Construcción de un almacén de datos especializado para el área de estadística del Distrito de Salud 17d07 que sirva de soporte para la toma de decisiones utilizando herramientas de integración de datos

Bastidas Santamaría, Juan Gabriel

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Postgrados

Maestría en Gestión de Sistemas de Información e Inteligencia de Negocios

Trabajo de titulación, previo a la obtención del título de Magíster en Gestión de Sistemas de Información e Inteligencia de Negocios

Msc. Campaña Ortega, Eduardo Mauricio

23 de noviembre 2020

Tesis_Juan_Bastidas_Maestria_BI_actualizado-signed-signed....

Scanned on: 22:45 January 29, 2022 UTC

Overall Similarity Score

Results Found

Total Words in Text

Identical Words	1241
Words with Minor Changes	442
Paraphrased Words	633
Ommited Words	0

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN

Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**Construcción de un almacén de datos especializado para el área de estadística del Distrito de Salud 17d07 que sirva de soporte para la toma de decisiones utilizando herramientas de integración de datos**" fue realizado por el señor **Bastidas Santamaría Juan Gabriel**, el mismo que ha sido revisado y analizado en su totalidad, por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 23 de noviembre de 2020

Firmado electrónicamente por:
EDUARDO MAURICIO
CAMPAÑA ORTEGA

Msc. Campaña Ortega, Eduardo Mauricio

Director

C.C.: 1708856701

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

RESPONSABILIDAD DE AUTORÍA

Yo, **Bastidas Santamaría Juan Gabriel**, con cédula de ciudadanía No. 1717434524, declaro que el contenido, ideas y criterios del trabajo de titulación, **"Construcción de un almacén de datos especializado para el área de estadística del Distrito de Salud 17d07 que sirva de soporte para la toma de decisiones utilizando herramientas de integración de datos"** es de mi autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 23 de noviembre de 2020

Firmado electrónicamente por:
JUAN GABRIEL BASTIDAS SANTAMARIA

Ing. Bastidas Santamaría, Juan Gabriel

C.C.: 1717434524

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORIZACIÓN DE PUBLICACIÓN

Yo, **Bastidas Santamaría Juan Gabriel**, con cédula de ciudadanía No. 1717434524, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **"Construcción de un almacén de datos especializado para el área de estadística del Distrito de Salud 17d07 que sirva de soporte para la toma de decisiones utilizando herramientas de integración de datos"** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 23 de noviembre de 2020

Firmado electrónicamente por:
JUAN GABRIEL BASTIDAS SANTAMARIA

Ing. Bastidas Santamaría, Juan Gabriel

C.C.: 1717434524

DEDICATORIA

Dedico este trabajo a:

A Dios por darme la oportunidad de seguir estudiando, por brindarme salud y fuerza para conseguir mis objetivos.

A mis padres quienes son los pilares fundamentales para mis logros personales y profesionales, especialmente a mi madre que siempre ha estado pendiente de los estudios de sus hijos.

A mi hermano por ser la persona que me ha enseñado que sin importar las dificultades siempre hay que seguir adelante, que con esfuerzo, dedicación y fe en Dios se pueden alcanzar grandes logros.

Bastidas Santamaría, Juan Gabriel

AGRADECIMIENTO

Gracias a Dios por la salud que me ha brindado para cumplir otra meta en mi vida de la mano de mis seres queridos.

Gracias a mis padres y hermano por sus enseñanzas y valores inculcados en mi vida, que han servido para ir con pasos firmes en búsqueda de mis objetivos.

Gracias a mi familia Salazar Santamaría y Marcelo Jurado por su apoyo incondicional en todo momento.

Gracias a mi novia Carolina Espinosa que durante el tiempo de mis estudios de postgrado estuvo apoyándome incondicionalmente.

Mi más sincero agradecimiento al Ing. Mauricio Campaña, que durante mis estudios de pregrado y postgrado ha sido un apoyo y ejemplo como persona y profesional.

Agradezco a todos los docentes que impartieron sus conocimientos y experiencias en la “Maestría en Gestión de Sistemas de Información e Inteligencia de Negocios”.

Firmado electrónicamente por:
JUAN GABRIEL BASTIDAS SANTAMARIA

Bastidas Santamaría, Juan Gabriel

Tabla de contenido

CERTIFICACIÓN 3

RESPONSABILIDAD DE AUTORÍA 4

AUTORIZACIÓN DE PUBLICACIÓN 5

DEDICATORIA 6

AGRADECIMIENTO 7

Glosario 16

Resumen 18

Abstract 19

Capítulo I 20

Introducción 20

- Planteamiento del problema 20
- Antecedentes 21
- Justificación e importancia 22
- Alcance y preguntas de investigación 23
- Objetivos 25
  - *Objetivo General 25*
  - *Objetivos Específicos 25*
- Hipótesis 25
  - *Señalamiento de variables 26*
  - *Categorización de las variables de investigación 26*

Capítulo II 27

Marco Teórico 27

Inteligencia de Negocios (BI) 27
Características de BI *28*
Beneficios de BI *29*
Componentes Básicos de BI *30*
Técnicas de Integración de Datos 30
ETL (extracción, transformación y carga) *32*
ELT (extracción, carga y transformación) *33*
Data Warehouse 35
Data Mart 35
Bases de Datos OLTP y OLAP 37
OLTP – On Line Transactional Processing *37*
OLAP – On Line Analytical Processing *37*
Data warehouse vs Data mart 38
Construcción de un almacén de datos 38
Metodologías: Kimball e Inmon *40*
Cuadro comparativo metodología Kimball vs metodología Inmon *46*
Análisis OLAP 46
Modelado Multidimensional 48
Tablas del Modelo Multidimensional *49*
Esquemas del Modelo Multidimensional *50*
Sistemas Gestores de Bases de Datos 52
PostgreSQL *52*
Oracle *53*
MySQL *53*

