

UNIVERSIDAD TÉCNICA DEL NORTE



Facultad de Ingeniería en Ciencias Aplicadas Carrera de Software

IMPLEMENTACIÓN DE INTELIGENCIA DE NEGOCIOS PARA APOYAR EN LA TOMA DE DECISIONES INFORMADAS DE LA EMPRESA "SARI POPULAR"

Trabajo de grado presentado ante la ilustre Universidad Técnica del Norte previo a la obtención del título de Ingeniero en Software.

Autor:

William Anderson Nazate Enríquez

Director:

Ph.D Iván Danilo García Santillán

Ibarra – Ecuador

2022



UNIVERSIDAD TÉCNICA DEL NORTE BIBLIOTECA UNIVERSITARIA

AUTORIZACIÓN DE USO Y PUBLICACIÓN A FAVOR DE LA UNIVERSIDAD TÉCNICA DEL NORTE

1. IDENTIFICACIÓN DE LA OBRA

En cumplimiento del Art. 144 de la Ley de Educación Superior, hago la entrega del presente trabajo a la Universidad Técnica del Norte para que sea publicado en el Repositorio Digital Institucional, para lo cual pongo a disposición la siguiente información:

DATOS DE CONTACTO			
CÉDULA DE IDENTIDAD:	100464153-4		
APELLIDOS Y NOMBRES:	WILLIAM ANDERSON NAZATE ENRÍQUEZ		
DIRECCIÓN:	San Antonio, barrio los soles		
EMAIL:	wanazatee@utn.edu.ec wnazate521@gmail.com		
TELÉFONO FIJO:	062 551-021	TELÉFONO MÓVIL:	0985588095

DATOS DE LA OBRA	
TÍTULO:	IMPLEMENTACIÓN DE INTELIGENCIA DE NEGOCIOS PARA APOYAR EN LA TOMA DE DECISIONES INFORMADAS DE LA EMPRESA "SARI POPULAR"
AUTOR:	WILLIAM ANDERSON NAZATE ENRÍQUEZ
FECHA:	26/09/2022
PROGRAMA:	<input checked="" type="checkbox"/> PREGRADO <input type="checkbox"/> POSGRADO
TÍTULO POR EL QUE OPTA:	Ingeniero en Software
ASESOR / DIRECTOR:	PhD. Iván García

2. CONSTANCIAS

El autor manifiesta que la obra objeto de la presente autorización es original y se la desarrolló, sin violar derechos de autor de terceros, por lo tanto, la obra es original y que es el titular de los derechos patrimoniales, por lo que asume la responsabilidad sobre el contenido de la misma y saldrá en defensa de la Universidad en caso de reclamación por parte de terceros.

Ibarra, a los 26 días del mes de septiembre de 2022

EL AUTOR:



William Anderson Nazate Enriquez
100464153-4

CERTIFICADO DEL DIRECTOR DE TRABAJO DE GRADO



UNIVERSIDAD TÉCNICA DEL NORTE
FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS

CERTIFICACIÓN DEL ASESOR

Certifico que la Tesis previa a la obtención del título de Ingeniero en Software con el tema: "IMPLEMENTACIÓN DE INTELIGENCIA DE NEGOCIOS PARA APOYAR EN LA TOMA DE DECISIONES INFORMADAS DE LA EMPRESA SARI POPULAR" ha sido desarrollada y terminada en su totalidad por el Sr. William Anderson Nazate Enríquez, con cédula de identidad Nro. 100464153-4 bajo mi supervisión, por lo cual firmo en constancia.

1002292603
IVAN DANILO
GARCIA
SANTILLAN

Firmado
digitalmente por
1002292603 IVAN
DANILO GARCIA
SANTILLAN
Fecha: 2022.09.20
08:27:38 -05'00'

PhD. Iván García

DIRECTOR DE TESIS

CERTIFICACIÓN INSTITUCIÓN BENEFICIARIA



Dir. Matriz: Av. Mariano Acosta 12-51
Telf.: (06) 2606 287 Librería: (06) 2603 749
E-mail: sararodriguezpp@hotmail.com
Dir. Bodega (Ventas al por mayor)
Pílanqui Av. José Tobar 11-25
(diagonal al parque Ciudad Blanca)
Telf.: (06) 2641 193 E-mail: ventaspp@hotmail.com
www.papeleriaspopular.com

CARTA DE AUSPICIO

Ibarra, 01 de Abril de 2021

Me permito informar a Ustedes que el señor William Anderson Nazate Enríquez, con cedula de ciudadanía Nro. 100464153-4. Estudiante de pregrado de la Universidad Técnica del Norte, ha sido aceptado en esta institución para que realice su trabajo de Grado con el tema: **"Implementación de inteligencia de Negocios para apoyar en la toma de decisiones informadas de la empresa "Sari Popular"**

"Sari Popular" brindará las facilidades e información necesaria, así como garantizará el avance y entrega de los resultados.

Agradezco su atención.

Atentamente,



Sara Alicia Rodríguez Chacón

CI: 0400893152

Gerente - General

DEDICATORIA

Le dedico el resultado del presente trabajo de titulación a toda mi familia y seres queridos, principalmente a mis padres que me apoyaron incondicionalmente desde el inicio de todo el proceso, con su sacrificio, su tiempo, su atención, su predisposición por acompañarme en todo momento.

También dedico este trabajo a todas las personas que aportaron, confiaron y creyeron en mí, en algún momento de todo este proyecto.

AGRADECIMIENTO

Me gustaría empezar esta sección de agradecimiento con mis familiares, mis padres como pilares, por haberme brindando su espacio, preocupación y apoyo en buenos y difíciles momentos.

Agradezco a todos mis amigos y las personas más cercanas, también a los compañeros con quienes formé más amistad en el transcurso de la carrera, por su ayuda de manera desinteresada, gracias por su buena voluntad y compañerismo.

Igualmente agradezco a la universidad conjuntamente con los docentes profesionales, tutores y asesores que aportaron con su guía, observaciones, experiencia y conocimiento.

TABLA DE CONTENIDOS

AUTORIZACIÓN DE USO Y PUBLICACIÓN A FAVOR DE LA UNIVERSIDAD TÉCNICA DEL NORTE	II
CERTIFICADO DEL DIRECTOR DE TRABAJO DE GRADO	IV
CERTIFICACIÓN INSTITUCIÓN BENEFICIARIA.....	V
DEDICATORIA	VI
AGRADECIMIENTO.....	VII
RESUMEN.....	XIV
ABSTRACT.....	XV
INTRODUCCIÓN.....	16
Tema.....	16
Problema.....	16
Antecedentes	16
Situación Actual	16
Prospectiva.....	17
Planteamiento del problema	17
Objetivos	17
Objetivo General	17
Objetivos Específicos.....	18
Alcance	18
Justificación	19
CAPÍTULO 1.....	21
Marco Teórico.....	21
1.1 Empresa “Sari Popular”	21
1.2 Inteligencia de Negocios	22
1.2.1 Calidad de datos	22
1.2.1.1 ISO/IEC 25012.....	25
1.2.2 Historia BI.....	28
1.2.3 Definición	29
1.2.4 Importancia	30
1.3 Herramientas BI y Data Mining	32
1.3.1 BI.....	32
1.3.2 Data Mining.....	33
1.4 Metodologías BI y Data Mining	35
1.4.1 Inteligencia de Negocios (Business Intelligence).....	35
1.4.2 Minería de datos (Data Mining)	39

1.5 Trabajos Relacionados.....	41
1.6 Matriz de Conceptos.....	43
CAPÍTULO 2.....	44
2.1 Definición de requerimientos e indicadores.....	44
2.1.1 Requerimientos en BI.....	44
2.1.2 Indicadores.....	44
2.1.3 Requerimientos e indicadores en "Sari Popular".....	45
2.2 Uso de la Norma ISO/IEC 25012 – Calidad de Datos.....	46
2.3 Procesos ETL y construcción de Data Warehouse.....	49
2.3.1 Análisis de Requerimientos.....	50
2.3.1.1 Preguntas del Negocio.....	50
2.3.1.2 Indicadores y Perspectivas.....	52
2.3.1.3 Modelo Conceptual.....	53
2.3.2 Análisis de Data Sources.....	54
2.3.2.1 Hechos e Indicadores.....	54
2.3.2.2 Mapeo.....	55
2.3.2.3 Granularidad.....	57
2.3.2.4 Modelo Conceptual Ampliado.....	59
2.3.3 Modelo lógico del Data Warehouse.....	60
2.3.3.1 Tipología.....	60
2.3.3.2 Tablas de Dimensiones.....	60
2.3.3.3 Tablas de Hechos.....	63
2.3.3.4 Uniones.....	64
2.3.4 Integración de datos.....	65
2.3.4.1 Carga Inicial.....	65
2.3.4.2 Actualización.....	72
2.4. Reportes, Dashboard y Minería de datos.....	73
2.4.3 Algoritmo A Priori.....	76
2.4.4 Aplicación del algoritmo A priori en Weka.....	77
2.5 Cuestionario SUS (System Usability Scale - Escala de usabilidad del sistema).....	78
CAPÍTULO 3.....	80
3.1 Definición de métricas de evaluación.....	80
3.2 Norma ISO 25012.....	80
3.3 Sistema BI y Preguntas del Negocio.....	80
3.4 Algoritmo A priori.....	83
3.5 Análisis de resultados de la aplicación del cuestionario SUS.....	84
3.6 Discusión.....	85
CONCLUSIONES.....	87

RECOMENDACIONES	88
REFERENCIAS	89
ANEXOS	91
Anexo A: Formato ejecutado de Test de usabilidad SUS.....	91

ÍNDICE DE FIGURAS

Fig. 1. Diagrama del Alcance	17
Fig. 2. Diagrama del Alcance	19
Fig. 3. Estructura Organizacional – Sari Popular	22
Fig. 4. Datos, Información, Conocimiento.....	23
Fig. 5. Divisiones Norma SQuaRE	24
Fig. 6. Características y conceptos ISO/IEC 25012	26
Fig. 7. Comparación entre herramientas BI	33
Fig. 8. Comparación herramientas Data Mining	35
Fig. 9. Metodología Hefesto	36
Fig. 10. Diagrama del ciclo de vida dimensional	37
Fig. 11. Comparación entre metodologías Data Mining	41
Fig. 12. Matriz de Conceptos parte 1	43
Fig. 13. Matriz de Conceptos parte 2.....	43
Fig. 14. Diagrama ETL	49
Fig. 15. Modelo conceptual	54
Fig. 16. Estructura de los archivos fuentes de datos.....	55
Fig. 17. Mapeo fuentes de datos.....	56
Fig. 18. Modelo conceptual ampliado	59
Fig. 19. Perspectiva clientes	61
Fig. 20. Perspectiva producto	61
Fig. 21. Perspectiva local	62
Fig. 22. Perspectiva tiempo.....	63
Fig. 23. Tabla de Hechos	64
Fig. 24. Uniones	64
Fig. 25. Esquema de carga inicial	65
Fig. 26. Fuente de datos de fechas, parte 1	65
Fig. 27. Fuente de datos de fechas, parte 2	66
Fig. 28. Sentencia SQL para cargar la dimensión de fechas	66
Fig. 29. Fuente de datos de los locales	66
Fig. 30. Sentencia SQL para cargar la dimensión de los locales.....	67
Fig. 31. Fuente de datos de los clientes	67
Fig. 32. Sentencia SQL para carga la dimensión de los clientes.....	67
Fig. 33. Fuente de datos de los productos	68
Fig. 34. Sentencia SQL para cargar la dimensión de los productos.....	68
Fig. 35. Estructura de hoja cabecera	69
Fig. 36. Productos por facturas	69
Fig. 37. Función que deja solamente el nombre del producto	70

Fig. 38. Códigos de productos eliminados.....	70
Fig. 39. Función que coloca el nombre del producto en cada factura	71
Fig. 40. Productos por facturas	71
Fig. 41. Datos de la tabla de hechos para consumidor final	72
Fig. 42. Datos de la tabla de hechos para demás clientes.....	72
Fig. 43. Sentencia SQL para cargar la tabla de hechos.....	72
Fig. 44. Conexión base de datos PostgreSQL con Power BI.....	73
Fig. 45. Data Warehouse – Modelo de datos en Power BI	74
Fig. 46. Medidas para ventas por producto	74
Fig. 47. Medida para calcular el número de facturas mayores a \$100	75
Fig. 48. Medidas para calcular la rentabilidad por producto.....	75
Fig. 49. Productos seleccionados para el algoritmo A priori	78
Fig. 50. Escala de aceptación SUS	79
Fig. 51. Pantalla principal – Sistema BI	81
Fig. 52. Ventas mensuales, por local, por producto y cantidad de facturas	82
Fig. 53. Ventas por cliente, ciudad, cantidad de clientes por local	82
Fig. 54. Rentabilidad por producto.....	83
Fig. 55. Resultados algoritmo A Priori	83
Fig. 56. Resultados recopilados por participante	84
Fig. 57. Resultados preguntas pares e impares.....	84
Fig. 58. Resultados cuestionario SUS, por participante y promedio.....	85

ÍNDICE DE TABLAS

Tabla 1. Actividades referentes para evaluar la Norma ISO/IEC 25012.....	27
Tabla 2. Comparación metodologías BI.....	39
Tabla 4. Requerimientos e indicadores “Sari Popular”	45
Tabla 5. Requisitos de evaluación, característica de Exactitud	46
Tabla 6. Resumen criterios de aceptación, característica de Exactitud	47
Tabla 7. Resumen criterios de aceptación, característica de Consistencia.....	47
Tabla 8. Diseño de la evaluación	47
Tabla 9. Ejecución de la Evaluación	48
Tabla 10. Valores obtenidos de la evaluación final	48
Tabla 11. Reestructuración de Preguntas del Negocio	50
Tabla 12. Perspectivas e indicadores	52
Tabla 13. Granularidad de la perspectiva clientes	57
Tabla 14. Granularidad de la perspectiva tiempo	58
Tabla 15. Granularidad de la perspectiva local	58
Tabla 16. Granularidad de la perspectiva producto.....	59
Tabla 18. Pantallas sistema BI y preguntas del negocio.....	81
Tabla 19. Principales hallazgos	84

RESUMEN

El presente proyecto se enfoca en el tema de inteligencia de negocios (en inglés “Business Intelligence”), esta disciplina cuenta con bases y objetivos centrados en analizar la información generada por las diferentes empresas, comprenderla y generar conocimiento relevante para apoyar en la toma de decisiones informadas (basadas en datos).

En el caso práctico, la empresa “Sari Popular”, perteneciente a la comercialización xce productos de papelería, ha brindado la oportunidad de manejar sus datos de ventas con fines académicos. Cuenta con asesores comerciales y demás colaboradores del departamento de ventas, por lo cual han sido los principales usuarios en los que se ha basado el proyecto. La información que fluye en la empresa es gestionada a través de un sistema llamado TINI. Por lo cual se pudieron obtener adecuadas fuentes de datos.

En este sentido, el proyecto se presenta como una herramienta adicional con la cual se pretende atender los requerimientos de información más importantes y ser implementado como un recurso estratégico, además que permite tener al alcance los indicadores necesarios para el de apoyo para la toma de decisiones.

Para efectos de la investigación se realizaron revisiones bibliográficas de los principales conceptos de Inteligencia de Negocios, metodologías para llevar a cabo un proyecto, además de estándares para controlar la calidad de datos como lo es la norma ISO/IEC 25012. Durante el proceso se implementaron diferentes herramientas/software, de los cuales principalmente para el resultado final se basa en Power BI, Python, PostgreSQL, las hojas de cálculo como Excel, Google Sheets. La metodología para realizar el proyecto es Hefesto.

Palabras Claves: Inteligencia de Negocios, metodología Hefesto, Power BI, Calidad de datos, ISO/IEC 25012.

ABSTRACT

This project focuses on the topic of business intelligence (Business Intelligence), this discipline has bases and objectives focused on analyzing the information generated by different companies, understanding it and generating relevant knowledge to support informed decision making (based on data).

In the practical case, the company "Sari Popular", belonging to the commercialization of stationery products, has provided the opportunity to manage its sales data for academic purposes. It has commercial advisors and other collaborators in the sales department, so they have been the main users on which the project has been based. The information flowing in the company is managed through a system called TINI. Therefore, it was possible to obtain adequate data sources.

In this sense, the project is presented as an additional tool with which it is intended to meet the most important information requirements and to be implemented as a strategic resource, in addition to having the necessary indicators available to support decision making.

For the purposes of the research, bibliographic reviews of the main concepts of Business Intelligence, methodologies to carry out a project, as well as standards to control data quality such as ISO/IEC 25012 were carried out. During the process different tools/software were implemented, of which mainly for the final result is based on Power BI, Python, PostgreSQL, spreadsheets such as Excel, Google Sheets. The methodology to carry out the project is Hephaestus.

Keywords: Business Intelligence, Hephaestus methodology, Power BI, Data Quality, ISO/IEC 25012.

INTRODUCCIÓN

Tema

Implementación de Inteligencia de Negocios para apoyar en la toma de decisiones informadas de la empresa "Sari Popular".

Problema

Antecedentes

La toma de decisiones en una empresa es de mucha importancia, puesto que las estrategias que se planteen tendrán directamente relación con los resultados de cada proyecto que se realice.

La toma de decisiones informadas por su parte, comprende, el tomar dichas decisiones en base a datos, hechos y métricas, lo que generará información de valor y permitirá guiar a los ejecutivos a tomar mejores decisiones alineadas con las metas y objetivos de cada organización.

La inteligencia de negocios surge como un medio para apoyar en la toma de decisiones de una empresa, aunque aparentemente este proceso se lo ha realizado en los últimos años, no lo es, tal vez sea por la tendencia de los datos, el auge de nuevas profesiones en torno a datos, que se da a conocer mayormente. Sin embargo, como menciona (Torres 2018) el Business Intelligence (BI) ya se ha estado usando desde finales de los años 80' para interpretar los datos y obtener información relevante.

Existen empresas que toman sus decisiones basadas en datos, otras están proceso y otras aún no lo hacen, todavía se manejan de acuerdo a intuición o experiencia personal, por lo que se tiene la incertidumbre presente, BI da soporte para garantizar resultados, como enuncian (Ayala and Ortiz 2017), esto ha permitido ganar ventaja competitiva, y llevar un mejor control del negocio para las empresas que lo aplican.

Situación Actual

La empresa "Papelería y Librería Sari Popular", se dedica a la distribución y comercialización de útiles escolares, artículos de oficina, bazar, arte, papelería, librería y tecnología, manejan un sistema de gestión empresarial TINÍ, el mismo que se compone de algunos módulos, entre ellos cuenta con Cubos OLAP (On-Line Analytical Processing), para analizar los datos desde el sistema mismo, aunque este proceso no se realiza.

Prospectiva

Teniendo en cuenta lo mencionado anteriormente, se ve la necesidad de fomentar una cultura de datos para tomar decisiones, a través de la implementación de un proceso de inteligencia de negocios como primera instancia, y avanzar en la adecuación de este cambio de esquema.

Planteamiento del problema

La empresa tiene una baja aplicabilidad de cultura de datos en su modelo organizacional, los datos que genera no están siendo aprovechados para extraer información y conocimiento importante, limitando así la visión de la situación actual del negocio. El sistema contable que se usa en la empresa, cuenta con la funcionalidad de cubos OLAP, pero no se toma en cuenta para la toma de decisiones, por lo que el desconocimiento de la importancia de fomentar una cultura basada en datos, y procesos de analítica podría afectar en la competitividad de la empresa. El problema puede verse representado gráficamente en el siguiente diagrama.

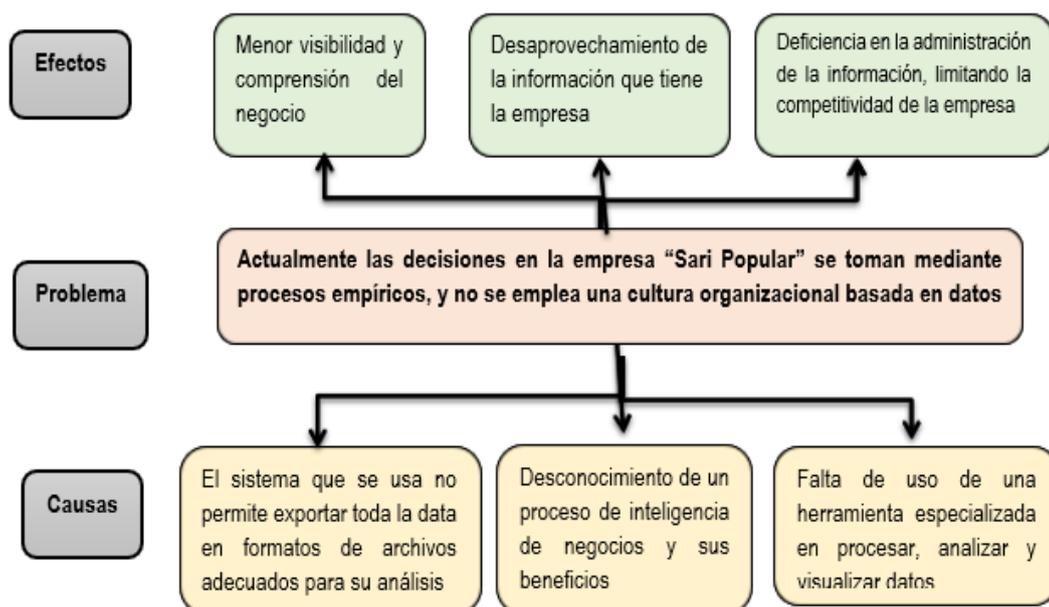


Fig. 1. Diagrama del Alcance

Objetivos

Objetivo General

Implementar una solución de Inteligencia de Negocios para apoyar en la toma de decisiones informadas de la empresa "Sari Popular".

Objetivos Específicos

- Elaborar un marco teórico respecto a calidad de datos, técnicas y herramientas de inteligencia de negocios.
- Implementar una metodología de inteligencia de negocios y buenas prácticas.
- Validar los resultados obtenidos con la herramienta de inteligencia de negocios y las preguntas del negocio planteadas.

Alcance

El presente proyecto tiene como finalidad integrar y procesar la información que genera la empresa “Sari Popular”, para desarrollar un entorno de inteligencia de negocios con componentes que permitirán realizar reportes, la construcción de un dashboard y minería de datos.

Debido a que se obtendrá una mejor visualización de los datos históricos, el movimiento de los artículos, encontrar asociaciones entre artículos vendidos y preguntas del negocio que sean necesarias para entender su situación actual y empezar a tomar mejores decisiones en el negocio.

