficas** 16](#_Toc181053390)

# **Resumen**

# **Abstract**

# **Introducción**

Según (Calle García, Alvarado Choez, Anzules Cruz, & Menéndez Parrales, 2024), 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.

Este fenómeno del flujo constante y variado de información, como refiere (Biroğul & Gültekin, 2016), 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. Seguidamente en (Kumar, 2023) se plantea que 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.

Todo esto desembocado en que la capacidad para tomar decisiones a partir de estos datos se vea afectada. 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 (Robles Angeles, Robles Angeles, Ángeles Hernández, & Acosta Miranda, 2023). 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(Dmitriyev, Mahmoud, & Marín-Ortega, 2022).

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(Biroğul & Gültekin, 2016). Finalmente, según (Robles Angeles et al., 2023), 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.

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 se deben hacer de forma manual directamente a las bases de datos debido a la falta de un repositorio de datos que permita la recopilación y manipulación de los 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 limita a los equipos consumidores de estos datos de poder hacer consultas más complejas o especializadas.

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 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(Vite, Villegas, & Pizarro, 2018). 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 (Romero-Chuquital & Melendres-Velasco, 2023).

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 (Romero-Chuquital & Melendres-Velasco, 2023). Este enfoque ha permitido a los gestores generar informes personalizados y optimizar procesos transaccionales diarios.

Asimismo, y según (Valle, Rocano, Siguenza-Guzman, Cattrysse, & Galarza, 2014), 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. En particular, se han desarrollado sistemas que utilizan técnicas avanzadas como la minería de datos para asegurar ingresos y evaluar proyectos bajo condiciones inciertas.

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(Llave, 2018). 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, según (Sulova, 2019) presentan desafíos como la gobernanza de datos y la calidad de los mismos, aspectos que requieren atención cuidadosa durante su implementación.

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

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

La inteligencia de negocios (BI, por sus siglas en inglés) se define como un marco integral que abarca un conjunto de herramientas y técnicas destinadas a transformar datos brutos en información valiosa. Esta información es crucial para la toma de decisiones estratégicas dentro de una organización [11]. A medida que las empresas generan cantidades masivas de datos, la inteligencia de negocios se ha convertido en una infraestructura crítica que permite a las organizaciones reaccionar de manera efectiva a las operaciones tanto operativas como estratégicas de sus clientes, mercados y otros grupos de interés [12].

Este proceso implica la extracción, análisis y presentación de datos mediante sistemas de información más entendibles por los clientes. Estos sistemas permiten a los gerentes obtener un conocimiento profundo sobre el comportamiento histórico del negocio y su entorno, lo que facilita decisiones informadas que mejoran el rendimiento organizacional y optimizan la utilidad económica mediante una reducción significativa de costos [13].

Como ya se mencionó, los sistemas de BI deben cumplir con la característica de permitir a los clientes entender estos datos y tomar decisiones en base a ellos, por lo que surge la necesidad de la creación de sistemas de apoyo a la toma de decisiones(DSS).

Los Sistemas de Apoyo a la Toma de Decisiones (DSS, por sus siglas en inglés) son aplicaciones prácticas de la inteligencia de negocios, diseñadas para asistir a los gestores en la toma de decisiones informadas basadas en datos analíticos. Estos sistemas integran diversas fuentes de información y aplican modelos analíticos avanzados para proporcionar recomendaciones y pronósticos que facilitan el proceso decisional en las organizaciones [14].

Los DSS se caracterizan por su capacidad para manejar grandes volúmenes de datos y transformarlos en información útil, permitiendo a los gestores evaluar diferentes escenarios y tomar decisiones estratégicas más acertadas. Al utilizar herramientas de análisis predictivo y modelos de simulación, los DSS ayudan a identificar tendencias, evaluar riesgos y oportunidades, y optimizar los recursos empresariales [15].

Según refiere [16] para logran una mayor eficiencia en el tratamiento de fuentes muy dispersas y heterogéneas, se hace necesario el empleo de herramientas ETL para la correcta recuperación y preparación de estos datos.

Los procesos de Extracción, Transformación y Carga (ETL) son fundamentales en la preparación de datos para su análisis en sistemas de inteligencia de negocios (BI) y sistemas de apoyo a la toma de decisiones (DSS). Estos procesos permiten integrar datos desde múltiples fuentes hacia un formato coherente y utilizable, lo que es esencial para convertir datos brutos en información valiosa [16].

Por las explicaciones de [17], estos sistemas, como su nombre lo indica, constan de 3 partes: Extracción, Transformación y Carga. La primera implica la recopilación de datos desde diversas fuentes, que pueden incluir bases de datos, archivos planos, servicios web, entre otros. Esta fase es crítica porque asegura que todos los datos relevantes sean capturados para el análisis posterior. Por otro lado, la Transformación es el proceso donde los datos extraídos se limpian y se convierten en un formato adecuado para el análisis. Esto puede incluir tareas como la depuración de datos duplicados, la normalización de formatos y la aplicación de reglas de negocio específicas. La transformación es vital para asegurar que los datos sean precisos y consistentes. Por último, la Carga es la fase final donde los datos transformados se almacenan en un sistema de destino, como un data warehouse o un data mart. La eficiencia en esta etapa garantiza que los datos estén disponibles para su análisis en tiempo real o casi real.