Como se puede visualizar en la imagen de la arquitectura del proyecto, las fuentes de datos parten del sistema TINI, ya que es el sistema en funcionamiento actualmente, conjuntamente con su base de datos y los archivos exportados en formato “.xlsx”.

Posteriormente se consolidará un data mart/warehouse con los datos pasados por el proceso de extracción, transformación y carga, en inglés ETL. Este data mart/warehouse proveerá la información más rápidamente para generar reportes, dashboard y minería de datos en los artículos.

De las herramientas a destacar se encuentran Power BI para el proceso ETL y las visualizaciones, WEKA permitirá ejecutar algoritmos de minería de datos, sin embargo, cabe mencionar que estas son las herramientas idealizadas inicialmente en el proyecto, sin descartar que en el desarrollo puedan agregarse otras herramientas con un desempeño puntual en alguna fase.

En cuanto a buenas prácticas, se seleccionará una metodología de inteligencia de negocios adecuada de acuerdo a la literatura adicionalmente, se hará uso del estándar ISO/IEC 25012, el mismo que está enfocado en la calidad de los datos, podría darse el caso de trabajar con datos defectuosos, nulos o cualquier otro factor adverso que afectaría

considerablemente a la eficiencia del desarrollo del proyecto. El alcance se puede ver representado gráficamente, mediante el siguiente diagrama.

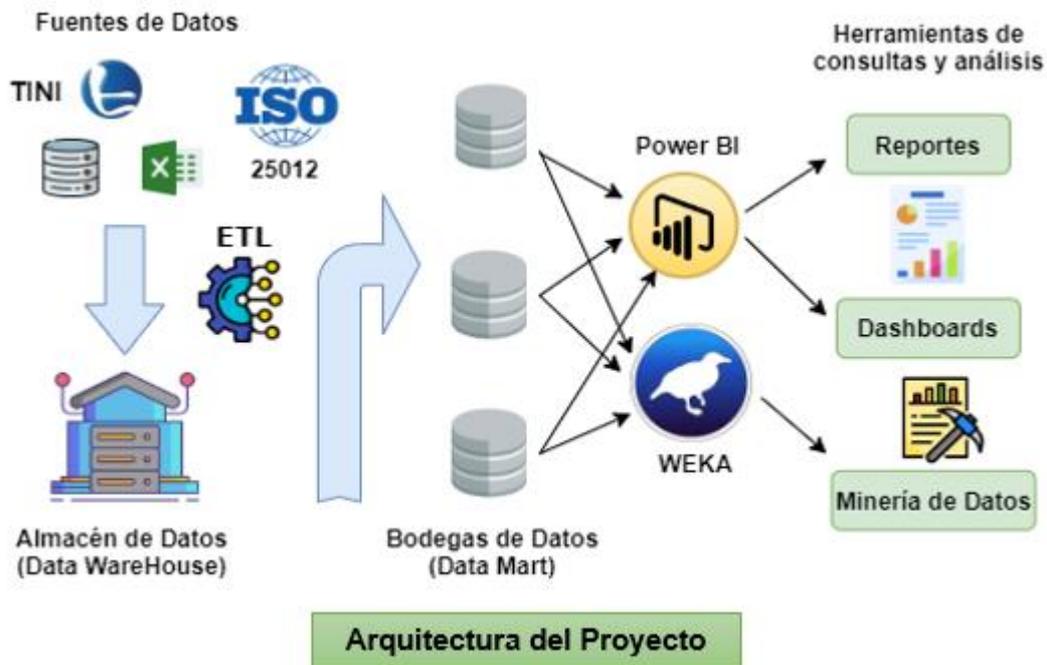


Fig. 2. Diagrama del Alcance

Justificación

En la actualidad las empresas están haciendo uso de los datos que generan, para extraer información, conocimiento valioso que les permita entender la situación actual de su negocio, conocer a sus clientes, su rendimiento, de igual forma, es importante para llegar a tener ventaja competitiva.

La realización de un proceso de inteligencia de negocios permitirá apoyar a la empresa a organizar su información, y tenerla disponible al instante para tomar decisiones estratégicas del negocio. Basándose al objetivo 9 "Industria, Innovación e Infraestructuras" de los Objetivos de Desarrollo Sostenible (ODS), el proyecto encaja en este contexto, puesto que promueve el uso de nuevas tecnologías para hacer uso eficiente de los recursos.

En el mismo contexto de este objetivo da a conocer que las tecnologías han estado en la primera línea como respuesta a la COVID-19, la pandemia impulsó la digitalización y las empresas tuvieron que modificar su modelo de negocio para mantenerse en este tiempo, por

lo que una meta que plantea alcanzar es fomentar la cultura de datos como innovación en el modelo organizacional de la empresa.

Este proyecto también se relaciona con el objetivo 8, “Trabajo decente y crecimiento económico”, ya que se presenta como un trabajo práctico de implementación de inteligencia de negocios en una empresa de un sector donde no es tan comúnmente conocido, como lo es la industria papelera. Además, que pretende alcanzar la meta de desarrollar destrezas y promover servicios profesionales que impulsen un crecimiento económico inclusivo en la sociedad.

Justificación tecnológica

Construir un data mart/warehouse que integre la información necesaria, a través de un proceso previo de extracción y transformación apoyado del estándar ISO/IEC 25012 para la calidad de los datos. Posteriormente dicha información servirá para proveerla rápidamente, y mostrarla visualmente en un dashboard, con el fin de ser un recurso de apoyo para la toma de decisiones estratégicas.

CAPÍTULO 1

Marco Teórico

1.1 Empresa “Sari Popular”

La historia de “Sari Popular” en un breve resumen es relatada o por (Sahona, 2014), donde informa:

Papelería Popular dirigida por la Sra. Sara Rodríguez quien empezó sus actividades económicas desde el 01 de junio del año de 1999. Decidió convertir a papelería popular en una sociedad en el año 2012, es entonces que “Sari Papelería Popular S.A. empieza a realizar sus primeras transacciones comerciales en el mercado en la (Venta Al Por Mayor y Menor de Artículos de Papelería) desde el 6 de marzo del 2012, con el acta de constitución de la compañía, otorgada en la ciudad de Quito y legalizada en el cantón Ibarra. En la ciudad de Otavalo el 5 de diciembre del dos mil once ante el abogado José Fabián Simbaña Ayabaca Notario Primero del Cantón Otavalo; se eleva a escritura pública la Constitución de La compañía “SARI PAPELERÍA POPULAR S.A.” Esta institución se rige bajo la base legal de la Superintendencia de Compañías. (p.100).

De tal manera que se conoce los inicios de la empresa. Años más tarde se conoce la realidad de la empresa más actualizada, redactado en el trabajo investigativo por parte de (Parra, 2020) expone lo siguiente:

Sari Papelería Popular de la ciudad de Ibarra, que inicia sus actividades comerciales en el año 2012, dedicándose a la venta de artículos de papelería. Constituida por dos socios, se encuentra en el mercado local por más de 8 años, ofreciendo artículos de papelería al por mayor y menor convirtiéndose en una de las empresas más grandes de la ciudad de Ibarra. (p.11).

En la página web (Derecho Ecuador, 2013) se encuentra establecido lo que es una sociedad anónima en Ecuador, señalando que, la sociedad anónima es una compañía cuyo capital, dividido en acciones negociables, está formado por la aportación de los accionistas que responden únicamente hasta el monto de sus acciones.

Es importante conocer la estructura organizacional de la empresa, ya que permite saber quiénes son los usuarios finales del proyecto, una vez obtenido la información de la estructura se elaboró el siguiente diagrama representativo.

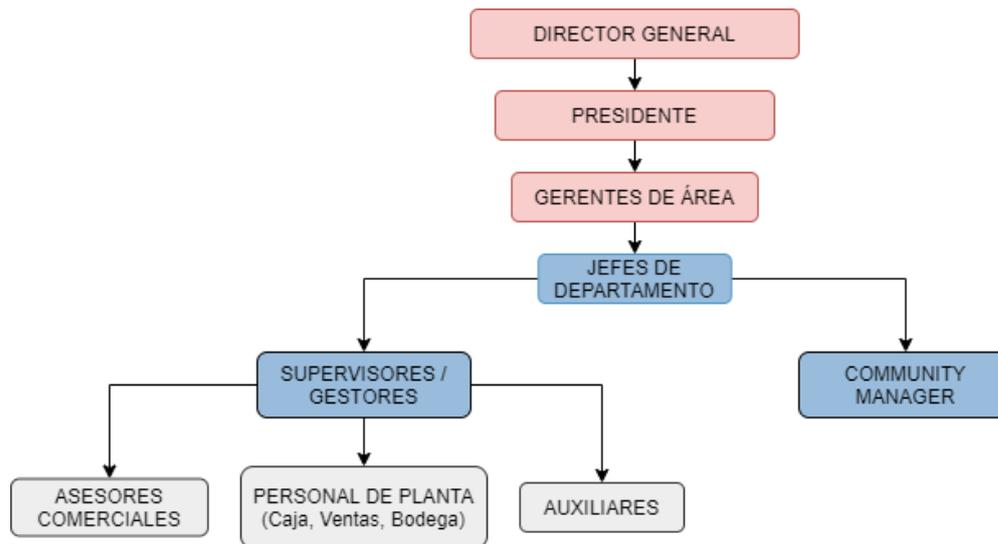


Fig. 3. Estructura Organizacional – Sari Popular

El contenido descrito, permite entender un poco más el contexto de la empresa “Sari Popular”, en la cual se llevó a cabo el proyecto.

1.2 Inteligencia de Negocios

1.2.1 Calidad de datos

Es entendible reconocer la importancia de los datos en proyectos Business Intelligence (BI), son un elemento fundamental, llegándose a conocer como los activos más preciados. Sari Papelería Popular de la ciudad de Ibarra, que inicia sus actividades comerciales en el año 2012, dedicándose a la venta de artículos de papelería. La empresa Sari Papelería Popular S.A. constituida por dos socios, se encuentra en el mercado local por más de 8 años, ofreciendo artículos de papelería al por mayor y menor convirtiéndose en una de las empresas más grandes de la ciudad de Ibarra.

(Logicalis, 2021) ha afirmado que el papel de los datos, más que importante, es absolutamente fundamental: en los datos reside la calve del éxito de cualquier proyecto BI. Sin embargo, y aunque esto sigue siendo así, a día de hoy la cantidad de fuentes de origen que aportan incesantemente nuevos datos a las bases o data warehouses corporativas es tan grande, y tan ingente el volumen y la variedad de estos mismos datos, que es necesario tomar en consideración otras magnitudes como la calidad, a parte de la cantidad, para que un proyecto BI ofrezca los resultados esperados: principalmente, un apoyo sólido y fiable a la toma de decisiones.

El mismo hecho que la cantidad de datos es de volumen grande se piensa en la inseguridad de los mismos, es decir, se tiene la idea de que no todos los datos son fiables. Con este contexto (Logicalis, 2021) menciona 3 factores básicos para reconocer la calidad en los datos que se esté por manipular:

- Integridad: los datos deben ser completos, eviten duplicidades y dispongan las medidas necesarias para evitar cruces e interferencias.
- Operatividad: tienen que ser lo suficientemente homogéneos, sólidos y consistentes para permitir una explotación adecuada de los mismos.
- Veracidad: que dispongan de un comprobado valor referencial.

Las fuentes de datos deben ser verificadas referente a la calidad, ya que si no se previene esta situación las consecuencias pueden dirigirse al fracaso de la implementación de BI, que incluso podría afectar a los usuarios finales desencadenando en la desconfianza de las aplicaciones de inteligencia de negocios que se desarrollen.

El conocimiento es "la información en su contexto para producir una comprensión procesable". Es el remedio más apreciado para la complejidad y la incertidumbre, es un nivel superior de abstracción que reside en la mente de las personas. Es más amplio y mucho más difícil de captar que los datos o la información. En el contexto de la gestión del conocimiento, el conocimiento es "la comprensión humana de un campo de interés especializado que se ha adquirido mediante el estudio y la experiencia".

Con lo descrito anteriormente se puede generalizar que existe un sentido de derivación, partiendo desde el conocimiento hacia la información y los datos finalmente. En la siguiente figura de una manera visual representa el contexto descrito.



Fig. 4. Datos, Información, Conocimiento
Fuente: Adaptado de (Chadha, 2021)

Es importante tener en cuenta los tres conceptos anteriormente mencionados, puesto que son la base para entender lo que son las decisiones informadas, mismas que contribuyen a conseguir el objetivo principal de la inteligencia de negocios, que es mejorar la toma de decisiones de una organización.

Referente a la calidad de los datos, las normas ISO/IEC 25000 forman una serie enfocada a este tema, (Fernández Sáenz, 2018) agrega que es conocida como SQuaRE (Systems and Software Quality Requirements and Evaluation, en inglés) y fue creada con el objetivo de mejorar y unificar las normas ya existentes relacionadas con las ISO/IEC 9126 e ISO/IEC 14598, es decir, el proceso de especificación de requisitos de calidad de software y el proceso de evaluación de calidad de software. A continuación, en la figura 1, se muestra la división respectiva.



Fig. 5. Divisiones Norma SQuaRE

Fuente: Tomado de (ISO 25000, 2021)

(Fernández Sáenz, 2018) señala los conceptos del significado de cada división:

- **División de Gestión de la Calidad:** Define todos los estándares que tienen en común todas las normas en la serie SQuaRE.
- **División de Modelo de Calidad:** Define un modelo de calidad incluyendo características para calidad externa, interna y en uso. También provee sub características y guía en el uso del modelo.
- **División de Medición de la Calidad:** Define un modelo de referencia para la medición de la calidad de producto de software, así como definiciones matemáticas de medidas de calidad.
- **División de Requerimientos de Calidad:** Ayuda a la especificación de requerimientos de calidad.

- **División de Evaluación de la Calidad:** Aporta requerimientos, recomendaciones y guía para la evaluación del producto de software.

1.2.1.1 ISO/IEC 25012

La Norma ISO/IEC 25012 corresponde a la división de modelos de calidad, en la cual define dos puntos específicos respecto a la calidad de datos. En el artículo (Calabrese et al., 2019) se explica más a detalle las categorías en cuestión:

- **Calidad de Datos Inherente:** Grado con el que las características tienen el potencial de satisfacer las necesidades establecidas y necesarias cuando los datos son utilizados bajo condiciones específicas. Desde el punto de vista inherente, la calidad de datos hace referencia a: x Valores de datos para el dominio y sus posibles restricciones (ej., Reglas de negocio con la calidad requerida por las características en una aplicación) x Relaciones entre valores de datos (ej., Consistencia) x Metadatos (ej., Autor, Formato, etc.).
- **Calidad de Datos Dependiente del Sistema:** Grado con el que la calidad de datos es alcanzada y preservada a través de un sistema informático cuando los datos son utilizados bajo condiciones específicas. Desde este punto de vista, la calidad de datos depende del dominio tecnológico en el que los datos se utilizan, y se alcanza mediante las capacidades de los componentes del sistema informático tales como: dispositivos hardware (ej., respaldo para alcanzar la Recuperabilidad) y software (ej., herramientas de migración para alcanzar la Portabilidad). Este punto de vista suele ser responsabilidad de los técnicos del sistema.

A continuación, en la siguiente figura se muestran las características de la Norma ISO/IEC 25012 con sus respectivos conceptos

Característica	Descripción
Exactitud	El grado en el cual el dato tiene atributos que representan el valor correcto de un concepto o evento en un contexto específico de uso.
Compleitud	El grado en el cual el dato asociado a una entidad tiene valores para todos los atributos esperados e instancias de entidad relacionadas, de acuerdo a un contexto específico de uso.
Consistencia	El grado en el cual el dato tiene atributos libres de contradicción y son coherentes con otros datos en un contexto específico de uso.
Credibilidad	El grado en el cual el dato tiene atributos considerados como verdaderos y creíbles por usuarios en un contexto específico de uso.

Característica	Descripción
Actualidad	El grado en el cual el dato tiene los atributos que son del período correcto en un contexto específico de uso.
Accesibilidad	El grado en el cual se puede acceder al dato en un contexto específico de uso.
Conformidad	El grado en el cual el dato tiene atributos que se adhieren a normas, convenciones o regulaciones vigentes y reglas relacionadas con la calidad de datos en un contexto específico de uso.
Confidencialidad	El grado en el cual el dato tiene los atributos que aseguran que éste es sólo accesible e interpretable por usuarios autorizados en un contexto específico de uso. El grado en el cual el dato tiene los atributos que aseguran que éste es sólo accesible e interpretable por usuarios autorizados en un contexto específico de uso.
Eficiencia	El grado en el cual el dato tiene los atributos que pueden ser procesados y proporciona los niveles esperados de funcionamiento (desempeño) usando las cantidades y los tipos de recursos apropiados en un contexto específico de uso.
Precisión	El grado en el cual el dato tiene atributos que son exactos o que proporcionan su discriminación en un contexto específico de uso.
Trazabilidad	El grado en el cual el dato tiene atributos que proporcionan un rastro de auditoría de acceso a los datos y de cualquier cambio hecho a los datos en un contexto específico de uso.
Comprensibilidad	El grado en el cual el dato tiene atributos que le permiten ser leído e interpretado por usuarios, y es expresado en lenguajes apropiados, símbolos y unidades en un contexto específico de uso.
Disponibilidad	El grado en el cual el dato tiene atributos que le permiten ser recuperados por usuarios autorizados y/o aplicaciones en un contexto específico de uso.
Portabilidad	El grado en el cual los datos tienen atributos que les permiten ser instalados, substituidos o movidos de un sistema a otro conservando la calidad existente, en un contexto específico de uso.
Recuperabilidad	El grado en el cual el dato tiene atributos que le permiten mantener y conservar un nivel especificado de operaciones y calidad, aún en caso de falla, en un contexto específico.

Fig. 6. Características y conceptos ISO/IEC 25012

Fuente: Adaptado de (González Pinzón & González Sanabria, 2013)

Entendiendo los conceptos, se ha tomado en cuenta la guía de evaluación desarrollada por (Calabrese et al., 2019), y a continuación, en la siguiente tabla se muestran las actividades.

Tabla 1. Actividades referentes para evaluar la Norma ISO/IEC 25012

Evaluación Normas ISO/IEC 25012	
Actividad 1	Establecer los requisitos de la evaluación
	Esta actividad consiste en establecer el propósito de la evaluación, identificando las partes interesadas en el producto, los riesgos posibles si los hubiese y el modelo de calidad a utilizar.
Actividad 2	Especificar la evaluación
	Dentro de esta actividad se especifican los módulos de evaluación (métricas, herramientas y técnicas) junto con los criterios de decisión a aplicar.
Actividad 3	Diseñar la evaluación
	En esta actividad se define el plan con las tareas que se deben realizar en la evaluación.
Actividad 4	Ejecutar la evaluación
	Actividad destinada a la ejecución de la evaluación, obteniendo las métricas de calidad y aplicando los criterios de decisión.
Actividad 5	Concluir la evaluación
	En esta última actividad se culmina la evaluación de la calidad, realizando un informe de resultados finales y conclusiones en base a los valores obtenidos.

Fuente: Adaptado de (Calabrese et al., 2019)

Es evidente la importancia de evaluar la calidad de datos en el desarrollo de un proyecto de BI, por ello existen guías y normas para llevar a cabo este proceso de mejor manera.

(Chadha, 2021) ha afirmado lo siguiente:

Los datos son sólo bits y números, hechos y cifras. Son discretos, autónomos y, aislados, no tienen ningún significado. Los datos son el valor de un atributo medible o calculable. También son hechos no organizados y no procesados. Son estáticos y un requisito previo a la información.

La palabra información deriva de informar, que significa "dar forma a". La información significa dar forma a los datos. También son hechos y cifras basadas en datos reformulados o procesados. A diferencia de los datos, la información significa entender las relaciones. Tiene significado, propósito y relevancia. Tiene forma porque está organizada por una razón, su enfoque es cualitativo.

1.2.2 Historia BI

La información en la actualidad ha tomado mayormente relevancia en el ámbito empresarial, generando conocimiento que las empresas usan para tomar mejores “decisiones informadas”, en el contexto se entenderían como decisiones basadas en información. El conocimiento obtenido posteriormente se convierte en planes de acción a realizar, por ello la información se ha convertido en un apoyo fundamental para gestionar sus recursos, optimizarlos y principalmente aumentar los ingresos.

La inteligencia de negocios se fundamenta en la recolección y procesamiento de datos, permitiendo producir información que sirve como recurso estratégico para validar las intuiciones que los líderes o responsables de las empresas tienen relacionado a la toma de decisiones. Podría considerarse que este principio se ha mantenido a lo largo de la historia.

Un contexto importante sobre la historia de BI es el que manifiesta (Scheps, 2008) por medio de su libro dando a conocer que los primeros ordenadores eran máquinas de tabulación, diseñadas y construidas para realizar cálculos puntuales, pero los científicos e inventores desarrollaron la capacidad de almacenamiento de información casi a la par que el crecimiento de la potencia de cálculo. A partir de los años 40's, ambas tecnologías se disparan, el almacenamiento masivo comenzó a tomar forma cuando se utilizaron las propiedades de la cinta magnética para almacenar patrones de información.

Evidentemente esto generaría montañas de datos almacenados literalmente, y para gestionar a los mismos, los programadores desarrollaron Sistemas de Gestión de Bases de Datos (SGBD) de creciente potencia y complejidad. En base a la redacción de (Pedraza, 2018) y (Signaturit, 2021), se resumen los siguientes acontecimientos entorno a la historia

- Década de 1970: SAP y otras empresas desarrollan aplicaciones de negocio generando los primeros modelos de bases de datos relacionales como respuesta a la creciente de la información.
- Década de 1980: Se consolida una revolución empresarial popularizando aún más las bases de datos y la estandarización del lenguaje SQL, sin embargo, se empezaron a notar los fracasos de los proyectos que intentaban implementar Data Warehouses o almacenes de datos, esto debido a que se los desarrollaba en base a los sistemas de bases de datos orientadas a transacciones que existían.
- Década de 1990: Conjuntamente Inmon y Kimball establecen guías para la correcta creación de DWs. Esta década se llega a considerar como Business

Intelligence 1.0. Se desarrollan y comercializan más herramientas BI, aunque no eran totalmente satisfactorias ya que tienen limitaciones importantes como la cantidad de datos que pueden analizar, además de no ser tan fáciles de usar.

- Año 2000 en adelante: Esta etapa es conocida como BI 2.0 conjuntamente con Social BI. Donde BI 2.0 hace referencia a que los usuarios finales pueden generar y compartir contenido desde las herramientas de análisis y Social BI se relaciona con la capacidad de integrar datos externos de otros sistemas permitiendo obtener más información estructurada y no estructurada.

El contexto de la historia descrito previamente puede ser tomado en cuenta como una pauta para comprender la necesidad de trabajar y aprovechar toda esa información que se conoce que existe. Por ello en la actualidad, BI sigue evolucionando con más herramientas y la consolidación de complementos como Big Data, Minería de Datos, Inteligencia Artificial permiten generar estrategias para las empresas que conocen los beneficios que pueden obtener, evidentemente que cada vez más dichas empresas se suman a esta industria.

1.2.3 Definición

Según (Bernabeu R. Dario, 2010) se puede describir el Business Intelligence (BI - Inteligencia de Negocios), como la actividad de: almacenar y procesar grandes cantidades de datos, para que, mediante la utilización de herramientas de software especializadas, sea sencillo el análisis y exploración de dichos datos, con el principal objetivo de obtener conocimiento (knowledge) orientado a tomar decisiones en tiempo real.

(Quimbia, 2017) menciona que se entiende por Business Intelligence al conjunto de metodologías, aplicaciones, prácticas y capacidades enfocadas a la creación y administración de información que permite tomar mejores decisiones a los usuarios de una organización.

El Business Intelligence (BI - por sus siglas en ingles), es una solución que se compone por una plataforma tecnológica que integra todas las fuentes de información tanto internas como externas de la organización para que se vean como una sola y luego ser visualizadas mediante herramientas de reporte, así como también de las preguntas claves del negocio y quizá lo más importante, la inteligencia analítica del usuario. (Agüero, 2019).

La definición por parte de (Andrea & Rojas, 2017) detalla que BI es una disciplina que integra información proveniente de diversas fuentes u orígenes aportando facilidad en los procesos de almacenamiento, abordaje, selección, y el tratamiento de datos históricos. Así mismo el trabajo realizado por (Guevara et al., 2016), tiene similitud además describe que BI no se considera como una metodología, ni un software/sistema, sino, se trata de la relación

de diferentes herramientas con fines de análisis de información para dar soporte en toma de decisiones empresariales.

De acuerdo a las definiciones descritas previamente, se podría contextualizar a la Inteligencia de Negocios como, un término en general que abarca; procesos y métodos de recopilación, almacenamiento y análisis de datos históricos con el fin de optimizar el rendimiento de las diferentes actividades que se llevan a cabo en una organización mediante la toma de decisiones.

1.2.4 Importancia

La Inteligencia de Negocios puede brindar variedad de beneficios a las organizaciones que la implementan, puesto que la información obtenida como resultado, genera conocimiento sobre las ventas, gestión interna de la empresa y demás, pero, sobre todo el entendimiento de la relación con el cliente.

También se pueden mencionar otras ventajas de acuerdo con lo que describe (Andrea & Rojas, 2017):

- Las empresas pueden identificar los clientes rentables y las razones de la lealtad de estos clientes, así como identificar futuros clientes potenciales.
- Analizar datos de flujo de clics para mejorar las estrategias de comercio electrónico.
- Determinar qué combinaciones de productos y líneas de servicio comprar y cuándo.
- Detectar y disuadir conductas fraudulentas, como los picos de uso cuando las tarjetas de crédito son robadas.

Para la empresa, centrarse en mejorar la atención a los clientes, es realizar un ejercicio de fortalecimiento del conocimiento y establecimiento de medios de comunicación muy fluidos para que este centro de servicios esté siempre informado y al tanto de las operaciones del servicio. (Santillán Suquitana, 2018).

Es importante mencionar que la inteligencia de negocios es de utilidad debido al control de información y conocimiento que se puede generar. La información es un activo muy importante en toda organización y es un aspecto clave para obtener una ventaja competitiva en el mundo de los negocios. Pero se requiere que el personal responsable de tomar decisiones tenga fácil y rápido acceso a la información útil y acorde con el contexto de la organización. (Ayala & Ortiz, 2017).

De manera general se conoce que la inteligencia de negocios o BI permite transformar datos en información útil, para tomar decisiones basadas en información objetiva. De acuerdo con (Actuaria, 2021), describe algunos beneficios más detallados que evidencian la importancia de BI, los mismos que a continuación se muestran:

- Reducción significativa de recursos en la elaboración de reportes: es posible automatizar los procesos de consolidación de información reduciendo la cantidad de tiempo que el personal dedica para la creación y actualización de reportes.
- Consolidación automática de información de múltiples fuentes: La información de distintos sistemas o archivos es consolidada con un sentido de negocio, lo que permite contar con un panorama más amplio.
- Generación de información a medida, según las necesidades del negocio: Una solución integral de inteligencia de negocios empodera al analista para que pueda obtener información sin tener que esperar a que las áreas de informática o tecnología de la información realicen cambios en las herramientas.
- Control oportuno y optimización de gestión de diversas operaciones: Los procesos de consolidación de datos, al ejecutarse automáticamente, permiten la actualización de información de manera más frecuente, incluso a tiempo real.
- Calidad de información: Al minimizar el trabajo manual de personas se evita una cantidad importante de errores.
- Información fácil de comprender y que permite hacer descubrimientos: Datos expresados en dashboards o cuadros de mando visuales, con gráficas fáciles de usar e interpretar, y con capacidades de filtrado múltiple de información mediante diferentes dimensiones, lo que permite ir mucho más allá de simples tablas.
- Puente hacia la analítica avanzada: Con la implementación de soluciones integrales de inteligencia de negocios, se prepara el ambiente para el siguiente paso que es el análisis avanzado de información. El machine learning o inteligencia artificial permiten ir más allá del análisis descriptivo de la información o de lo ocurrido en el pasado, para tratar de predecir lo que podría ocurrir en el futuro y automatizar aún más ciertos procesos de la organización.

1.3 Herramientas BI y Data Mining

1.3.1 BI

Las herramientas BI, responden a su término en inglés Business Intelligence y están relacionadas con los datos que generan las compañías. Gracias a estas herramientas las empresas pueden obtener información que posteriormente permitirán generar informes para la toma de decisiones que irán en beneficio del negocio. Existe un sinnúmero de herramientas BI, dentro del mercado las más relevantes son:

- Plataformas y herramientas empresariales.
- Software de BI.

Las plataformas de Business Intelligence se utilizan principalmente para el análisis de otras aplicaciones que suelen funcionar a partir de cálculos muy complejos; pero, además, estas plataformas se usan para crear aplicaciones más sencillas a las que los usuarios pueden acceder rápidamente.

QlikView es una plataforma enfocada al análisis visual de datos y aplicaciones interactivas que tiene por objetivo mejorar el proceso de acceso a los datos de cara al usuario. Como, por ejemplo, acceder a ciertas visualizaciones 'limpias' y fáciles de comprender, diseños de gráficos llamativos, entre otros.

Microsoft Power BI: se destaca, además, por su extraordinaria integración con otras herramientas de la propia marca. Desde el trabajo mediante voz con Cortana hasta datos alojados en Azure, pasando por SQL Server o Excel, donde podremos ampliar los gráficos y diagramas habituales de este programa de una forma masiva.

Tableau: con esta herramienta pueden crearse dashboards visuales y responder preguntas de negocio de forma sencilla, agrupando datos de distintas fuentes y cargando descripciones emergentes con información adicional, excluir valores atípicos o visualizando datos en varias dimensiones geográficas al mismo tiempo.

Pentaho: Con el procesamiento de datos en tiempo real, Pentaho puede ser una gran opción como herramienta de análisis para el back-end. Destacando como una herramienta de código abierto, puede proporcionar una gran flexibilidad y recientemente se convirtió en una nueva empresa comercial que podría producir soluciones aún más robustas.

A continuación, en la posterior figura se visualiza una comparación entre estas herramientas.

<i>QLIK</i>	<i>TABLEAU</i>	<i>POWER BI</i>
DESCRIPCIÓN: <ul style="list-style-type: none"> • Es una plataforma enfocada al análisis visual de datos y aplicaciones interactivas. • Mejora el proceso de acceso a los datos de cara al usuario 	DESCRIPCIÓN: <ul style="list-style-type: none"> • Es una herramienta BI interactiva que sirve para la visualización interactiva de los datos. • Los usuarios pueden interactuar: comparando datos, filtrándolos o creando una conexión entre unas variables y otras. 	DESCRIPCIÓN: <ul style="list-style-type: none"> • Es una herramienta BI con servicio en la nube, con fácil implementación. • Permite subir, compartir y tener acceso a informes desde cualquier dispositivo ya sea ordenador, Tablet o Smartphone. • Se puede acceder a orígenes de datos locales, de bases de datos y servicio en la nube.
VENTAJAS: <ul style="list-style-type: none"> • Destaca las relaciones entre entidades para el usuario. • El motor asociativo subyacente, realiza uniones naturales en tiempo real, en función de la selección del usuario • Los conocimientos de datos se pueden generar rápidamente a partir de un usuario competente. 	VENTAJAS: <ul style="list-style-type: none"> • Los usuarios pueden explorar datos sin ninguna experiencia de programación. • Ofrece la capacidad de conectarse a una multitud de fuentes de datos. • Se puede utilizar la API de esta herramienta para extraer datos. • Puede funcionar como front-end de herramientas que permiten realizar análisis avanzados. 	VENTAJAS: <ul style="list-style-type: none"> • Permite extraer informes de manera autónoma. • Usando la herramienta Quick Insights, puede producir gráficos personalizados. • Integración del análisis avanzado a través de scripts y objetos visuales de R, Microsoft Azure Machine Learning y Azure Stream Analytics.

Fig. 7. Comparación entre herramientas BI
Fuente: Tomado de: (Santillán Suquitana, 2018)

1.3.2 Data Mining

RapidMiner, antes conocida como YALE, siglas de “Yet Another Learning Environment”, es un data mining software muy conocido. De acuerdo con una encuesta de KDnuggets realizada en el año 2014, esta tool se constituía como la herramienta de data mining más usada. Destaca por permitir el acceso gratuito y por su fácil manejo dado que no requiere un conocimiento elaborado en programación, sin olvidar la gran selección de operadores que ofrece. Especialmente son los startups las que recurren a ella.

De la misma forma, el programa puede importar tablas Excel, archivos SPSS y masas de datos de diferentes bases de datos e integra los programas de data mining WEKA y R. Todo ello pone de relieve el carácter polivalente de este software.

WEKA (Waikato Environment for Knowledge Analysis) es un software de código abierto desarrollado por la Universidad de Waikato en la primera mitad de los años noventa. Basada en Java y compatible con Windows, macOS y Linux, la interfaz gráfica de usuario

facilita el acceso a este software que también ofrece conexión a bases de datos SQL, siendo capaz de procesar en ellas los datos solicitados.

Asimismo, presenta un sinnúmero de funciones de aprendizaje automático y secunda tareas tan relevantes del data mining como el análisis de clústeres, de correlación o de regresión, así como la clasificación de datos, punto fuerte este último del data mining software al usar redes de neuronas artificiales, árboles de decisión y algoritmos ID3 o C4.5. No obstante, este programa no es tan potente en aspectos como el análisis de clústeres, en el que solo se ofrecen los procedimientos más importantes.

KNIME (Konstanz Information Miner), desarrollado por la universidad de Constanza, se puso a disposición de los usuarios como software de código abierto pese a crearse desde el principio con objetivos comerciales.

Escrito en Java y preparado con Eclipse, en la actualidad KNIME se considera una herramienta de gran popularidad entre la comunidad internacional de programadores y, si se compara con otros programas de data mining, destaca por una amplia gama de funciones: con más de 1000 módulos y paquetes de aplicaciones preparados, esta herramienta permite descubrir estructuras ocultas de datos. Además, se pueden ampliar sus módulos con otras soluciones adicionales de pago.

SAS (Statistical Analysis System) es un producto de SAS Institute, una de las mayores empresas privadas de software en todo el mundo. SAS constituye la data mining tool principal en el análisis en el sector de los negocios y, de hecho, se considera como el programa más adecuado para grandes empresas, aunque también sea el software con un coste económico mayor de todos los aquí descritos.

El prestigio de este software se debe a que utiliza tecnología punta en la realización de pronósticos y presenta una visualización interactiva de los datos, de inestimable ayuda en grandes presentaciones. Con este programa de data mining se dispone básicamente de todos los elementos necesarios para llevar a cabo una minería de datos con éxito.

En la siguiente imagen se muestra una figura con el resumen de la comparación entre las principales herramientas mencionadas.

	Características	Lenguaje de Programación	Sistema Operativo	Precio/Licencia
RapidMiner	Apto para todos los procesos. Destaca el análisis predictivo.	Java	Windows, MacOS, Linux	Freeware, diferentes versiones de pago.
Weka	Muchos métodos de clasificación.	Java	Windows, MacOS, Linux	Software libre (GPL)
SAS	Caro, pero potente para grandes empresas.	Lenguaje SAS	Windows, MacOS, Linux	Freeware limitado a instituciones públicas, el precio se establece tras solicitud, diferentes modelos disponibles.
Knime	Software de data mining de código abierto que ha democratizado el acceso a los análisis predictivos.	Java	Windows, MacOS, Linux	Software libre (GPL) a partir de la versión 2.1

Fig. 8. Comparación herramientas Data Mining

Fuente: Adaptado de (IONOS, 2018)

1.4 Metodologías BI y Data Mining

1.4.1 Inteligencia de Negocios (Business Intelligence)

Para el desarrollo de un proyecto de inteligencia de negocios, se puede encontrar distintas metodologías, por lo tanto, se debe analizar la que más se apega a cada proyecto y organización. Se imponen entre la mayoría tres metodologías, la de Hefesto, Kimball e Inmon.

- **Metodología Hefesto**

La metodología HEFESTO se adapta perfectamente a cualquier ciclo de vida de desarrollo de software, debido a que se basa en los requerimientos de los usuarios y su estructura es capaz de adaptarse con facilidad a las modificaciones del negocio. (Vélez, 2008) Además, los objetivos de esta metodología son fáciles y sencillos de comprender, ya que se enfoca en proporcionar lineamientos para la construcción de Data Warehouse de una forma ordenada, comprensible y útil, para que quién la revise sepa lo que está haciendo.

La siguiente figura muestra como está estructurada la metodología Hefesto.



Fig. 9. Metodología Hefesto

Fuente: Tomado de (Bernabeu & García, 2017)

- **Metodología Kimball (Bottom-Up)**

La metodología planteada por Kimball permite simplificar la complejidad que conlleva el desarrollo de una solución DW/BI. Kimball brinda un enfoque de menor a mayor, además cuenta con herramientas prácticas que facilitan a la implementación de un Data Warehouse. (Molina, 2015).

Kimball usa un modelo dimensional para atender a las necesidades de los departamentos en distintas áreas de la empresa. El ciclo de vida del proyecto de DW, está conformado por cuatro principios básicos:

- Centrado en el negocio: El enfoque que se debe hacer es en la identificación de los requerimientos del negocio y mediante esto desarrollar relaciones sólidas con el negocio.
- Construcción de una infraestructura de información adecuada: Se debe diseñar una sólida base de información que sea fácil de usar y de alto rendimiento, donde se reflejará una extensa gama de requerimientos del negocio identificado en la empresa.
- Realizar entregas en incrementos significativos: Crear el DW en incrementos entregables en plazos de 6 a 12 meses, en esto la metodología es similar a las metodologías ágiles de desarrollo de software.
- Ofrecer una solución completa: Brindar todos los elementos necesarios para proporcionar valor a los usuarios de los negocios, también se incluye la entrega de aplicaciones para informes, capacitaciones, soporte y documentación. (Burgos, F.; Michilena, 2015).

El diagrama del ciclo de vida dimensional de Kimball se presenta en el siguiente diagrama:

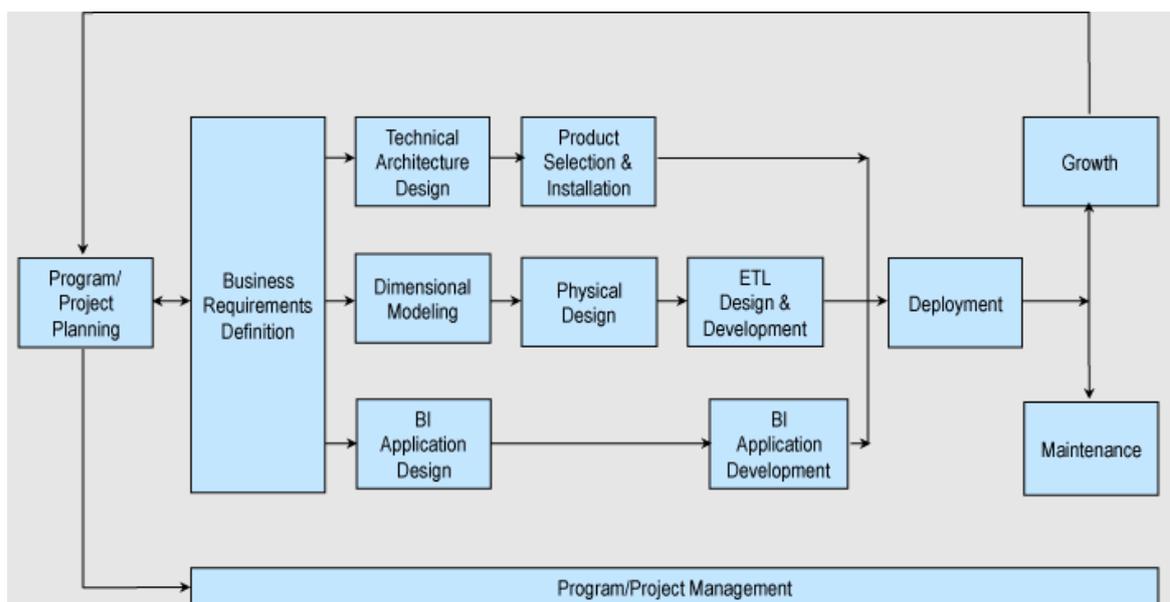


Fig. 10. Diagrama del ciclo de vida dimensional

Fuente: Tomado de (Mendoza, 2017)

- **Metodología Inmon (Top-Down)**

El enfoque que sigue la metodología Bill Inmon hace referencia de mayor a menor, el cual señala que la forma de construcción del Data Warehouse es teniendo una visión global para después pasar al detalle (Salazar Tataje, 2017).

La metodología Inmon es también conocida como una metodología descendiente o “Top-Down” que consiste en llevar cabo una relación entre las sucesivas etapas de estructuración de modo que se produzca una relación entre ellas mediante entradas y salidas de información.

Bill Inmon menciona que un Data Warehouse debe cumplir con las siguientes características:

- Dirigido a un área: en lugar de operaciones de la compañía se debe manejar los datos sobre un área específica.
- Integrado: acoplamiento de diferentes fuentes de datos de manera coherente.
- Variable en el tiempo: todos los datos pertenecen a un periodo de tiempo determinado.
- No volátil: no se eliminan los datos. (Agüero, 2019).

Esta metodología trabaja de manera iterativa contrario al clásico desarrollo de sistemas, debido a que con lo primero que se trabaja son los datos para luego ser integrados, probados y que sean programados de acuerdo con ellos, para evaluar los resultados y por lo tanto entender los requerimientos.

Para mayor apreciación de la información obtenida tras un proceso de investigación, a continuación, se muestra la tabla comparativa entre las Hefesto, Inmon y Kimball.

Tabla 2. Comparación metodologías BI

	Hefesto	Innom	Kimball
Enfoque General	Lineamientos de construcción de Data Warehouse	Metodología descendiente (Top-Down)	Abajo hacia arriba (Bottom-up)
Objetivo	Entregar una solución útil, sencilla y que sea fácil de entender	Entregar una solución basada en métodos probados	Entregar una solución que facilite la labor de los usuarios finales, para consultar directamente los datos con una tasa de respuesta razonable
Flexibilidad	Alto	Alto	Medio
Accesibilidad al usuario final	Alta	Baja	Alta
Herramientas	Modelo Conceptual	Tradicionales (Diagramas Entidad Relación y Sistemas de interpretación de Data)	Modelado Dimensional, se aparta del tradicional modelado relacional
Plazos	Cambios constantes	Continuos y discretos	Cambia lentamente
Más usada	Medio	Alta	Alta

1.4.2 Minería de datos (Data Mining)

- **CRISP-DM**

En el proceso de DM se encuentra varios modelos relevantes entre ellos se tiene a la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) siendo un proceso jerárquico que se basa en un conjunto de 4 niveles de abstracción (Alarcón et al., 2013). CRISP-DM es una de las metodologías más usadas en los proyectos de implementación de minería de datos.

Los estándares que se debe seguir en la metodología CRISP-DM dentro de las etapas aplicados al proyecto de análisis de datos, es igual a lo realizado por la ingeniería de software

cuando se habla de las fases de desarrollo de software. Esta metodología permite aplicar clasificación, clusterización, predicción de valores o análisis de dependencias o asociaciones. (Abad, 2020).

- **KDD**

Este proceso integra varias etapas hasta llegar a obtener el conocimiento para la toma de decisiones, inicia con la recopilación de la información de los diferentes orígenes de datos, esta etapa es conocida como etapa de selección de datos y es el pilar para que el proceso completo sea exitoso. La siguiente fase es el procesamiento de datos, cuyo objetivo es seleccionar, limpiar y transformar los datos, para obtener una vista minable que permita aplicar las diferentes técnicas de minería de datos (Vila, 2019).

- **SEMMA**

(Sample, Explore, Modify, Model, and Assess) creada por el Instituto SAS, que la define como una organización lógica del conjunto de herramientas funcionales de SAS Enterprise Miner para llevar a cabo las tareas centrales de la minería de datos. SEMMA ofrece y permite la comprensión, organización, desarrollo y mantenimiento de proyectos de Minería de Datos y consta de cinco etapas:

- **Muestra:** Es una etapa opcional. Una parte de un gran conjunto de datos se toma lo suficientemente grande como para extraer información significativa y lo suficientemente pequeña como para manipularla rápidamente.
- **Explorar:** Ayuda a obtener la comprensión y las ideas, así como refinar el proceso de descubrimiento mediante la búsqueda de tendencias y anomalías.
- **Modificar:** Modifica los datos mediante la creación, selección y transformación de variables para enfocar el proceso de selección del modelo. Aquí se buscan valores atípicos y se reduce el número de variables.
- **Modelo:** En esta etapa el software busca automáticamente la combinación de datos.
- **Evaluar:** Evalúa la confiabilidad y la utilidad de los hallazgos y estima el desempeño.