Sin embargo, como vimos en el proceso de Carga, para poder llevarlo a cabo es fundamental contar con un lugar bien estructurado para el almacenamiento de estos datos. Durante la anterior investigación inicial detectaron 3 “tipos de almacenes”: Data Warehause, Data lakes y Dama Marts.

Un **Data Warehouse** es un sistema centralizado diseñado para almacenar grandes volúmenes de datos estructurados provenientes de diversas fuentes. Su principal objetivo es facilitar el análisis y la generación de informes. Este organiza los datos en esquemas predefinidos, como el esquema estrella o copo de nieve, lo que permite realizar consultas complejas de manera eficiente. Este tipo de almacenamiento es ideal para el análisis histórico y la toma de decisiones estratégicas, ya que ofrece una visión consolidada y limpia de los datos [18].

Por su parte un **Data Mar**t es una versión más pequeña y especializada de un data warehouse. Está diseñado para atender las necesidades específicas de un departamento o área dentro de una organización, como ventas o finanzas. Los data marts suelen contener subconjuntos de datos relevantes para ese departamento, permitiendo un acceso más rápido y eficiente a la información necesaria para análisis específicos. Esto reduce la complejidad y el tiempo de respuesta en comparación con un data warehouse completo [19].

Finalmente, Un **Data Lake** es un repositorio de almacenamiento que permite guardar grandes cantidades de datos en su formato nativo, ya sean estructurados, semiestructurados o no estructurados. A diferencia del data warehouse, los data lakes no requieren una estructura predefinida para los datos, lo que los hace más flexibles y escalables. Son ideales para el análisis exploratorio y el procesamiento de big data, ya que permiten almacenar datos brutos que pueden ser transformados y analizados posteriormente según las necesidades del negocio [18].

|  |  |  |  |
| --- | --- | --- | --- |
| Característica | Data Warehouse | Data Mart | Data Lake |
| Formato de Datos | Estructurado | Estructurado | Estructurado, semiestructurado y no estructurado |
| Escalabilidad | Limitada | Limitada | Alta |
| Flexibilidad | Baja | Media | Alta |
| Objetivo Principal | Análisis histórico y generación de informes | Análisis específico por departamento | Almacenamiento y análisis exploratorio de big data. |
| Tiempo de Respuesta | Lento (debido a la complejidad) | Rápido (enfocado en necesidades específicas) | Variable (depende del tipo de análisis) |

Gracias a esta tabla que resume las características de cada uno, y teniendo en cuenta las características del negocio, resulta más conveniente la implementación de un Data Lake sobre los otros sistemas de almacenamiento, pues permite el manejo de datos más heterogéneos y es más flexible que las demás.

* 1. **Soluciones existentes** y **limitaciones**

E

## **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**

Biroğul, S., & Gültekin, H. (2016). Importance of Business Intelligence Solution on Decision-Making Process of Companies. *International Journal of Applied Mathematics, Electronics and Computers, 4*, 86-89.

Calle García, A. J., Alvarado Choez, K. J., Anzules Cruz, J. J., & Menéndez Parrales, A. G. (2024). NUEVAS HERRAMIENTAS TECNOLÓGICAS EN LA RECOPILACIÓN DE DATOS DE MERCADO. *Ciencia y Desarrollo*.

Dmitriyev, V., Mahmoud, T., & Marín-Ortega, P. M. (2022). SOA enabled ELTA: approach in designing business intelligence solutions in Era of Big Data. *International Journal of Information Systems and Project Management*.

Kumar, S. (2023). Data Silos A Roadblock for AIOps. *ArXiv, abs/2312.10039*.

Llave, M. R. (2018). Data lakes in business intelligence: reporting from the trenches. *Procedia computer science, 138*, 516-524.

Robles Angeles, L. J., Robles Angeles, M. B., Ángeles Hernández, L., & Acosta Miranda, M. L. (2023). El poder de los datos: Cómo ETL reformuló la gestión de datos. *Programación Matemática y Software*.

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.

Sulova, S. (2019). *The Usage of Data Lake for Business Intelligence Data Analysis.* Paper presented at the International Conference Information and communication technologies in business and education.

Valle, V. A. H., Rocano, W. R. P., Siguenza-Guzman, L., Cattrysse, D., & Galarza, V. H. S. (2014). *Diseño e Implementación de un Sistema de Soporte de Decisiones para el Centro de Documentación Regional “Juan Bautista Vázquez”*.

Vite, V. J., Villegas, S. C., & Pizarro, G. (2018). Aplicación de un Spatial Data Warehouse en la gestión de proyectos de vinculación: Caso de Estudio. *Journal of Science and Research: Revista Ciencia e Investigación, 3*(9), 19-24.