La siguiente figura muestra con detalle un resumen de las principales metodologías para minería de datos.

KDD, SEMMA Y CRISP son metodologías orientadas a procesos de Data Mining o Minería de datos las tres cuentan con ventajas y desventajas con respecto al uso moderno empresarial de la minería de datos siendo cada una modelada para ser solución o proceso de un problema o estrategia			
	KDD	SEMMA	CRISP
Enfoque	orientado a la identificación de patrones más favorables para cierta tarea	Orientado al desarrollo del proceso de MD o minería de datos.	Orientado a los objetivos empresariales.
uso	productos enfocados en patrones de datos	Ligado a productos SAS	Metodología abierta y gratuita
Metodología	metodología de patrones arquitectónicos orientados a datos	Metodología aún no definida	Metodología de gestión de proyectos
Complejidad	Más complejo de implementar que los otros dos, tiene una cantidad considerable de fases a desarrollar	Simple y bastante agil, sus fases están más implementadas a desarrollo ágil	Es el menos complejo de entender y aplicar, cuenta con una curva de adaptabilidad muy amplia para cualquier desarrollador
# de fases de desarrollo	9	6	6
Siglas	knowledge Discovery in Databases	Sample, Explore, Modify, Model and Access	Cross-industry Standard Process
Relevancia actual	Baja	Media	Alta

Fig. 11. Comparación entre metodologías Data Mining
Fuente: Tomado de (Cristancho, 2021)

1.5 Trabajos Relacionados

(Abad, 2020), desarrolló un modelo matemático para definir una tarifa diferencial en el uso del aparcamiento del Campus Central de la Universidad de Cuenca, puesto que muchos parqueaderos cobraban tarifas baratas, esto ocasionaba que la ciudad y el Campus se colapsara en tráfico. Las metodologías que usó fueron Hefesto para crear un almacén de datos (Data Warehouse) y posteriormente usarlo con otra herramienta de inteligencia de negocios, la cual es la minería de datos. En la minería de datos, la metodología implementada fue CRISP-DM, con las técnicas de clasificación y agrupamiento. La plataforma Grafana Labs le permitió crear el dashboard y los gráficos necesarios para visualizar los datos.

(Molina, 2015), implementó un aplicativo de inteligencia de negocios para la empresa Importadora Tomebamba S.A. El proyecto ayudó a la optimización del tiempo en la toma de decisiones ya que presenta información de calidad en el momento deseado. Se llevó a cabo

mediante la creación de un Data Mart para el área de ventas. Usó la metodología Kimball. Para la integración de los datos se empleó Pnetaho y también menciona la importancia de complementar el proyecto con minería de datos, por lo que la técnica seleccionada fue la de regresión y la ejecutó en el software Weka.

(Vele, 2020), implantó un sistema de Inteligencia de Negocios para automatizar los procesos ETL y el sistema web de reportería. Esto permitió brindarles a los usuarios finales la información en menos tiempo para que puedan tomar sus decisiones, así como también generar nuevos productos, entre otras ventajas. La metodología para la gestión del proyecto web fue Scrum, para la parte de BI se usó Kimball y CRISP-DM como referencia.

(Silva Peñafiel et al., 2019), presentaron una investigación enfocada en la selección adecuada de metodologías para la creación de Data Warehouse, tomó en cuenta varias características propias de cada metodología, desde su integración, el cumplimiento de objetivos, entre otros. Los métodos de investigación puestos a prueba fueron Ralph Kimball, Hefesto, SAS Rapid Data Warehouse. Como resultado se obtuvo que la metodología Hefesto permitió identificar ágilmente y también comprender los objetivos que se desea alcanzar con la implementación de un Data Warehouse, en las etapas iniciales permite definir claramente los requerimientos del usuario, involucrándolo en cada etapa de los siguientes procesos.

(Setiawan et al., 2017), realizaron un estudio en base a una empresa de Papelería, para conocer los procesos de negocio, la compra de bienes a los proveedores y las ventas al cliente final, y determinaron que estos procesos se pueden integrar automáticamente con la base de datos, también se concluyó que se pueden proporcionar informes según el gerente lo requiera. Las técnicas implementadas fueron por medio del uso de reglas de asociación de minería de datos y sistemas de información.

(Gatzioura & Sánchez-Marrè, 2015), desarrollaron recomendaciones de gran importancia basadas en patrones de artículos, buscando proporcionar más información sobre los hábitos de compra de sus clientes. Usaron técnicas de asociación (con el algoritmo A priori), extraídas de la base de datos transaccional, así presentaron como ventaja las recomendaciones de compra de artículos.

(Mazón Olivo et al., 2018), identificaron los indicadores clave de rendimiento (KPI), en una empresa comercializadora de cacao. La metodología CRISPD-DM se usó para la creación del dashboard, presentando los datos de manera visual y apoyar a la toma de decisiones. También usaron una técnica de minería de datos, la de regresión para pronosticar tendencias del rendimiento de la cosecha y los ingresos de ventas.

(Mazon-Olivo et al., 2019), implementaron inteligencia de negocios y minería de datos por medio de un ambiente web que permita analizar indicadores claves de desempeño, para los procesos de producción y ventas. La metodología CRISP-DM fue usada para el desarrollo, también Pentaho Server para el procesamiento analítico y algoritmos de regresión. Orientado a la toma de decisiones.

1.6 Matriz de Conceptos

El instrumento de matriz de conceptos se implementó con los trabajos y artículos científicos que más se asimilaban al objetivo y desarrollo del presente proyecto, a continuación, en las siguientes figuras se identifican las técnicas, algoritmos, herramientas y metodologías que se usaron.

	TECNICA				ALGORITMO					
	clasificación	regresión	agrupamiento	asociación	KNN	RNA	RF	k-MEANS	SVM	A priori
(Abad, 2020)	X		X		X			x		
(Molina, 2015)		X								
(Vele, 2020)										
(Silva Peñafiel et al., 2019)										
(Setiawan et al., 2017)				X						X
(Gatzioura & Sánchez-Marrè, 2015)				X						X
(Mazón Olivo et al., 2018)		X								
(Mazon-Olivo et al., 2019)		X								

Fig. 12. Matriz de Conceptos parte 1

	HERRAMIENTA						METODOLOGÍA BI				METODOLOGÍA DM	
	R shiny	Tableau	Pentaho	Grafana	Weka	DW	Kimball	Imon	Hefesto	SAS	KDD	CRISP-DM
(Abad, 2020)				X					X			X
(Molina, 2015)			X		X		X					
(Vele, 2020)						X	X					X
(Silva Peñafiel et al., 2019)							X		X	X		
(Setiawan et al., 2017)												
(Gatzioura & Sánchez-Marrè, 2015)												
(Mazón Olivo et al., 2018)	X											X
(Mazon-Olivo et al., 2019)	X		X									X

Fig. 13. Matriz de Conceptos parte 2

CAPÍTULO 2

2.1 Definición de requerimientos e indicadores

2.1.1 Requerimientos en BI

Es importante tener en cuenta el concepto que contextualiza a los requerimientos, por ello, en la descripción de (EvaluandoSoftware, 2021), se define al objeto de la ingeniería de requerimientos como un medio comunicación entre el cliente y el proveedor (desarrollador), para establecer el alcance funcional y técnico del proyecto a realizar. Los requerimientos tienen gran impacto en la gestión de los proyectos de software, por lo se consideran un elemento fundamental.

El ámbito del presente proyecto es de Inteligencia de Negocios y (Cavazos, 2008) en su publicación hace referencia que la toma de requerimientos en este tipo de proyectos, parte con personas de la empresa en los que está pensado enfocar el proyecto. Es decir, si se va a realizar un almacén de datos para apoyar la toma de decisiones de ventas, se debe levantar los requerimientos con las personas que estén a cargo de este departamento o que tengan una influencia importante en el mismo.

2.1.2 Indicadores

(Cuartin, 2020), manifiesta que los indicadores son unidades de medición usados para evaluar el rendimiento de procesos internos de la empresa, también cubren aspectos de rentabilidad, productividad, calidad de servicio, etc. En términos generales podrían considerarse las siguientes categorías de indicadores:

- **Indicadores de proceso:** Monitorean la cantidad y clases de actividades. Ejemplo: “Cantidad y clases de servicios prestados”, “Cantidad de personas capacitadas”. “Cantidad de personas de materiales producidos”.
- **Indicadores de resultados:** Evalúan si las actividades planificadas como si los objetivos, se alcanzaron o los resultados propuestos. Ejemplo: “Medición de conocimiento, actitudes por medio de encuestas.” “Calidad y beneficios prestados según entrevistas individuales”.

- **Indicadores históricos:** Muestran información recolectada y analizada acerca de temas específicos en un período de tiempo. Ejemplo: “Monto total de ventas por producto en el primer semestre del año 2021”.

Entre los beneficios de los indicadores pueden resaltar:

- Estimar con precisión el tiempo en proyectos a realizar.
- Identificar clientes rentables.
- Conocer el nivel de rendimiento del personal individual y en equipo.
- Tomar acciones estratégicas según los análisis y resultados obtenidos.

2.1.3 Requerimientos e indicadores en "Sari Popular"

La toma de requerimientos se llevó a cabo con el asesor comercial de la papelería “Sari Popular”, por medio de una reunión en la que se solicitó la creación de una lista sobre “Preguntas del Negocio”. En este caso dichas preguntas son los requerimientos históricos, con los que se desea conocer información sobre productos, clientes, entre otros, y los indicadores son lo que se puede medir, a continuación, se puede visualizar en la tabla 4.

Tabla 4. Requerimientos e indicadores “Sari Popular”

Requerimientos ■	Indicadores ■
1) Monto total de ventas de cada categoría en el año 2021.	2) Monto total de ventas de cada categoría en el año 2022.
3) Monto de venta generado por venta de hojas, resmas y cajas de papel bond en determinado periodo de tiempo.	
4) Cantidad de clientes que compran en cada local en determinado tiempo.	5) Cantidad de facturas de más de \$100 que se emitieron en 2021.
6) Cantidad de facturas de más de \$100 que se emitieron en 2022.	
7) El producto más vendido en cada zona en determinado periodo.	8) El producto menos vendido en cada zona en determinado tiempo.
9) El producto con mayor rentabilidad en el año 2021.	10) El producto con mayor rentabilidad en el año 2022.

Este esquema general de requerimientos e indicadores obtenidos sirven para tener en cuenta hacia que información está orientado el proyecto, más adelante se podrá ver más a

detalle cómo se usó este bosquejo mediante la aplicación de la metodología de inteligencia de negocios seleccionada.

2.2 Uso de la Norma ISO/IEC 25012 – Calidad de Datos

Para el proceso de evaluación de la calidad de los datos se tomó en cuenta las tablas de dimensiones que conforman el Data Mart, adicionalmente se usó la guía del trabajo desarrollado por (Calabrese et al., 2019), la misma que a su vez toma referencia de las pautas establecidas en la Norma ISO/IEC 25040. A continuación, se empieza a implementar la guía.

Establecer los requisitos de la evaluación: El propósito de realizar este proceso es determinar la calidad de las fuentes de datos que se obtuvo para realizar la solución de BI. De las características que tiene el estándar ISO/IEC 25012 se ejecutaron las de *Exactitud (Semántica y Sintáctica)* y *Consistencia*.

Especificar la evaluación: Tomando como ejemplo las características seleccionadas anteriormente, se planteó una métrica por la característica de Consistencia y la característica de Exactitud que ya cuenta con sus respectivas subcategorías. A continuación, en la tabla se establece los criterios de aceptación para la característica de Exactitud.

Tabla 5. Requisitos de evaluación, característica de Exactitud

Tipo	Categoría	Descripción
Semántica	Inaceptable	Si valor (%) ≥ 0 y < 0.4
	Mínimamente aceptable	Si valor (%) ≥ 0.4 y < 0.6
	Rango Objetivo	Si valor (%) ≥ 0.6 y < 0.9
	Excede los requerimientos	Si valor (%) ≥ 0.9
Sintáctica	Inaceptable	Si valor (%) ≥ 0 y < 0.1
	Mínimamente aceptable	Si valor (%) ≥ 0.1 y < 0.3
	Rango Objetivo	Si valor (%) ≥ 0.3 y < 0.6
	Excede los requerimientos	Si valor (%) ≥ 0.6

Tabla 6. Resumen criterios de aceptación, característica de Exactitud

Resumen		
Resumen de Criterios mínimos de aceptación (Exactitud)	Inaceptable	Semántica: Inaceptable. Sintáctica: Inaceptable.
	Mínimamente aceptable	Semántica: Mínimamente aceptable. Sintáctica: Mínimamente aceptable.
	Rango Objetivo	Semántica: Rango Objetivo. Sintáctica: Mínimamente aceptable.
	Excede los requerimientos	Semántica: Excede los requerimientos. Sintáctica: Excede los requerimientos.

La siguiente tabla hace referencia a los criterios de aceptación de la característica de consistencia.

Tabla 7. Resumen criterios de aceptación, característica de Consistencia

Tipo	Categoría	Descripción
Resumen de Criterios de aceptación mínimos (Consistencia)	Inaceptable	Si valor (%) ≥ 0 y < 0.4
	Mínimamente aceptable	Si valor (%) ≥ 0.4 y < 0.7
	Rango Objetivo	Si valor (%) ≥ 0.7 y < 0.9
	Excede los requerimientos	Si valor (%) ≥ 0.9

Diseñar la evaluación: En esta sección se define la manera en que se evaluará, en el caso práctico de la Dimensión Cliente.

Tabla 8. Diseño de la evaluación

Característica	Documentación
Exactitud Semántica	Atributos sobre los cuales se analizará la semántica: Cliente -> Se espera que haga referencia a una persona natural o una empresa Ciudad -> Se espera que mencionen a ciudades reales del país
Exactitud Sintáctica	Atributos sobre los cuales se analizará la sintáctica: Cliente -> Deben estar en letras mayúsculas Ciudad -> Deben estar en letras mayúsculas
Consistencia	Los datos de los clientes deben estar coherentemente asignados

Ejecutar la evaluación:

En la siguiente tabla se establecen los resultados del proceso de ejecución llevado a cabo.

Tabla 9. Ejecución de la Evaluación

Característica	Documentación
Campo: Cliente	Valor: 0.85
Campo: Ciudad	Valor: 0.95
Semántica: Rango Objetivo y Excede los requerimientos	
Campo: Cliente	Valor: 1
Campo: Ciudad	Valor: 1
Sintáctica: Excede los requerimientos	
Consistencia	Valor: 0.80 => Excede Rango Objetivo

Finalización de la evaluación:

Conforme se obtuvieron los resultados, en la siguiente tabla se muestran las respectivas definiciones.

Tabla 10. Valores obtenidos de la evaluación final

Característica	Documentación
Exactitud	Rango Objetivo y Excede los requerimientos
Consistencia	Excede los requerimientos
Resultados de la Evaluación	Rango Objetivo / Excede los requerimientos

Como se puede observar las características de la norma ISO/IEC 25012 aportan la calidad en la fuente principal del presente proyecto, dicha fuente son los datos que se tienen a disposición, llegando a estar en una escala adecuada por lo que es viable continuar.

2.3 Procesos ETL y construcción de Data Warehouse

En esta sección se desarrolla un almacén de datos para un determinado departamento de la empresa (Data Mart), con proyección a la consolidación de un Data Warehouse, que sería la unión de varios Data Mart con otros departamentos. La metodología seleccionada fue Hefesto, debido a su simplicidad de guía y desarrollo. A continuación, la figura representa el pipeline del ETL realizado.

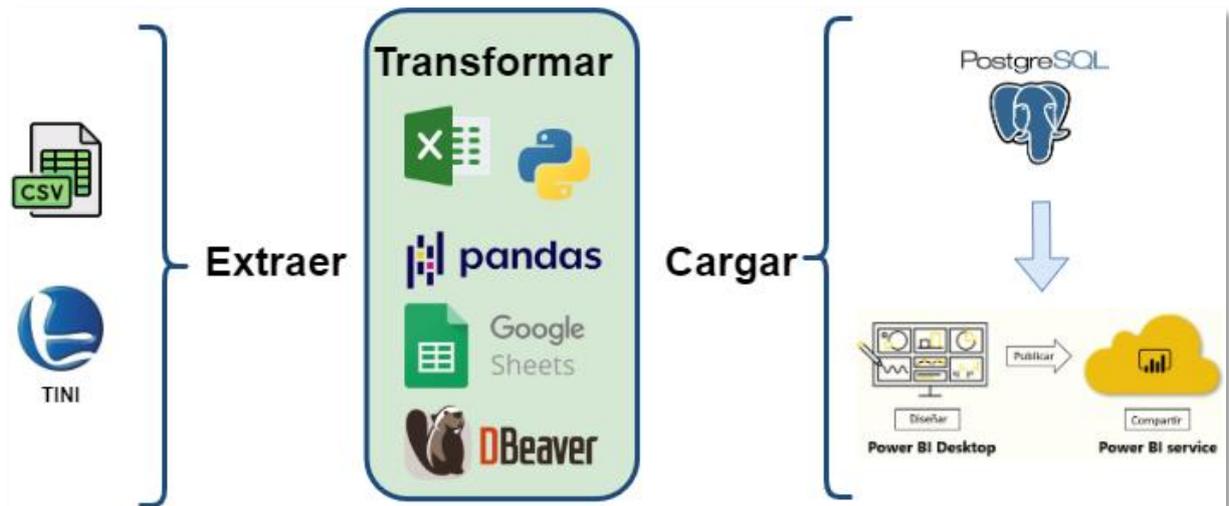


Fig. 14. Diagrama ETL

De manera breve, se indica los procesos realizados:

Extraer:

Se obtuvo las fuentes de datos del software TINI, el punto de venta que maneja la empresa, los archivos fueron adquiridos en formato .csv y .xlsx.

Transformar:

Para la transformación se llevó diferentes procesos con herramientas tales como Excel, Python, Pandas (una librería para manejar el tipo de archivos de las fuentes), Google Sheets también se implementó para manejar las hojas de cálculo y Dbeaver, un visor de bases de datos, con múltiples vistas para manejar toda la base.

Cargar:

En esta parte se cargó en PostgreSQL, un gestor de base de datos muy conocido y robusto. Una vez almacenados los datos se realizaron los reportes y dashboards en Power BI Desktop, para finalmente publicar en el servicio de Power BI (aplicación en la nube), para mantener actualizados los datos y que los usuarios finales puedan consumir los dashboards.

Como siguiente paso se llevó a cabo la metodología Hefesto.

2.3.1 Análisis de Requerimientos

Como inicio de la metodología se identificaron los requerimientos para el desarrollo, por lo que se definió comenzar la construcción del Data Warehouse enfocado en el departamento de ventas. Por lo tanto, se desarrolló cada paso establecido según lo menciona la metodología seleccionada para el proyecto.

2.3.1.1 Preguntas del Negocio

El objetivo principal de esta fase es identificar las necesidades de información por parte de los usuarios, definir las preguntas que guiarán hasta el final de la metodología para ser respondidas y apoyar en la toma de decisiones. Cabe recalcar que las preguntas generadas no deben ser tan simples, sino más elaboradas, que involucren varias entidades como clientes, tiempo, productos, etc.

Las preguntas del negocio se mostraron en la Tabla 4, aunque en esta sección se estructuraron de una mejor manera, para tener mayor visibilidad y comprensión de lo que se está buscando y principalmente en base a los datos con los que se cuenta. En la siguiente tabla se resumen los ajustes.

Tabla 11. Reestructuración de Preguntas del Negocio

Reestructuración de Preguntas del Negocio
<p>Estas preguntas, podrían ser sólo una</p> <ol style="list-style-type: none">1) Monto total de ventas de cada categoría en el año 20212) Monto total de ventas de cada categoría en el año 2022 <p>Resumida en “Monto total de ventas de cada categoría en un período de tiempo.”</p> <p><u>Razón: Los datos obtenidos son de los meses de Abril a Octubre del 2021, ya que la base de datos transaccional habría sido modificada por el proveedor del sistema que se maneja en “Sari Popular”.</u></p>
<p>Esta pregunta:</p> <ol style="list-style-type: none">3) Monto de venta generado por venta de hojas, resmas y cajas de papel bond en determinado periodo de tiempo. <p>Se podría cambiar por: “Monto de venta generado por productos (o tipos de productos) en un determinado periodo de tiempo.”</p>

Razón: Se podría agrupar este tipo de productos

Esta pregunta:

- 4) Cantidad de clientes que compran en cada local en determinado tiempo.

Se mantuvo.

Estas preguntas:

- 5) Cantidad de facturas de más de \$100 que se emitieron en 2021.
- 6) Cantidad de facturas de más de \$100 que se emitieron en 2022.

Se podría reemplazar por: "**Cantidad de facturas de más de \$100 que se emitieron en un determinado periodo de tiempo.**"

Razón: Los datos recolectados son de los meses de Abril a Octubre del 2021, ya que la base de datos transaccional habría sido modificada por el proveedor del sistema que se maneja en "Sari Popular".

Estas preguntas:

- 7) El producto más vendido en cada zona en determinado período.
- 8) El producto menos vendido en cada zona en determinado tiempo.

Se podría reemplazar por: "**Ventas por producto, local y periodo de tiempo.**"

Razón: El término "zona" se interpreta de mejor manera como "local", que tiene la empresa.

Estas preguntas:

- 9) El producto con mayor rentabilidad en el año 2021.
- 10) El producto con mayor rentabilidad en el año 2022.

Se podría reemplazar por: "**El producto con mayor rentabilidad en un periodo de tiempo.**"

Razón: Los datos recolectados son de los meses de Abril a Octubre del 2021, ya que la base de datos transaccional habría sido modificada por el proveedor del sistema que se maneja en "Sari Popular".

De tal modo que resumidas las preguntas que se van a responder en el desarrollo del proyecto son las siguientes:

- 1) Monto total de ventas de cada categoría en un período de tiempo
- 2) Monto de venta generado por productos (o tipos de productos) en determinado periodo de tiempo
- 3) Cantidad de clientes que compran en cada local en determinado tiempo
- 4) Cantidad de facturas de más de \$100 que se emitieron en un determinado periodo de tiempo
- 5) Ventas por producto, local y periodo de tiempo
- 6) El producto con mayor rentabilidad en un periodo de tiempo

2.3.1.2 Indicadores y Perspectivas

Una vez definidas las preguntas del negocio, se procede a su descomposición en perspectivas e indicadores para su utilización. En el libro escrito por (Bernabeu & García, 2017) se definen los términos de la siguiente manera:

- Los **Indicadores**, son valores numéricos y representan lo que se desea analizar concretamente, por ejemplo: saldos, importes, promedios, cantidades, sumatorias, fórmulas, etc.
- Por otro lado, las **Perspectivas** se refieren a las entidades mediante las cuales se quieren examinar los Indicadores, con el fin de responder a las preguntas predefinidas, por ejemplo: clientes, proveedores, sucursales, países, productos, rubros, etc.

Siendo así, de acuerdo a cada pregunta se irá definiendo las perspectivas e indicadores. La tabla 12 representa el proceso desarrollado.

Tabla 12. Perspectivas e indicadores

Indicadores y Perspectivas	
<p>Pregunta 1: <u>Monto total de ventas</u> de cada <u>categoría</u> en un período de <u>tiempo</u></p>	
Indicador	Perspectivas

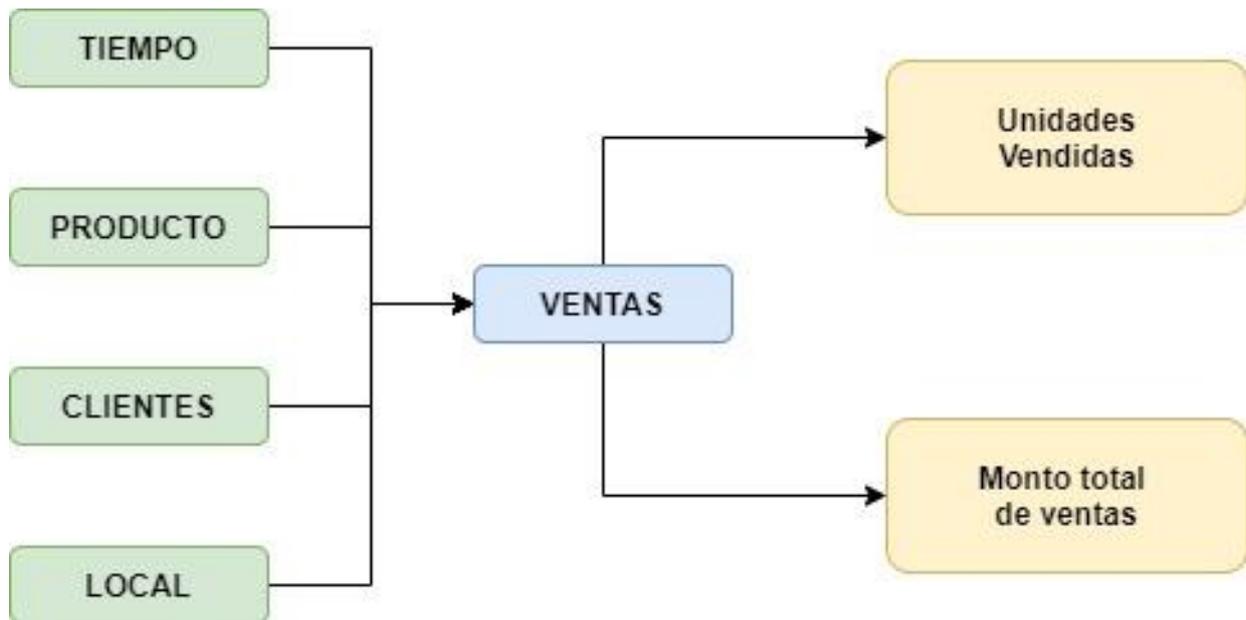


Fig. 15. Modelo conceptual

2.3.2 Análisis de Data Sources

En esta sección se analizarán las fuentes de datos con los que se cuenta

2.3.2.1 Hechos e Indicadores

En esta sección se da a conocer cómo serán calculados los indicadores, y los hechos como componentes. Adicionalmente se define las funciones de agregación que se usarán, por ejemplo; SUM (sumatoria), COUNT (conteo), AVG (promedio), entre otras. A continuación, se establecen en el caso práctico:

- Indicador: Unidades Vendidas
 - Hechos: Unidades (de producto) vendidas
 - Función de agregación: SUM

Aclaración: Representa la sumatoria de los productos (unidades) vendidos.

- Indicador: Monto total de las ventas
 - Hechos: Unidades Vendidas * Precio de venta
 - Función de agregación: SUM

Aclaración: Representa la sumatoria de ventas totales de cada producto de acuerdo a sus unidades multiplicadas por su respectivo precio.

Es importante enfatizar que los indicadores de la pregunta 3 (“Cantidad de clientes que compran en cada local en determinado tiempo”) y pregunta 4 (“Cantidad de facturas de más de \$100 que se emitieron en un determinado periodo de tiempo”) no fueron tomados en cuenta en el modelo conceptual ya que esta consulta se puede obtener en tiempo de ejecución utilizando una herramienta de visualizador gráfico.

2.3.2.2 Mapeo

Los recursos obtenidos fueron archivos con extensiones “.xlsx”, es decir archivos del software “Excel”. Es importante mencionar que el sistema transaccional de la empresa en la que se llevó a cabo el proyecto se maneja con el lenguaje COBOL, con un sistema de archivos para su funcionamiento, por lo que no se cuenta con un diagrama entidad-relación para comprender de mejor manera la estructura de los datos, que soportan las transacciones de la empresa. En la siguiente imagen se muestran los archivos obtenidos como recursos, con sus respectivos campos, dichos archivos representarían el proceso de ventas.



Fig. 16. Estructura de los archivos fuentes de datos

Por lo expuesto, es importante recalcar que la solución BI propuesta se genera por los recursos obtenidos y el mapeo se planteó de la siguiente manera:

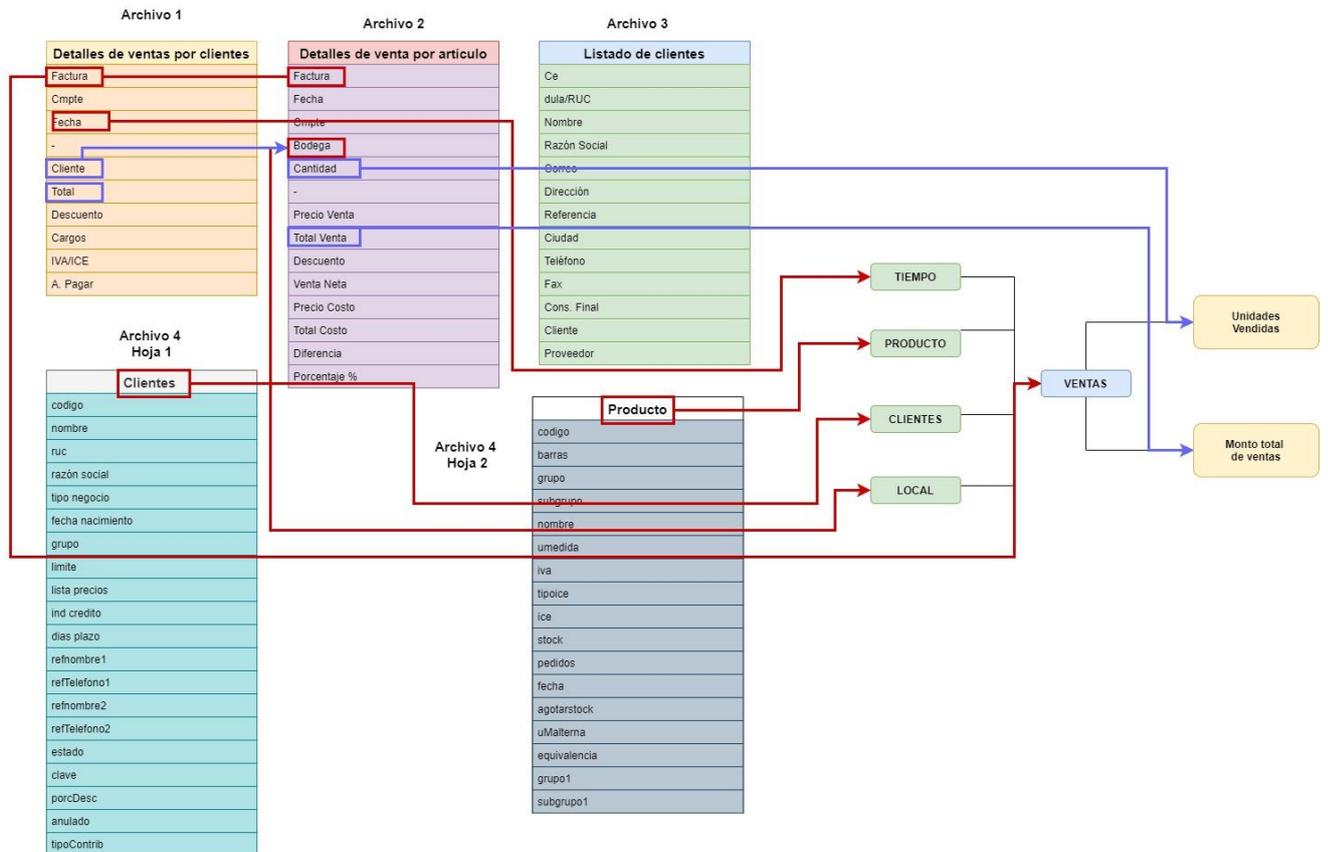


Fig. 17. Mapeo fuentes de datos

Descripción del mapeo:

- La perspectiva **Tiempo** se relaciona con el campo **Fecha** del archivo 1, el mismo que contiene la fecha de la factura.
- La perspectiva **Categoría** se relaciona con el campo **Grupo** del archivo 4, el mismo que contiene las características del producto.
- La perspectiva **Producto** se relaciona con todo el contenido del archivo 4 (Hoja 2) y con los valores del precio de venta y costo del archivo 2.
- La perspectiva **Clientes**, se relaciona con todo el contenido del archivo 4 (Hoja 1).
- La perspectiva **Local**, se relaciona con el campo **Bodega**, que contiene el archivo 2, se podría decir que es el “detalle de la factura”.

- El atributo **Factura** del archivo 1 y 2 se relaciona con el campo “**idFactura**” en la tabla de hechos. En este caso especial, el identificador/código de una factura es una dimensión degenerada. Este tipo de dimensiones se crean en las tablas de hechos, ya que solo interesa un único campo, como en este ejemplo, representa la transacción de la venta.
- El indicador **Unidades Vendidas** se relaciona con el campo **Cantidad**, del archivo 2 (contiene la venta de cada producto), generando la fórmula: SUM (Cantidad).
- El indicador **Monto Total de Ventas**, se relaciona con el campo **Total Venta**, del archivo 2, generando la fórmula: SUM (Total Venta).
- El indicador **Cantidad de facturas de más de \$100**, se relaciona con el campo **Total** del archivo 1, generando la fórmula: COUNT (IF Total > 100).
- Le indicador **Cantidad de clientes que compran en un local**, se relaciona el campo **Bodega** (archivo 2), generando la fórmula: COUNT (Bodega).

2.3.2.3 Granularidad

En este apartado se determina el detalle de cada campo que conformará las diferentes perspectivas, es decir, los seleccionados de la lista que tenían los archivos. En la siguiente serie de tablas están representadas las perspectivas con sus campos que las conforman.

Tabla 13. Granularidad de la perspectiva clientes

Perspectiva Clientes	
Nombre	Descripción
IdCliente	Es la clave primaria de la tabla clientes, y representa solamente a un determinado cliente.
Cliente	Nombre o razón social del cliente.
Ubicación	Lugar de residencia el cliente.

La estructura de la perspectiva tiempo representada en la Tabla 14 es simple, el punto a recalcar es que el nombre de la semana se omitió, puesto que en los filtros de los reportes solo el número sería más usable.

Tabla 14. Granularidad de la perspectiva tiempo

Perspectiva Tiempo	
Nombre	Descripción
idFecha	Clave primaria y representa la fecha completa.
diaNumero	Representa el número del día.
diaNombre	Representa Número del trimestre.
semanaNumero	Representa numéricamente la semana.
mesNumero	Representa numéricamente el mes.
mesNombre	Es el nombre del mes.
trimestreNumero	Representa numéricamente el trimestre.
trimestreNombre	Es el nombre del trimestre.
anio	Es el número del año.

Tabla 15. Granularidad de la perspectiva local

Perspectiva Local	
Nombre	Descripción
idLocal	Es la clave primaria de la tabla locales, y representa solamente a un determinado local.
Nombre	Hace referencia a la designación de cada local.

Tabla 16. Granularidad de la perspectiva producto

Perspectiva Producto	
Nombre	Descripción
idProducto	Es la clave primaria de la tabla productos, y representa solamente a un determinado producto.
Producto	Es un nombre descriptivo de un producto.
Categoría	Es un nombre descriptivo de un determinado conjunto de productos.
Subcategoría	Es un nombre descriptivo de una determinada categoría de productos.
PrecioVenta	Es un valor numérico decimal que representa el precio de venta de un producto.
PrecioCosto	Es un valor numérico decimal que representa el precio de costo de un producto.

2.3.2.4 Modelo Conceptual Ampliado

Aquí se plasman los procesos realizados anteriormente, y se unen para ampliar el Modelo Conceptual del proceso de ventas. En la siguiente figura se muestra cada perspectiva con sus campos y los indicadores.

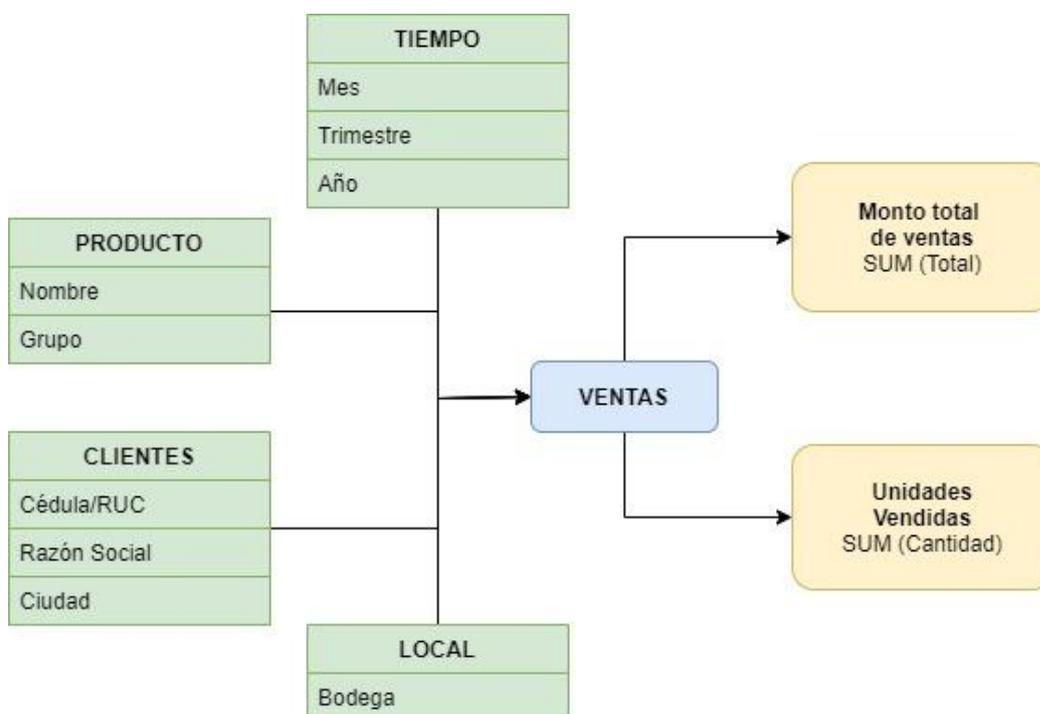


Fig. 18. Modelo conceptual ampliado

2.3.3 Modelo lógico del Data Warehouse

Como siguiente paso se creará el modelo lógico donde se establecerá la estructura de los datos que van a ser procesados y almacenados en un sistema gestor de base de datos.

2.3.3.1 Tipología

De entre las opciones existentes de esquemas “Estrella”, “Copo de nieve”, “Constelación”, se seleccionó el esquema Estrella por su sencilla comprensión y aplicación.

2.3.3.2 Tablas de Dimensiones

A continuación, se muestra el proceso de generación de las tablas de dimensiones que como base tienen las perspectivas que ya se definieron anteriormente. Siendo otra característica recomendable que se agregue un campo como clave principal, de tipo subrogada, este concepto es basado en lo que manifiesta (Bernabeu & García, 2017), ya que ocupan menos espacio, son más sencillamente manejables, más estructurados para la dimensión del Data mart o Data Warehouse entre otros beneficios.

Algunos campos de las dimensiones que gramaticalmente usan tilde, así como también que llevan la letra “ñ”, fueron omitidos para evitar cualquier inconveniente. con palabras Finalmente se modifican los nombres de los campos si su significado no es tan intuitivo.

Son tomados en cuenta aquellos campos que desde una perspectiva visual se los ve materializados en los reportes que fueron construidos posteriormente, por ello no se visualizaría en un dashboard tan amigable y usable una cédula de cliente o un código de producto.

Perspectiva Clientes:

- La tabla de dimensión se referenció con el nombre “*dimClientes*”.
- Se le agregó una clave principal con el nombre “*idCliente*”.
- Se le modificó el nombre del campo “*Razón Social*” por “*cliente*”.
- Se le modificó el nombre del campo “*Ciudad*” por “*ubicación*”.

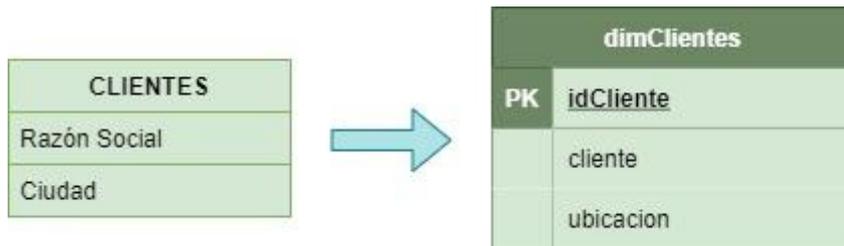


Fig. 19. Perspectiva clientes

Perspectiva Producto:

- La tabla de dimensión se referenció con el nombre “*dimProductos*”.
- Se le agregó una clave principal con el nombre “*idProducto*”.
- Se le modificó el nombre del campo “*Nombre*” por “*producto*”.
- Se le modificó el nombre del campo “*Grupo*” por “*categoría*”.
- Se le modificó el nombre del campo “*Subgrupo*” por “*subcategoría*”.
- Se le modificó el nombre del campo “*Grupo*” por “*categoría*”.
- Se le modificó el nombre del campo “*Precio Venta*” que viene del archivo 2 de las fuentes de datos, por “*precioVenta*”.
- Se le modificó el nombre del campo “*Precio Costo*” que viene del archivo 2 de las fuentes de datos, por “*precioCosto*”.

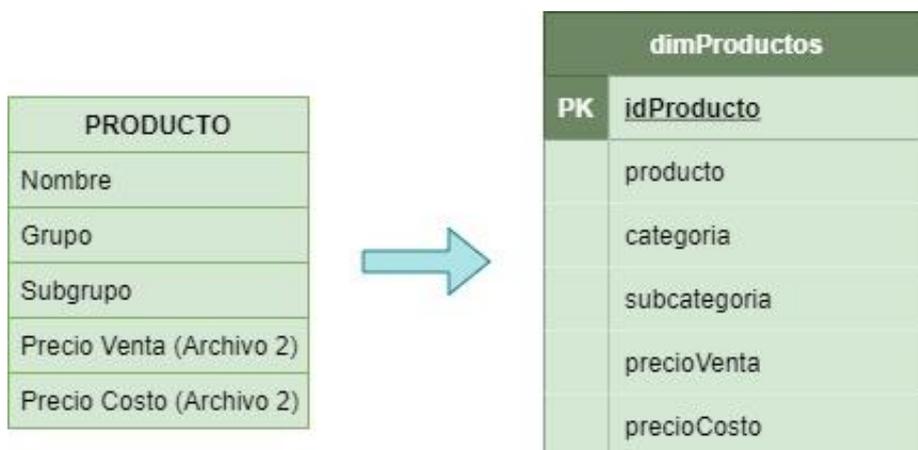


Fig. 20. Perspectiva producto

Perspectiva Local:

- La tabla de dimensión se referenció con el nombre “*dimLocales*”.
- Se le agregó una clave principal con el nombre “*idLocal*”.
- Se le modificó el nombre del campo “*Bodega*” por “*nombre*”.

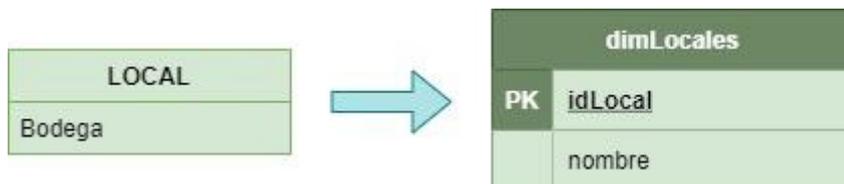


Fig. 21. Perspectiva local

Perspectiva Tiempo:

- La tabla de dimensión se referenció con el nombre “*dimFechas*”.
- Se le agregó una clave principal con el nombre “*idFecha*”.
- Se le procesó la fecha y la “*Fecha completa*” se referenció como “*fecha*”.
- Se le procesó la fecha y el “*Día*” se referenció como “*diaNumero*” y “*diaNombre*”.
- Se le procesó la fecha y la “*Semana*” se referenció como “*semanaNumero*”.
- Se le procesó la fecha y el “*Mes*” se referenció como “*mesNumero*” y “*mesNombre*”.
- Se le procesó la fecha y el “*Trimestre*” se referenció como “*trimestreNumero*” y “*trimestreNombre*”.
- Se le procesó la fecha y el “*Año*” se referenció como “*anio*”.



Fig. 22. Perspectiva tiempo

2.3.3.3 Tablas de Hechos

A continuación, se definió la tabla de hechos con sus respectivos componentes de acuerdo al modelo conceptual predefinido.

- La tabla de Hechos se referenció el nombre de *“factVentas”*.
- Su clave principal es la combinación de las claves primarias de cada dimensión, en este caso: *“idCliente”*, *“idProducto”*, *“idLocal”*, *“idFecha”* y la dimensión degenerada *“idFactura”*.
- Se crearon dos hechos que corresponden a los indicadores establecidos anteriormente:
 - Monto Total de Ventas, se referenció con el nombre de *“total”*.
 - Unidades Vendidas, se referenció con el nombre de *“cantidad”*.



Fig. 23. Tabla de Hechos

2.3.3.4 Uniones

Por último, en esta etapa se realizó las uniones de las dimensiones con la tabla de hechos, como se visualiza en la siguiente imagen.

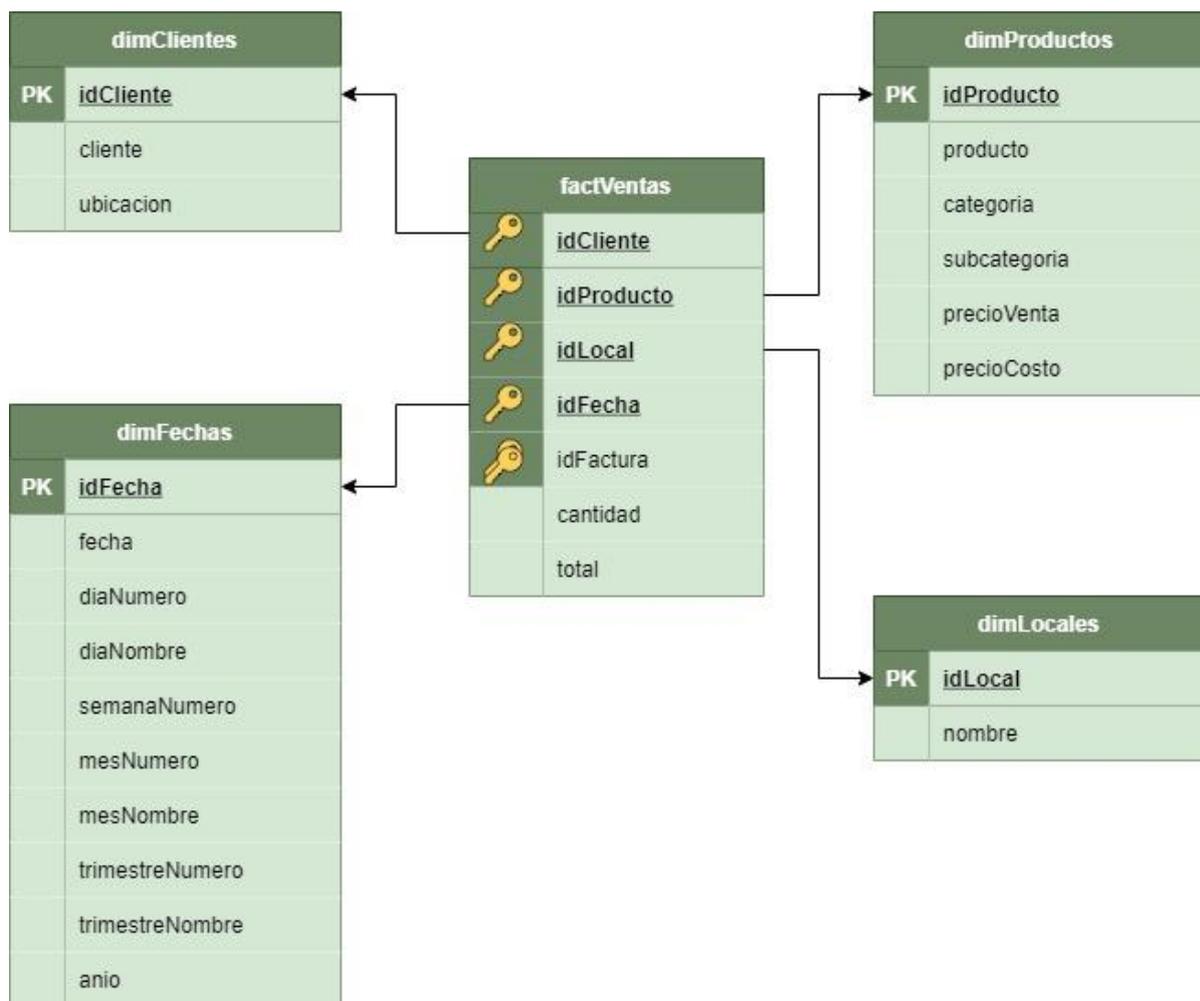


Fig. 24. Uniones

2.3.4 Integración de datos

Con el Modelo Lógico desarrollado se empieza a cargar los datos a través de procesos ETL. Primero se cargaron los datos de las tablas de dimensiones y luego de la tabla de hechos.

2.3.4.1 Carga Inicial

Cada tabla dimensión y tabla de hechos tuvo su propio proceso de carga, también es importante recalcar que la selección de los parámetros/características para cada dimensión se dio en base a lo que el usuario final observaría, por ello se necesita definir, qué es lo más relevante y que se puede presentar en el reporte final. Un punto importante a mencionar es que la carga de las dimensiones se realizó cuando ya se llevó previamente un proceso de extracción y transformación, donde se prepararon los datos.

La carga inicial se muestra a continuación, en la figura.

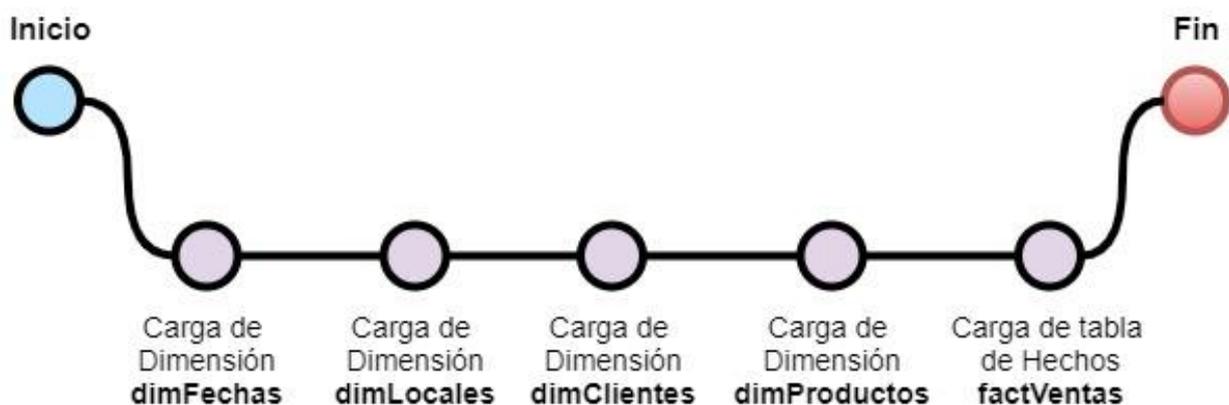


Fig. 25. Esquema de carga inicial

Carga de dimensión dimFechas: Para la carga de la dimensión Fechas se organizó de acuerdo a los campos preestablecidos en el modelo conceptual. Mediante Excel se planteó la información de la siguiente manera.

'idFecha';	'fecha';	'diaNumero';	'diaNombre';	'semanaNumero';
20210401;	'01/04/2021';	1;	'Viernes';	13;
20210402;	'02/04/2021';	2;	'Sábado';	13;
20210403;	'03/04/2021';	3;	'Domingo';	13;
20210404;	'04/04/2021';	4;	'Lunes';	13;
20210405;	'05/04/2021';	5;	'Martes';	14;

Fig. 26. Fuente de datos de fechas, parte 1

'mesNumero';	'mesNombre';	'trimestreNumero';	'trimestreNombre';	'anio';
4;	'Abril';	2;	'2 Trimestre';	2021;
4;	'Abril';	2;	'2 Trimestre';	2021;
4;	'Abril';	2;	'2 Trimestre';	2021;
4;	'Abril';	2;	'2 Trimestre';	2021;
4;	'Abril';	2;	'2 Trimestre';	2021;

Fig. 27. Fuente de datos de fechas, parte 2

Finalmente se aplicó la sentencia *SQL Insert*, para enviar los datos a la dimensión tal como lo muestra la figura con el primer ejemplo.

```

INSERT INTO public."dimFechas"(
  "idFecha", "fecha", "diaNumero", "diaNombre", "semanaNumero",
  "mesNumero", "mesNombre", "trimestreNumero", "trimestreNombre", "anio")
VALUES (20210401, '01/04/2021', 1, 'Viernes', 13, 4, 'Abril', 2, '2 Trimestre', 2021);

```

Fig. 28. Sentencia SQL para cargar la dimensión de fechas

Carga de dimensión dimLocales: En esta dimensión se utilizó la columna que viene en el reporte de ventas por producto, llamada “Bo” y representa el número de local, se lo puede ver en la siguiente figura.

Sari Papeleria Popular S.A.										
Detalle de Ventas por Producto										
Periodo-reporte: 2021/04/01 al 2021/04/30										
Factura	Fecha	Cmpte.	Bo	Canti.	P.Venta	Dcto.	P.Costo	Difer.	%	
AP30010002		BASTIDOR 20X15 CELINA			Unidad/Medida: UN					
FC1000008---	21/04/06	EI0000009455	1	-	-	-	-	-	-	-
FC7000001---	21/04/16	EI0000013291	3	-	-	-	-	-	-	-
FC1000011---	21/04/19	EI0000014186	2	-	-	-	-	-	-	-
FC1000013---	21/04/23	EI0000015742	1	-	-	-	-	-	-	-
FC1000013---	21/04/23	EI0000015772	1	-	-	-	-	-	-	-
FC7000001---	21/04/26	EI0000016079	2	-	-	-	-	-	-	-
FC1000013---	21/04/27	EI0000016672	3	-	-	-	-	-	-	-
FC1000014---	21/04/27	EI0000016784	1	-	-	-	-	-	-	-
FC1000014---	21/04/27	EI0000016787	2	-	-	-	-	-	-	-
FC1000014---	21/04/28	EI0000017135	2	-	-	-	-	-	-	-
FC1000014---	21/04/29	EI0000017351	1	-	-	-	-	-	-	-
FC7000002---	21/04/29	EI0000017434	1	-	-	-	-	-	-	-
FC1000014---	21/04/30	EI0000017623	3	-	-	-	-	-	-	-
FC1000014---	21/04/30	EI0000017948	1	-	-	-	-	-	-	-

Fig. 29. Fuente de datos de los locales

El personal de la empresa comunicó el nombre de cada local con su número identificativo respectivamente. Como paso final se ejecutó la sentencia *SQL insert*, como lo muestra la figura



```
INSERT INTO public."dimLocales"("idLocal", nombre) VALUES(1,'Matriz');
INSERT INTO public."dimLocales"("idLocal", nombre) VALUES(2,'Bodega Principal');
INSERT INTO public."dimLocales"("idLocal", nombre) VALUES(3,'Sucursal 1');
```

Fig. 30. Sentencia SQL para cargar la dimensión de los locales

Carga de dimensión dimClientes: Para esta dimensión se seleccionaron determinados campos de un reporte de clientes, que contienen la estructura de la figura

Sari Papeleria Popular S.A.		
Lista de Clientes		
Nombre	Razon Social	Dirección
QUINDE WILFRIDO	QUINDE WILFRIDO	IBARRA
TACURI LUIS	TACURI LUIS	IBARRA
VEGA ALEX	VEGA ALEX	QUITO
ELENA ANDACHI	ELENA ANDACHI	BABAHoyo
RUEDA HILDA MARIA	RUEDA HILDA MARIA	IBARRA
PAGAY JOAQUIN	PAGAY JOAQUIN	CAYAMBE
CARMEN DOMINGUEZ	CARMEN DOMINGUEZ	IBARRA
JAIME DELGADO	JAIME DELGADO	IBARRA
PURCACHI CESAR	PURCACHI CESAR	IBARRA
MUÑOZ SILVIA	MUÑOZ SILVIA	QUITO
MARTINEZ PEREZ ROSA ELVIRA	MARTINEZ PEREZ ROSA ELVIRA	TULCAN
GARNICA JUAN	GARNICA JUAN	IBARRA

Fig. 31. Fuente de datos de los clientes

Luego de un proceso de limpieza de datos y ordenamiento, se generó el comando *SQL insert*, de la siguiente manera, con algunos ejemplos.



```
INSERT INTO public."dimClientes"(cliente, ubicacion) VALUES('A.I.A.I','IBARRA');
INSERT INTO public."dimClientes"(cliente, ubicacion) VALUES('ABAD FREIRE IMPORAF LTDA.','GUAYAQUIL');
INSERT INTO public."dimClientes"(cliente, ubicacion) VALUES('ABANT MEDIOS','QUITO');
INSERT INTO public."dimClientes"(cliente, ubicacion) VALUES('ABDON CERVANTES','IBARRA');
INSERT INTO public."dimClientes"(cliente, ubicacion) VALUES('ABEDRABBO REEM','IBARRA');
```

Fig. 32. Sentencia SQL para carga la dimensión de los clientes

Carga de dimensión dimProductos: Se seleccionaron determinados campos del reporte de productos, como se muestra en la siguiente figura.

grupo	subgrupo	nombre
ARTE	BASTIDOR	BASTIDOR 20X15 CELINA
ARTE	BASTIDOR	BASTIDOR TRIANGULO 40CM CELINA
ARTE	OLEO	OLEO LEFRANC 150ML AMARILLO LIMON
ARTE	OLEO	OLEO LEFRANC 150ML AMARILLO MEDIO
BAZAR	FIELTRO/FOMY	FOMY LAMINA A4 ESCARCHADO ADHESIVO
BAZAR	FIELTRO/FOMY	PALETAS BARRILITO FOMY COLORES X72
BAZAR	FIELTRO/FOMY	FOMY MOLDEABLE BARRILITO X15
DIDACTICO	ROMPECABEZA	JD RELOJ MUCHACHO TRABAJADOR
DIDACTICO	AJEDREZ	AJEDREZ PLAST IMANTADO CREATIV PEQ 1327
PAPELERIA	CUADERNO	CUAD PASTA DURA AC 200H PRIMAVERA H CU
PAPELERIA	PAPELES	PAPEL CONTINUO 1P TROQ 9 1/2X11 X1400
OFICINA	MAQUINARIA	GUILLOTINA LANCER METAL A4 MESA
OFICINA	ACSESORIOS	ORGANIZADOR ESCRITORIO LANCER X5 PZS
TECNOLOGIA	AUDIFONO	AUDIFONO WIRELESS TRAIN 600BT UNDER
TECNOLOGIA	MICROSD	MEMORIA MICROSD HP 16GB
TECNOLOGIA	PENDRIVE	MEMORY FLASH HP 64GB 150/165

Fig. 33. Fuente de datos de los productos

La columna nombre representa al producto como tal, el grupo es la categoría y el subgrupo es la subcategoría. Luego de un ordenamiento por producto se generó y ejecutó el comando *SQL insert*, como se muestran algunos ejemplos, a continuación, en la figura 34.

```

INSERT INTO public."dimProductos"(producto, categoria, subcategoria, precioVenta, precioCosto)
VALUES ('ACRILICO FRANCO 60ML AZUL VIOLETA', 'ARTE', 'ACRILICO', 0.75, 0.68);

INSERT INTO public."dimProductos"(producto, categoria, subcategoria, precioVenta, precioCosto)
VALUES ('FOMY PLIEGO ESCARCHADO 50X70 VERDE', 'BAZAR', 'FIELTRO/FOMY', 0.69, 0.59);

INSERT INTO public."dimProductos"(producto, categoria, subcategoria, precioVenta, precioCosto)
VALUES ('ABACO B/MAS 50 MULLOS PIZARRA', 'DIDACTICO', 'ABACO', 2.75, 1.95);

INSERT INTO public."dimProductos"(producto, categoria, subcategoria, precioVenta, precioCosto)
VALUES ('PAPEL COUCHE 65X90 300G', 'PAPELERIA', 'PAPELES', 0.27, 0.17);

INSERT INTO public."dimProductos"(producto, categoria, subcategoria, precioVenta, precioCosto)
VALUES ('PERFORADORA STUDMARK MEDIANA 4524', 'OFICINA', 'PERFORADORA', 3.71, 2.64);

INSERT INTO public."dimProductos"(producto, categoria, subcategoria, precioVenta, precioCosto)
VALUES ('AUDIFONO HUAWEI EARPHONES', 'TECNOLOGIA', 'AUDIFONO', 2.1, 1.5);

INSERT INTO public."dimProductos"(producto, categoria, subcategoria, precioVenta, precioCosto)
VALUES ('ACRILICO ARTESCO 30ML BRILLANTE ROJO', 'ARTE', 'ACRILICO', 0.42, 0.37);

INSERT INTO public."dimProductos"(producto, categoria, subcategoria, precioVenta, precioCosto)
VALUES ('ABACO PLASTICO ESCOLAR CARACOL 86710', 'DIDACTICO', 'ABACO', 2.63, 2.27);

INSERT INTO public."dimProductos"(producto, categoria, subcategoria, precioVenta, precioCosto)
VALUES ('ACETATO A3 COPIADORA MANUAL LANCER', 'PAPELERIA', 'PAPELES', 0.13, 0.08);

```

Fig. 34. Sentencia SQL para cargar la dimensión de los productos

Carga de tabla de hechos factVentas: Finalmente se procede a realizar la carga de la tabla de hechos. Se tomaron en cuenta 2 archivos, el primero que se podría considerar como la cabecera de una venta y el segundo, el detalle respectivamente. En la hoja “cabecera” se usó la fórmula “BuscarV” de Excel para obtener el “idCliente” de la dimensión clientes por cada factura, dando así la siguiente estructura.

Factura	Fecha	Cliente	idCliente
FC100006864	21/04/01	CONSUMIDOR FINAL	3536
FC100006865	21/04/01	CONSUMIDOR FINAL	3536
FC100006866	21/04/01	MALES FRANCO PABLO EDISSON	8997
FC100006867	21/04/01	MALES FRANCO PABLO EDISSON	8997
FC100006870	21/04/01	RUALES LUZGARDA	13249
FC100006871	21/04/01	CONSUMIDOR FINAL	3536
FC100006872	21/04/01	CONSUMIDOR FINAL	3536
FC100006873	21/04/01	ACHINA PATRICIA	33
FC100006874	21/04/01	HERRERA BAEZ LUISA EMILIA	6738
FC100006878	21/04/01	CHASIQUIZA CAMPAÑA JOSE LUIS	3064
FC100006879	21/04/01	CONSUMIDOR FINAL	3536
FC100006882	21/04/01	CARLOSAMA JANETH	2480
FC100006887	21/04/01	RIVERA LUIS FABIAN	12856
FC100006888	21/04/01	CONSUMIDOR FINAL	3536
FC100006889	21/04/01	CONSUMIDOR FINAL	3536
FC100006890	21/04/01	LOPEZ PAOLA	8619
FC100006891	21/04/01	LOPEZ PAOLA	8619
FC100006907	21/04/01	CONSUMIDOR FINAL	3536
FC100006908	21/04/01	DIAZ COLIMBA GUADALUPE JUDITH	4259
FC100006909	21/04/01	DIAZ COLIMBA GUADALUPE JUDITH	4259

Fig. 35. Estructura de hoja cabecera

En la hoja detalle como primera instancia se desplazó aparte en un nuevo archivo la columna del producto por factura, luego de eliminadas las filas en blanco, la figura 36 muestra el ejemplo de cómo se encontraba estructurado el archivo.

idFactura	
AP30010002	BASTIDOR 20X15 CELINA
FC1000008007	
FC1000008365	
FC1000014962	
AP30010008	BASTIDOR 100X100 CELINA
FC1000007658	
AP30010014	BASTIDOR 40X30 LANCER
FC1000008128	
FC1000009659	
AP30010026	BASTIDOR 50X50 CELINA
FC1000007182	
FC1000009824	
FC1000010917	
FC1000013897	
AP30010034	BASTIDOR 70X70 CELINA
FC1000007621	
AP30010038	BASTIDOR CANVAS 15X20
FC1000013712	
AP30010041	BASTIDOR 30X30 CELINA MANDALA V/FIG
FC1000009478	
FC7000001349	
FC1000012179	

← Producto
 } Facturas en las que aparece

Fig. 36. Productos por facturas

Posteriormente se implementó una función desarrollada con Python y la librerías Pandas para quitar el código por defecto de cada producto, ya que como lo sugiere la metodología Hefesto es mejor usar una clave subrogada. La función es la siguiente.

```

"""
    Elimina los códigos de los productos
    df: es el archivo que tiene la columna idFactura
"""
def eliminarCodigoProducto(df):
    for i in range(len(df)):
        # 14 es el número de caracteres antes del nombre del producto
        if len(df.iloc[i,0])>=14:
            dato = df.iloc[i,0]
            # actualiza para tener solamente el nombre del producto
            df.at[i,'idFactura'] = dato[17:]
        # guardar en un nuevo archivo
    df.to_excel("codigosEliminados.xlsx", index=False)

```

Fig. 37. Función que deja solamente el nombre del producto

Los resultados de la función anterior se muestran en la figura 38.

idFactura
BASTIDOR 20X15 CELINA
FC1000008007
FC1000008365
FC1000008542
BASTIDOR 100X100 CELINA
FC1000007658
BASTIDOR 40X30 LANCER
FC1000008128
FC1000009659
BASTIDOR 50X50 CELINA
FC1000007182
FC1000009824
FC1000010917
FC1000013897
BASTIDOR 70X70 CELINA
FC1000007621
BASTIDOR CANVAS 15X20
FC1000013712
BASTIDOR 30X30 CELINA MANDALA V/FIG
FC1000009478
FC7000001349
FC1000012179

Fig. 38. Códigos de productos eliminados

Luego se pasó a ejecutar la siguiente función, la misma que coloca el código del producto

```

"""
Pasa el código de cada fila a su adelante
"""
def passProductsToInvoices(df2):
    inicial = df2.loc[0,'idFactura']
    for i in range(len(df2)):
        # Las facturas empiezan siempre con "FC"
        if str(df2.loc[i,'idFactura']).startswith("FC"):
            # Crea una columna llamada "Nuevo"
            df2.loc[i,'Nuevo'] = inicial
        else:
            df2.loc[i,'Nuevo'] = "-"
            inicial = df2.loc[i,'idFactura']
    df2.to_excel("productosFacturas.xlsx", index=False)

```

Fig. 39. Función que coloca el nombre del producto en cada factura

El resultado se muestra a continuación.

idFactura	Nuevo
BASTIDOR 20X15 CELINA	-
FC1000008007	BASTIDOR 20X15 CELINA
FC1000008365	BASTIDOR 20X15 CELINA
FC1000008542	BASTIDOR 20X15 CELINA
BASTIDOR 100X100 CELINA	-
FC1000007658	BASTIDOR 100X100 CELINA
BASTIDOR 40X30 LANCER	-
FC1000008128	BASTIDOR 40X30 LANCER
FC1000009659	BASTIDOR 40X30 LANCER
BASTIDOR 50X50 CELINA	-
FC1000007182	BASTIDOR 50X50 CELINA
FC1000009824	BASTIDOR 50X50 CELINA
FC1000010917	BASTIDOR 50X50 CELINA
FC1000013897	BASTIDOR 50X50 CELINA
BASTIDOR 70X70 CELINA	-
FC1000007621	BASTIDOR 70X70 CELINA
BASTIDOR CANVAS 15X20	-
FC1000013712	BASTIDOR CANVAS 15X20
BASTIDOR 30X30 CELINA MANDALA V/FIG	-
FC1000009478	BASTIDOR 30X30 CELINA MANDALA V/FIG
FC7000001349	BASTIDOR 30X30 CELINA MANDALA V/FIG
FC1000012179	BASTIDOR 30X30 CELINA MANDALA V/FIG

Fig. 40. Productos por facturas

Como se puede observar se obtuvieron los productos en cada factura. Con estos datos se obtuvo el "idProducto" y demás datos que conforman los campos de la tabla de hechos. Un paso adicional fue filtrar el "idCliente" para separar todas las transacciones del cliente facturado como "consumidor final" de los otros clientes. A continuación, se muestra la estructura de los campos con algunos ejemplos.

idCliente	idProduct	idLocal	idFecha	idFactura	cantidad	total
3536	694	1	20210406	FC10000080--	3	3.75
3536	690	1	20210405	FC10000076--	2	24.55
3536	702	1	20210412	FC10000096--	1	1.83
3536	706	2	20210403	FC10000071--	1	3.93
3536	4206	2	20210405	FC10000076--	1	2.99
3536	16	1	20210405	FC10000075--	1	2.41
3536	5546	1	20210404	FC10000073--	10	16.52
3536	165	1	20210414	FC10000102--	3	2.01
3536	164	3	20210417	FC10000114--	5	6.47
3536	27	1	20210414	FC10000103--	2	0.98

Fig. 41. Datos de la tabla de hechos para consumidor final

idCliente	idProduct	idLocal	idFecha	idFactura	cantidad	total
9994	694	3	20210426	FC70000018---	2	2.50
2927	694	1	20210427	FC10000139--	1	1.25
7973	694	1	20210427	FC10000140--	1	1.25
15231	694	3	20210429	FC70000020--	1	1.25
453	702	1	20210406	FC10000081--	2	3.66
4534	706	2	20210415	FC10000109--	1	3.93
11022	706	1	20210422	FC10000128--	4	15.71
3729	711	1	20210427	FC10000137--	1	2.01
15288	698	3	20210413	FC70000013--	1	2.68
896	698	1	20210429	FC10000145--	1	2.68
8506	699	1	20210403	FC10000072--	4	8.19

Fig. 42. Datos de la tabla de hechos para demás clientes

Finalmente se guardaron los archivos con formato “.csv” y se ejecutó la sentencia SQL para cargar los datos directamente, como se encuentra en la siguiente figura.

```

COPY "factVentas"
FROM 'D:\Desktop\Carga de Datos\factVentas\nombreMes\insert_consumidorFinal.csv/insert_clientes.csv'
DELIMITER ','
CSV HEADER;

```

Fig. 43. Sentencia SQL para cargar la tabla de hechos

De esta manera se cargaron todas las dimensiones del modelo que finalizan esta etapa y permiten seguir con la metodología.

2.3.4.2 Actualización

Para las políticas de actualización se tomó en cuenta los siguientes puntos:

- La carga de datos será actualizada mensualmente.
- Durante el período que transcurre el mes se irán recolectando los datos y los reportes generados por el sistema informático TINI.

- Diariamente se irán desarrollando actividades para automatizar algunas tareas que intervienen en todo el proceso ETL.
- Las tablas dimensión o hechos se podrán reestructurar de manera que la empresa solicite información adicional que no se pueda obtener con el modelo de datos que actualmente está en funcionamiento.

2.4. Reportes, Dashboard y Minería de datos

En este apartado se presentan el proceso que se llevó a cabo para generar los reportes.

Una vez obtenido el almacén de datos con sus dimensiones y hechos respectivas se formó la conexión entre PostgreSQL y PowerBI mediante el conector que el mismo software de BI contiene, la figura 26 muestra lo descrito.



Fig. 44. Conexión base de datos PostgreSQL con Power BI

De esta forma se obtiene la conexión del Data Warehouse almacenado en el gestor de base de datos PostgreSQL. Es importante mencionar que lo mencionado es una forma de conexión, sin embargo, en ocasiones se puede presentar algún inconveniente para actualizar los datos, de ser este el caso una solución sería conectar la base de datos con Power BI mediante un ODBC. Como evidencia de ello en la figura 45 se muestra el modelado de datos, tal cual se desarrolló a través de la metodología Hefesto.

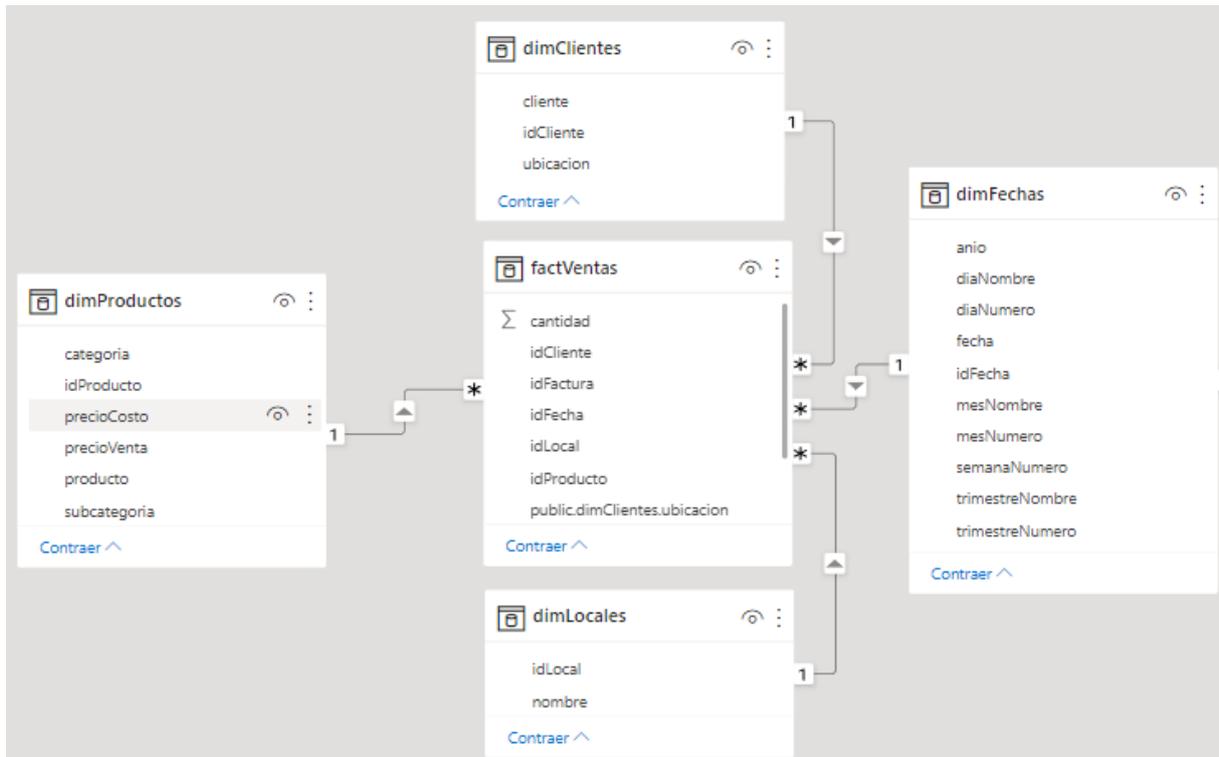


Fig. 45. Data Warehouse – Modelo de datos en Power BI

El modelo de datos proporciona la estructura para generar las visualizaciones necesarias, sin embargo, falta agregar un componente importante, conocidas también como “Medidas”. Las medidas en Power BI proporcionan diferentes tipos de fórmulas para llevar a cabo un fin como puede ser la sumatoria de una columna, el filtro, cálculos con fechas, etc. La figura 46 muestra las primeras medidas implementadas.

```

FC V x P = CALCULATE(SUM(factVentas[total]),
    FILTER('factVentas',LEFT('factVentas'[idFactura],2)="FC"))

VC V x P = CALCULATE(SUM(factVentas[total]),
    FILTER('factVentas',LEFT('factVentas'[idFactura],2)="VC"))

Total FC - VC = COALESCE([FC V x P] - [VC V x P],0)

```

Fig. 46. Medidas para ventas por producto

FC V x P y VC V x P: Cumplen un similar funcionamiento, en el campo “idFactura” de la tabla de hechos “factVentas” filtran los que empiezan por “FC” los mismos que representan una factura como tal y los separan de los que empiezan con “VC” ya que

son notas de crédito (devolución de una venta) por lo que no sumarían al total de ventas.

Total FC – VC: Es entendible su funcionamiento porque calcula realmente el total de ventas y además lleva la función “Coalesce”, para evitar que Power BI por defecto coloque “Blank”, y en lugar de ello ponga el valor de 0, en caso de que un resultado fuese nulo.

Las medidas en Power BI se ejecutan con fórmulas DAX (Data Analysis Expressions - Expresiones de análisis de datos). Continuando, en la figura 47 se muestran la medida que fue realizada para calcular la cantidad de facturas mayores a \$100.

```
Mayores a 100 =
VAR Cant = CALCULATE([Filas],
    FILTER('fv_facturas',LEFT('fv_facturas'[idFactura],2)="FC"), fv_facturas[Total]>100)
RETURN
    IF(ISBLANK(Cant),0,Cant)
```

Fig. 47. Medida para calcular el número de facturas mayores a \$100

Mayores a 100: Se filtró solamente el monto total de ventas de las facturas (“FC”) mayores a 100. Como la granularidad mínima en la dimensión fecha va desde cada día es muy probable que no todos los días se generen facturas con monto total mayor a 100. Por ello si son valores en blanco en la etiqueta se mostraría “0” y de esta manera se evitaría que aparezca “Blank”.

La última sección de medidas implementadas fue desarrollada para calcular la rentabilidad por producto, en la siguiente figura 48 muestra dichas medidas.

```
Ventas = CALCULATE(
    SUMX(factVentas,factVentas[cantidad]*factVentas[public.dimProductos.precioVenta]),
    FILTER('factVentas',LEFT('factVentas'[idFactura],2)="FC"))

Costos = CALCULATE(
    SUMX(factVentas,factVentas[cantidad]*factVentas[public.dimProductos.precioCosto]),
    FILTER('factVentas',LEFT('factVentas'[idFactura],2)="FC"))

Utilidad = [Ventas] - [Costos]

Margen = DIVIDE([Utilidad],[Ventas],0)
```

Fig. 48. Medidas para calcular la rentabilidad por producto

Ventas: Multiplica la cantidad de un producto por su respectivo precio de venta en una transacción de la tabla de hechos y filtra solamente las facturas “FC”.

Costos: Multiplica la cantidad de un producto por su respectivo precio de costo en una transacción de la tabla de hechos y filtra solamente las facturas.

Utilidad: Se obtuvo tras restar el total de ventas menos el total de costos.

Margen: Se usó la fórmula “Divide” para controlar los casos de división entre cero, y se dividió entre la utilidad por las ventas para obtener el margen porcentual.

La serie de procesos realizado permitió tener el ambiente listo para desarrollar los reportes y los dashboards contestando las preguntas del negocio que se plantearon al inicio del proyecto. El sistema de inteligencia de negocios se mostró como tal en el capítulo 3, de los resultados.

2.4.3 Algoritmo A Priori

En este apartado se implementó algoritmo A Priori, para profundizar en el tema se hace mención al texto de (García & Acevedo, 2010) donde especifica que este algoritmo busca primero todos los conjuntos frecuentes unitarios (contando sus ocurrencias directamente en la base de datos), se mezclan estos para formar los conjuntos de ítems candidatos de dos elementos y seleccionan entre ellos los frecuentes.

Considerando la propiedad de los conjuntos de ítems frecuentes, se vuelve a mezclar estos últimos y se seleccionan los frecuentes (hasta el momento ya han sido generados todos los conjuntos de ítems frecuentes de tres o menos elementos). Así sucesivamente se repite el proceso hasta que en una iteración no se obtengan conjuntos frecuentes.

El uso común se basa en encontrar grupos de artículos o productos que aparecen frecuentemente en la transacción de compra de los clientes, obteniendo esta información se podría generar nuevas formas de promocionar la mercancía.

Un tema ligado al algoritmo A Priori son las reglas de asociación, que intentan mostrar la dependencia de un producto con otro. En otro contexto, el objetivo se definir de modo que si se ha comprado un producto “A” entonces es probable que se compre un producto “B”. Dentro de las reglas de asociación destacan los siguientes elementos:

Antecedente (Si): Es un elemento que se encuentra dentro de los datos.

Consecuente (Entonces): Es un elemento que se encuentra en combinación con el antecedente.

Apoyo/Soporte: Indica la frecuencia con la que los productos aparecen en los datos y proporciona protección para las transacciones "A" y "B". Puede calcularse hallando el número de transacciones en las que aparece un conjunto de productos. De otra forma:

$$\text{Soporte} = N^{\circ} \text{ de veces ocurrido "A"} / N^{\circ} \text{ Total de transacciones}$$

Confianza: Se refiere a la probabilidad de que un artículo "B" se compre también si se compra el artículo "A". De otra forma:

$$\text{Confianza} = N^{\circ} \text{ de veces ocurrido "A" y "B"} / \text{Total de ocurrencias de "A"}$$

Lift: Indica la probabilidad de que se compre el artículo "B" cuando se compra el artículo "A", controlando al mismo tiempo la popularidad del artículo "B". Se puede calcular dividiendo la confianza entre el soporte. De otra forma:

$$\text{Lift} = \text{Confianza de "A" y "B"} / \text{Soporte}$$

2.4.4 Aplicación del algoritmo A priori en Weka

Para la aplicación del algoritmo se utilizó el programa Weka, el mismo que está compuesta por una serie de herramientas gráficas de visualización y diferentes algoritmos para el análisis de datos y modelado predictivo.

En vista de la inmensa cantidad de datos se optó por encontrar las relaciones entre los productos más vendidos y así mismo en las 50 transacciones que tenían más productos. Los productos obtenidos mediante la visualización del reporte de ventas por producto fueron los que se muestran en la figura 49.

@attribute	CUAD_ESPIRAL_AC_100H
@attribute	CUAD_ESPIRAL_AC_100H
@attribute	PAPEL_FOLDER_A4
@attribute	PAPEL_BOND_CAJA_X250
@attribute	PAPEL_BOND_CAJA
@attribute	B_BIC_PUNTA_FINA
@attribute	CARP_TRANS_OFICIO
@attribute	CUAD_ESPIRAL_AC_100H
@attribute	PAPEL_FOLDER_A4
@attribute	CUAD_ESPIRAL_AC_100H
@attribute	COLOR_NORMA_LARGO
@attribute	PAPEL_BOND_CAJA
@attribute	CUAD_ESPIRAL_AC
@attribute	CUAD_COSIDO_100H
@attribute	CUAD_ESPIRAL_JR
@attribute	B_BIC_PUNTA_FINA
@attribute	PAPEL_BOND_CAJA_X250
@attribute	CUAD_ESPIRAL_AC
@attribute	B_BIC_PUNTA_FINA
@attribute	MARC_X_12

Fig. 49. Productos seleccionados para el algoritmo A priori

Con este recurso en la sección de resultados se interpreta la ejecución del algoritmo.

2.5 Cuestionario SUS (System Usability Scale - Escala de usabilidad del sistema)

Se planteó ejecutar el cuestionario SUS para evaluar la usabilidad y satisfacción del Sistema BI desarrollado para los usuarios la empresa Sari Popular. Se encuentra estructurado por 10 ítems, los cuales se detallan en *Anexo A*:

El cuestionario propuesto fue respondido específicamente por el asesor comercial y los colaboradores del departamento de ventas, siendo un total de 5 personas, ya que son quienes intervienen directamente en el proceso de toma de decisiones empresariales.

Para el modo de evaluación, el cuestionario SUS hace uso de una escala Likert, la misma que genera un número, que representa una medida compuesta de la usabilidad del sistema global sometido a estudio. La categorización de las opciones a responder con su equivalente numérica, se encuentra de la siguiente manera:

- Muy en desacuerdo: 1 punto
- En desacuerdo: 2 puntos
- Neutral: 3 puntos
- De acuerdo: 4 puntos
- Muy de acuerdo: 5 puntos

El siguiente paso es sumar los puntos de las preguntas impares (positivas) y restar 5 del total para obtener un resultado que se definirá como “X”, es necesario restar 5 porque en realidad es una escala de 0 a 4, ya que son 5 posibilidades.

A 25 restar las preguntas pares (negativas) para obtener un resultado definido como “Y”. Finalmente sumar (X+Y) y para llevar la escala a 100, se debe multiplicar por 2.5. Es importante aclarar que el resultado final es una nota, no un porcentaje. Finalmente, el resultado se compara con la figura

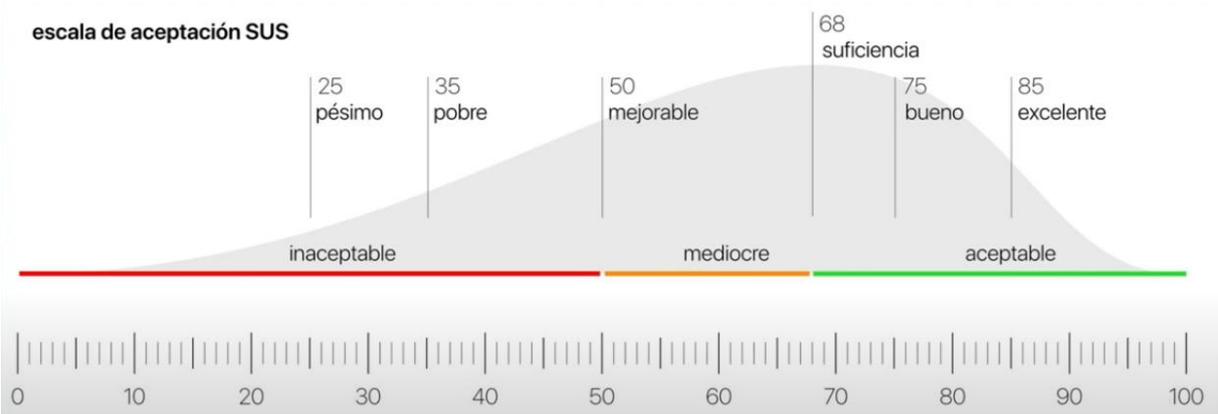


Fig. 50. Escala de aceptación SUS

Fuente: Tomado de (Hernández, 2021)

Como se puede ver en la anterior imagen, una equivalencia de los resultados para interpretar y entender la satisfacción del sistema. Los resultados de la aplicación del cuestionario SUS se encuentran en el capítulo 3.

CAPÍTULO 3

Resultados

3.1 Definición de métricas de evaluación

En este apartado se definen las métricas para evaluar el proceso y los resultados obtenidos del proyecto. Con el objetivo de conocer las impresiones generadas por parte de los usuarios finales, entre diversas opciones se tomó como instrumento el cuestionario SUS (System Usability Scale), traducido; Escala de usabilidad del sistema. Es uno de los cuestionarios más reconocidos por su propósito sencillo y concreto que con un número mínimo de cuestiones/preguntas, permite evaluar la eficacia, eficiencia y satisfacción de los usuarios al usar un sistema o aplicación tecnológica informática.

3.2 Norma ISO 25012

El modelo de evaluación propuesto y desarrollado se realizó en base a 3 características que fueron consideradas para el proyecto; La Exactitud (sintáctica y semántica) y la consistencia. Tras la ejecución del proceso apoyado de la guía (Calabrese et al., 2019), se pudo determinar que las fuentes de datos obtenidos como recursos, no fueron 100% puros y limpios, lo cual es considerablemente normal.

Tomando en cuenta la parte técnica, la aplicación del estándar permitió construir las diferentes dimensiones del Data Mart con integridad, es decir con una estructura lo suficientemente compacta.

La utilización del estándar aportó valor al proyecto en un sentido de buena práctica, ya que está fundamentada científicamente, y permitió generar más calidad.

3.3 Sistema BI y Preguntas del Negocio

La construcción de los dashboards y reportes se organizaron según muestra la siguiente imagen.



Fig. 51. Pantalla principal – Sistema BI

En la tabla 18 se resume las preguntas del negocio que fueron planteadas al inicio del proyecto y en que pantalla del sistema BI se encuentran solventadas.

Tabla 18. Pantallas sistema BI y preguntas del negocio

Pantallas del sistema BI	Preguntas del negocio
1.- Ventas <ul style="list-style-type: none"> Mensuales Por local Por producto Cantidad de Facturas 	P1. Monto total de ventas de cada categoría en un periodo de tiempo. P2. Monto total de ventas por productos (o tipos de productos en determinado periodo de tiempo). P4. Cantidad de facturas de más de \$100 que se emitieron en un determinado periodo de tiempo. P5. Ventas por producto, local y periodo de tiempo.
2.- Ventas <ul style="list-style-type: none"> Por cliente Por ciudad Cantidad de clientes 	P3. Cantidad de clientes que compran en cada local en un determinado tiempo.
3.- Cantidad de Facturas	P6. El producto con mayor rentabilidad en un periodo de tiempo.

Como se pudo observar las preguntas respondidas, a continuación, en una serie de figuras se mostrarán los dashboards y reportes construidos en el sistema BI.

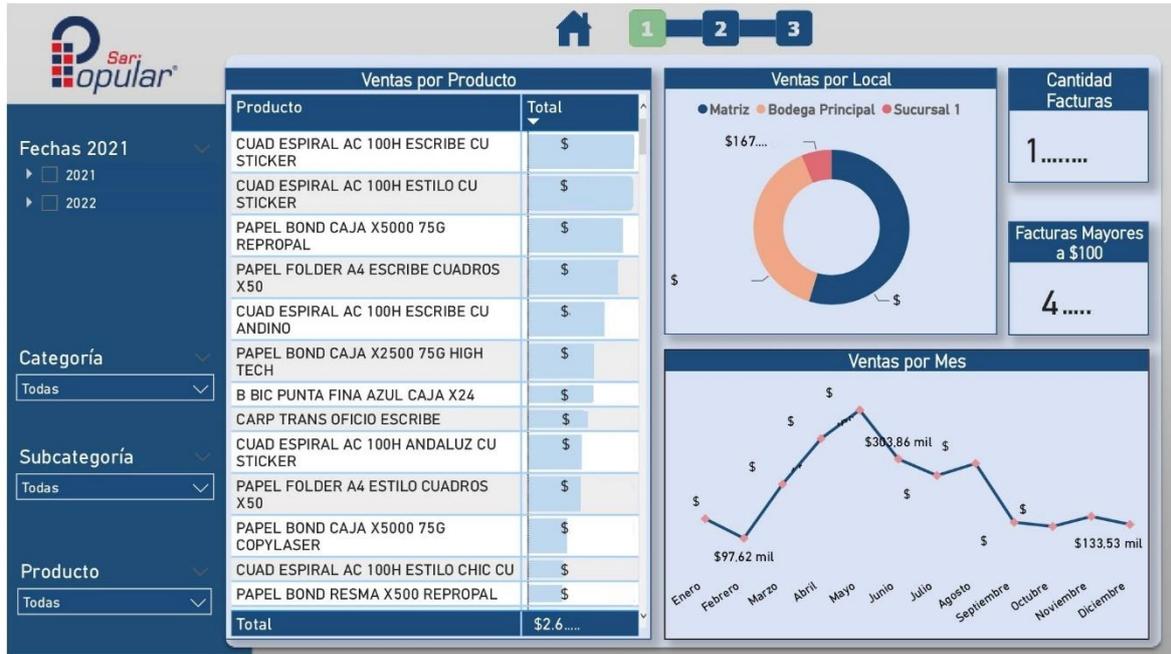


Fig. 52. Ventas mensuales, por local, por producto y cantidad de facturas



Fig. 53. Ventas por cliente, ciudad, cantidad de clientes por local

Producto	Ventas	Costos	Utilidad	Margen
TONER COPIADORA RICOH 2550	\$75.89	\$4.46	\$71.43	94.12 %
PIZARRA T/L KENART MAGNETICA	\$1.260.86	\$82.34	\$1.178.52	93.47 %
ESPIRAL 14MM X120				93.33 %
ESPIRAL 07MM X100	\$718.2	\$53.2	\$665	92.59 %
ESPIRAL 09MM X100				92.59 %
APLIQUE CORAZON HOLOGRAMA X1MEDIANO				88.89 %
CARTULINA FINE V/C RETAZOS GRANDES				88.89 %
MAQUETA PLANTA ACUARIO	\$112.95	\$12.55	\$100.4	88.89 %
VINCHA CARPETA X 1				88.89 %
B PELIKAN POINTEC NEGRO	\$16.72	\$2.28	\$14.44	86.36 %
B PELIKAN POINTEC ROJO	\$8.58	\$1.17	\$7.41	86.36 %
ESPIRAL 45MM X16	\$200.9	\$27.44	\$173.46	86.34 %
ESPIRAL 29MM X35				85.39 %
ESPIRAL 20MM X70	\$707.85	\$110.11	\$597.74	84.44 %
ESPIRAL 25MM X45	\$259.7	\$40.81	\$218.89	84.29 %
ACETATO A4 COPIADORA MANUAL LANCER	\$3.121.02	\$520.17	\$2.600.85	83.33 %
ESPIRAL 33MM X25				81.91 %
JD TABLA ENHEBRAR SET X24				81.61 %
LIENZO 1.8X1M				80.73 %
IMAN RECTANGULAR 27X17 GRANDE	\$87.84	\$17.08	\$70.76	80.56 %
PAPEL BOND A3 75G X1	\$23.85	\$4.77	\$19.08	80.00 %

Fig. 54. Rentabilidad por producto

Es importante aclarar que algunos valores solamente estarán con el símbolo representativo de moneda “\$”, es decir anonimizados, debido a que se encuentra de por medio un acuerdo de confidencialidad de datos de la empresa “Sari Popular”. Donde se obtuvo el permiso de publicar ciertos valores, no todos completamente y así mismo en el cliente solo el primer nombre/apellido.

3.4 Algoritmo A priori

En el caso práctica se tomó 50 transacciones de la tabla de hechos del Data Warehouse, dichas transacciones hacen referencias a las facturas emitidas, y se tomó en cuenta los 10 productos más vendidos, de tal manera que los resultados fueron los siguientes.

1. CUAD_ESPIRAL_AC_100H_ANDALUZ=N CUAD_ESPIRAL_AC_200H=N 32 ==> PAPEL_BOND_CAJA_X2500_EXCELENT_COPY=N 31 <conf:(0.97)> lift:(1.13)
2. PAPEL_BOND_CAJA_X2500_HIGH_TECH=N PAPEL_BOND_CAJA_X2500_EXCELENT_COPY=N 37 ==> CUAD_ESPIRAL_AC_200H=N 35 <conf:(0.95)> lift:(1.13)
3. PAPEL_BOND_CAJA_X2500_HIGH_TECH=N CUAD_ESPIRAL_AC_200H=N 37 ==> PAPEL_BOND_CAJA_X2500_EXCELENT_COPY=N 35 <conf:(0.95)> lift:(1.1)
4. CUAD_ESPIRAL_AC_100H_ESTILO=N CUAD_ESPIRAL_AC_200H=N 35 ==> PAPEL_BOND_CAJA_X2500_EXCELENT_COPY=N 33 <conf:(0.94)> lift:(1.1)
5. PAPEL_BOND_CAJA_X5000_COPYLASER=N PAPEL_BOND_CAJA_X2500_EXCELENT_COPY=N 33 ==> CUAD_ESPIRAL_AC_200H=N 31 <conf:(0.94)> lift:(1.12)
6. PAPEL_BOND_CAJA_X5000_COPYLASER=N CUAD_ESPIRAL_AC_200H=N 33 ==> PAPEL_BOND_CAJA_X2500_EXCELENT_COPY=N 31 <conf:(0.94)> lift:(1.09)
7. PAPEL_BOND_CAJA_X2500_HIGH_TECH=N CUAD_ESPIRAL_AC_100H_ANDALUZ=N 32 ==> PAPEL_BOND_CAJA_X2500_EXCELENT_COPY=N 30 <conf:(0.94)> lift:(1.09)
8. CUAD_ESPIRAL_AC_200H=N 42 ==> PAPEL_BOND_CAJA_X2500_EXCELENT_COPY=N 39 <conf:(0.93)> lift:(1.08)
9. CUAD_ESPIRAL_AC_100H_ANDALUZ=N 37 ==> PAPEL_BOND_CAJA_X2500_EXCELENT_COPY=N 34 <conf:(0.92)> lift:(1.07)
10. CUAD_COSIDO_100H=N 36 ==> PAPEL_BOND_CAJA_X2500_EXCELENT_COPY=N 33 <conf:(0.92)> lift:(1.07)

Fig. 55. Resultados algoritmo A Priori

Resumiendo, en una tabla, se obtuvo las siguientes apreciaciones:

Tabla 19. Principales hallazgos

Principales hallazgos	
1	Si se adquiere un cuaderno espiral de 100 hojas y 200 hojas independientemente de la marca es probable que se adquiera una caja de papel bond x2500
2	Si se adquiere un cuaderno cosido también es probable que se compre una caja de papel bond x2500
3	En general el producto de la caja de papel bond (x2500) es la que por lo general se compra en conjunto con los otros artículos más solicitados.

3.5 Análisis de resultados de la aplicación del cuestionario SUS

Con el fin de conocer la retroalimentación del proyecto se planteó una encuesta al personal del departamento de ventas de la empresa, específicamente fueron 5 colaboradores quienes se involucran en el apoyo del proceso de toma de decisiones.

En la siguiente figura se representa los valores obtenidos por la encuesta, una vez recopilados y organizados por cada participante.

Participante	Preguntas (Valoración)									
	P.1	P.2	P.3	P.4	P.5	P.6	P.7	P.8	P.9	P.10
1	5	2	5	1	4	2	5	1	5	2
2	5	2	4	3	4	2	4	2	4	4
3	4	3	4	2	4	2	4	2	4	1
4	5	2	5	2	4	1	5	1	4	1
5	4	3	4	3	4	2	4	2	3	2

Fig. 56. Resultados recopilados por participante

Las preguntas resaltadas de color verde son las pares, las mismas que tienen un sentido positivo y las rojas, lo opuesto. Como siguiente paso se aplicó las fórmulas tanto para las preguntas positivas como negativas, según lo indica las instrucciones del cuestionario.

Suma preguntas impares		Suma preguntas pares	
X1	19	Y1	17
x2	16	Y2	12
x3	15	Y3	15
x4	18	Y4	18
x5	14	Y5	13

Fig. 57. Resultados preguntas pares e impares

Finalmente se aplicó la última fórmula para obtener la nota final por usuario y el promedio de todos.

Resultados	
R1	90
R2	70
R3	75
R4	90
R5	67.5
Promedio	78.5

Fig. 58. Resultados cuestionario SUS, por participante y promedio

Como se puede observar la nota final promedio de todos los participantes es 78.5, y comparando este valor con el rango de la escala de usabilidad (SUS) se encuentra en una categoría de Bueno, lo cual es una nota aceptable para el proyecto.

3.6 Discusión

En base a la revisión de literatura de previos trabajos de grado enfocados en el tema de inteligencia de negocios se pudo observar que hay una mayoría considerable de proyectos que no implementan o hacen mención del proceso de calidad de datos mediante un estándar como la norma ISO 25012. Sin embargo, eso no significa que el resultado de dichos proyectos sea de menor relevancia para el principal objetivo, el mismo que es apoyar en la toma de decisiones informadas.

En otro punto, también cabe recalcar que los artículos científicos y tesis revisadas usan metodologías para el desarrollo de sus respectivos proyectos de inteligencia de negocios. Son evidencias de cómo establecer un proyecto tecnológico de este tipo, si bien la metodología de Kimball es la más común, seguida por la Inmon, puedo brindar un comentario personal de que la metodología creada por (Bernabeu & García, 2017), es lo suficientemente didáctica y comprensible para desarrollar Data Marts y Data Warehouses. Desde el inicio hasta el final, muestra a detalle cada indicación a seguir, por lo cual se ha obtenido buenos y satisfactorios resultados.

Adicionalmente el trabajo más similar al presente proyecto ha sido el desarrollado por (Abad, 2020), ya que sus bases son prácticamente las mismas, inteligencia de negocios y minería de datos, además de que el ejemplo mencionado realiza la parte final del proyecto mediante microservicios y desarrollo web. Aunque considero que es un proyecto robusto y

completo de software, podría mencionar que, desde mi punto de vista, es mejor separar y definir claramente las herramientas tecnológicas para cada proyecto. Es decir, aunque se pueda usar frameworks de desarrollo web para generar reportes, o visualizaciones, el resultado personalmente, creo que no sería el más óptimo, ya que un software de inteligencia de negocios trae herramientas y complementos con el fin de aportar mayor valor al proyecto, como DAX (Data Analysis Expressions) de Power BI, que permite consultar modelos de datos complejos, o Power Service para mantener y tener actualizada la información. Por lo tanto, podría ser un punto crucial, a considerar.

Como parte de limitaciones se pueden considerar ciertos puntos.

Uno de ellos es principalmente de la fuente de datos, el proyecto partió desde reportes en archivos de Excel, sin embargo, la metodología usada recomienda trabajar con el modelo entidad – relación de la base de datos y obtener los reportes mediante consultas SQL.

Debido a un cambio de estructura en la base de datos según manifestaron los asesores, se debió comenzar a tratar los datos desde Abril 2021.

Previamente al desarrollo del proyecto se firmó un acuerdo de confidencialidad, puesto que el trabajo se publicará, la empresa solicitó no exponer los resultados abiertamente y por ello se anonimizaron algunos.

El proceso ETL desarrollado aún cuenta con varios pasos sin automatizar por completo.

En la sección de minería de datos como complemento de una solución BI solamente se implementó el algoritmo A priori.

CONCLUSIONES

- Para el diseño y construcción de un Data Warehouse se puede tomar en cuenta la metodología Hefesto, ya que se estructura de una manera organizada y sencilla de ejecutar en cada una de sus respectivas fases.
- La implementación de la Norma ISO/IEC 25012 es una herramienta útil y práctica para la evaluación de la calidad de datos, siendo un pilar importante en el proyecto, puesto que la calidad de los mismos garantiza los resultados esperados.
- La inteligencia de negocios con sus artefactos y herramientas son un soporte que las empresas pueden tomarlo como un recurso estratégico, por ello, es importante a dar a conocer el valor que un proyecto de BI aporta, tanto a nivel de innovación como de administración en las organizaciones para que pueda obtener mayor relevancia y apoyo de la alta gerencia.
- El entendimiento del negocio es un factor clave que se logró comprender a lo largo del desarrollo del proyecto, ya que esto marca las pautas, la orientación, los objetivos y la estimación de los resultados que favorablemente se esperan.
- Los proyectos de inteligencia de negocios involucran diversidad de factores, de los cuales se puede destacar la información en el tiempo debido, el personal de diferentes áreas que interactuará, la comunicación entre otros.

RECOMENDACIONES

- En base a la metodología o procesos que se elijan para el desarrollo de un proyecto de inteligencia de negocios, es recomendable que se revise previamente los recursos necesarios en cuestión de las fuentes de datos, que se pueda partir directamente de un diagrama entidad – relación, como lo recomienda la metodología para que el proyecto tenga los resultados y expectativas esperadas.
- Obtener información de años anteriores con una estructura de datos general, en caso de que se llegue a volver a cambiar. Una parte esencial de un proyecto de inteligencia de negocios es contar con la información del pasado, en este caso solamente se tiene de 2 años prácticamente, porque desde este periodo se pudo desarrollar con una misma estructura de datos.
- Se podría acordar y mediar con cualquier empresa que se trabaje un proyecto de BI para que la exposición de datos pueda ser un poco más accesible, porque en sí, los proyectos se realizan con fines académicos.
- Automatizar en la medida posible los procesos repetitivos que se realicen para mantener y gestionar el Data Warehouse, puesto que, desde la obtención de datos, existe un proceso manual que sería mejor ejecutarlo automáticamente desde el sistema transaccional al sistema generado por la solución de inteligencia de negocios.
- Sería muy importante también desarrollar más algoritmos de minería de datos, de tipo descriptivo y predictivos para complementar y formar un proyecto más robusto de inteligencia de negocios.
- Para la correcta construcción de un Data Warehouse es importante definir el alcance de un proceso a la vez, como en el presente proyecto se realizó en base al proceso de ventas, y de igual manera determinar la granularidad de la base de datos analítica en la etapa inicial.
- Se recomienda continuar con la evaluación de la calidad de datos con estándares, técnicas y procedimientos adecuados, ya que este factor influirá mayormente en los resultados del proyecto.

REFERENCIAS

- Abad, F. (2020). *Microservicios aplicados a la integración y análisis de datos orientados al cálculo de una tarifa diferencial para aparcaderos de vehículos privados. Caso de estudio: Campus Central de la Universidad de Cuenca.*
- Alarcón, D. A., Salomón, E., Profesor, D., Díaz, G. D., & Santiago, S. (2013). *Análisis de Patrones de Compra de Tiendas Retail Utilizando Business Intelligence*. 97. [http://repositorio.uchile.cl/bitstream/handle/2250/115078/Adasme A., Daniel.pdf?sequence=1](http://repositorio.uchile.cl/bitstream/handle/2250/115078/Adasme_A.,_Daniel.pdf?sequence=1)
- Bernabeu, D., & García, M. (2017). *Hefesto DataWarehousing. Guía completa de aplicación teórico-práctica; metodología de Data Warehouse*. (2a ed.).
- Burgos, F.; Michilena, J. (2015). Universidad de guayaquil. *La Evasión Tributaria E Incidencia En La Recaudación Del Impuesto a La Renta De Personas Naturales En La Provincia Del Guayas, Periodo 2009-2012.*
- Calabrese, J., Esponda, S., Pasini, A., Boracchia, M., & Pesado, P. (2019). Guía para evaluar calidad de datos basada en ISO/IEC 25012. *XXV Congreso Argentino de Ciencias de la Computación*, 694–706. <https://core.ac.uk/download/pdf/301104068.pdf>
- Cavazos, E. (2008, mayo 19). *Toma de Requerimientos en una solución de Business Intelligence*. <https://gravitar.biz/bi/toma-de-requerimientos-en-solucion-business-intelligence/>
- Chadha, A. (2021, octubre 9). *Defining Knowledge, Information, Data*. <https://www.phpkb.com/kb/article/defining-knowledge-information-data-239.html>
- Cristancho, J. (2021). *Propuesta de Metodología para el Desarrollo de Proyectos de Analítica Prescriptiva a partir de un Metaanálisis*. <https://emea.mitsubishielectric.com/ar/products-solutions/factory-automation/index.html>
- Cuartin, A. (2020, noviembre 24). *¿Qué son indicadores? Tipos y beneficios de uso*. <https://blog.lemontech.com/que-son-indicadores-y-cuantos-tipos-existen-ejemplos/>
- EvaluandoSoftware. (2021, marzo 16). *Gestión de requerimientos de un proyecto de software*. <https://www.evaluandosoftware.com/gestion-requerimientos-proyecto-software-empresarial/>
- García, J. A., & Acevedo, Á. M. (2010). Análisis para predicción de ventas utilizando minería de datos en almacenes de ventas de grandes superficies. En *Universidad Tecnológica Pereira* (Vol. 0, Número 15, pp. 1–9). <https://www.golder.com/insights/block-caving-a-viable-alternative/>
- González Pinzón, M. F., & González Sanabria, J. S. (2013). Aplicación del estándar ISO / IEC 9126-3 en el modelo de datos conceptual entidad-relación. *Uptc*, 22(35), 113–125. <https://revistas.uptc.edu.co/index.php/ingenieria/article/download/2519/2374>
- Guevara, C., Ortega, S., Guevara, V., & Quiña, J. (2016). Business Intelligence aplicado al proceso de seguimiento de graduados de la Universidad Técnica del

- Norte. *Universidad Técnica del Norte*, 17. [https://documentos.redclara.net/bitstream/10786/1064/1/Business Intelligence aplicado al proceso de seguimiento de graduados de la Universidad Técnica.pdf](https://documentos.redclara.net/bitstream/10786/1064/1/Business%20Intelligence%20aplicado%20al%20proceso%20de%20seguimiento%20de%20graduados%20de%20la%20Universidad%20T%C3%A9cnica.pdf)
- Hernández, F. (2021). *Escala cuestionario SUS*. https://www.youtube.com/watch?v=ZwtWGXZfWPk&list=PLQpJzM-q2XbQVSODwvvec1T0b_p7uFgNu&index=3
- IONOS. (2018, enero 30). *Software de data mining: las mejores herramientas*. <https://www.ionos.es/digitalguide/online-marketing/analisis-web/software-de-data-mining-las-mejores-herramientas/>
- ISO 25000. (2021). *NORMAS ISO 25000*. <https://iso25000.com/index.php/normas-iso-25000>
- Molina, D. E. (2015). *Desarrollo de un aplicativo Bussiness Intelligence para la empresa importadora TOMBAMBA S.A.*
- Parra, M. del C. (2020). *Análisis de la Situación Financiera de la Empresa "Papelería Popular" S.A de la Ciudad de Ibarra.*
- Pedraza, A. (2018, marzo 29). *Historia de la Inteligencia de Negocios* . <https://www.gestiopolis.com/historia-de-la-inteligencia-de-negocios/>
- Sahona, K. (2014). *Manual de procedimientos administrativos, contables y financieros, para la empresa Sari Papelería Popular S.A de la Ciudad de Ibarra, Provincia Imbabura.*
- Santillán Suquitana, S. R. (2018). Implementación de una Solución de Inteligencia de Negocios para la Gestión de Indicadores de Soporte Tecnológico de la Empresa Farmaenlace CIA. LTDA. En *Universidad Técnica Del Norte*.
- Signaturit. (2021, abril 14). *¿Qué es Business Intelligence (BI) y qué herramientas existen?* <https://blog.signaturit.com/es/que-es-business-intelligence-bi-y-que-herramientas-existen>
- Vélez, M. (2008). *Reinstalación y operación en línea del equipo de pasteurización usando inteligencia de negocios con la metodología hefesto.*
- Vila, D. (2019). *Detección De Patrones De Deserción Estudiantil Utilizando Técnicas Predictivas De Clasificación Y Regresión De Minería De Datos, Para La Gestión Académica De La Universidad Técnica Del Norte*. 104.

ANEXOS

Anexo A: Formato ejecutado de Test de usabilidad SUS



Encuesta dirigida a los colaboradores de la empresa "Sari Popular"

Para cada una de las siguientes afirmaciones, marque la casilla que mejor describa sus reacciones ante el sistema de Inteligencia de Negocios desarrollado

* Obligatoria

1. **Creo que me gustaría usar este sistema con frecuencia ***

	Muy en desacuerdo	En desacuerdo	Neutral	De acuerdo	Muy de acuerdo
Elija una opción	<input type="radio"/>				

2. **Encontré el sistema innecesariamente complejo ***

	Muy en desacuerdo	En desacuerdo	Neutral	De acuerdo	Muy de acuerdo
Elija una opción	<input type="radio"/>				

3. **Pienso que el sistema era fácil de usar ***

	Muy en desacuerdo	En desacuerdo	Neutral	De acuerdo	Muy de acuerdo
Elija una opción	<input type="radio"/>				

4. **Creo que necesitaría la asistencia de un técnico para poder usar este sistema ***

	Muy en desacuerdo	En desacuerdo	Neutral	De acuerdo	Muy de acuerdo
Elija una opción	<input type="radio"/>				

5. **Encontré que las diversas funciones de este sistema estaban bien integradas ***

	Muy en desacuerdo	En desacuerdo	Neutral	De acuerdo	Muy de acuerdo
Elija una opción	<input type="radio"/>				

6. **Pienso que había demasiada inconsistencia en este sistema ***

	Muy en desacuerdo	En desacuerdo	Neutral	De acuerdo	Muy de acuerdo
Elija una opción	<input type="radio"/>				

7. **La mayoría de la gente aprenderá a usar este sistema muy rápidamente ***

	Muy en desacuerdo	En desacuerdo	Neutral	De acuerdo	Muy de acuerdo
Elija una opción	<input type="radio"/>				

8. **Me pareció que el sistema era muy incómodo de usar ***

	Muy en desacuerdo	En desacuerdo	Neutral	De acuerdo	Muy de acuerdo
Elija una opción	<input type="radio"/>				

9. **Me sentí muy confiado usando el sistema ***

	Muy en desacuerdo	En desacuerdo	Neutral	De acuerdo	Muy de acuerdo
Elija una opción	<input type="radio"/>				

10. **Necesito aprender muchas cosas antes de poder ponerme en marcha con este sistema ***

	Muy en desacuerdo	En desacuerdo	Neutral	De acuerdo	Muy de acuerdo
Elija una opción	<input type="radio"/>				