Herramientas para procesos ETL 54
Informática PowerCenter *54*
IBM InfoSphere DataStage *55*
Pentaho Data Integration *56*
Cuadro comparativo entre Informática PowerCenter, IBM DataStage y Pentho DI *57*
Herramientas de soporte a la toma de decisiones 58
Power BI *58*
Tableau *59*
Qlik *60*
Cuadro comparativo entre Power BI, Tableau y Qlik *61*
Capítulo III 62
Construcción de la Solución y Visualización de los Resultados 62
Planificación del proyecto 62
Definición del proyecto *62*
Objetivos y alcance del proyecto *62*
Situación actual *63*
Estrategia de implementación *63*
Selección de la metodología de desarrollo *63*
Tiempo e Inversión *64*
Roles *65*
Definición de requerimientos del negocio 65
Requerimientos Funcionales *66*
Requerimientos No Funcionales *67*
Modelo Físico de la base de datos transaccional sgmas en PostgreSQL *68*

Diseño técnico de la arquitectura....69
Selección de productos e instalación....70
Modelado Dimensional....70
Diseño del Modelo Estrella....71
Modelo gráfico de alto nivel....77
Diseño Físico....78
Diseño e Implementación....79
Proceso de extracción, transformación y carga (ETL)....79
Creación del trabajo ETL_SALUD....80
Unir data rdacaa pras....80
Clean fact y dimensiones....82
Cargar dimensiones....82
Cargar hecho....90
Resultado de carga de dimensiones y hecho....91
Especificación de Aplicaciones de BI....92
Desarrollo de Aplicaciones de BI....93
Reportes basados en los requerimientos funcionales....93
Implementación....99
Mantenimiento y Crecimiento....100
Gestión del Proyecto....100
Capítulo IV....101
Conclusiones y Recomendaciones....101
Conclusiones....101

Recomendaciones 102

Bibliografía 103

LISTADO DE TABLAS

Tabla 1. Diferencias entre data warehouse y data mart 38

Tabla 2. Cuadro comparativo Kimball vs Inmon 46

Tabla 3. Cuadro comparativo herramientas ETL 57

Tabla 4. Cuadro comparativo herramientas de visualización de datos 61

Tabla 5. Costo de hardware 64

Tabla 6. Costo de software 64

Tabla 7. Recurso humano 64

Tabla 8. Otros gastos 64

Tabla 9. Roles 65

Tabla 10. Requerimientos funcionales 66

Tabla 11. Requerimientos no funcionales 67

Tabla 12. Selección de productos e instalación 70

Tabla 13. Dimensiones 71

Tabla 14. dim_agenda_medica 72

Tabla 15. dim_canton 72

Tabla 16. dim_especialidad 73

Tabla 17. dim_grupo_cultural 73

Tabla 18. dim_motivo_atencion 73

Tabla 19. dim_paciente 73

Tabla 20. dim_parroquia....74

Tabla 21. dim_provincia....74

Tabla 22. dim_personal_medico....74

Tabla 23. dim_registro_paciente....75

Tabla 24. dim_sexo....75

Tabla 25. dim_tiempo....75

Tabla 26. Tabla de hechos o fact....76

Tabla 27. Tecnología disponible para la implementación....100

LISTADO DE FIGURAS

Figura 1. Diagrama Causa-Efecto....21

Figura 2. Business Intelligence (BI)....28

Figura 3. Extracción, Transformación y Carga (ETL)....33

Figura 4. Extracción, Carga, Transformación (ELT)....34

Figura 5. Data Mart y Data Warehouse....36

Figura 6. Botton-Up (Kimball)....42

Figura 7. Proceso Kimball, Business Dimensional....42

Figura 8. Top-Down (Inmon)....44

Figura 9. Operadores OLAP....47

Figura 10. Modelo Multidimensional....48

Figura 11. Esquema Estrella....50

Figura 12. Esquema Copo de Nieve....51

Figura 13. Modelo Físico Base de Datos Transaccional SGMAS....68

Figura 14. Arquitectura técnica Data Mart 69
Figura 15. Modelo Estrella Data Mart 71
Figura 16. Modelo bubble del indicador turno 77
Figura 17. Diseño físico Data Mart 78
Figura 18. Pantalla de entrada de spoon 79
Figura 19. Job - ETL_SALUD 80
Figura 20. Job - unir data rdacaa pras 80
Figura 21. Transformación - Cargar data rp 81
Figura 22. SQL - Clean fact y dimensiones 82
Figura 23. Job – Cargar dimensiones 83
Figura 24. Transformación – dim provincia 83
Figura 25. Transformación – dim cantón 84
Figura 26. Transformación – dim parroquia 84
Figura 27. Transformación – dim paciente 85
Figura 28. Transformación – dim registro paciente 86
Figura 29. Transformación – dim especialidad 87
Figura 30. Transformación – dim personal médico 87
Figura 31. Transformación – dim agenda médica 88
Figura 32. Transformación – motivo atención 89
Figura 33. Transformación – dim tiempo 89
Figura 34. Transformación – fact turno 90
Figura 35. Tablas de dimensiones y tabla de hecho 91
Figura 36. Datos cargados dim paciente 92

Figura 37. Datos cargados fact turno....92

Figura 38. Total de pacientes registrados por tipos de fuentes de datos 93

Figura 39. Cantidad de pacientes registrados por nacionalidad....94

Figura 40. Cantidad de Pacientes registrados por grupo étnico....95

Figura 41. Cantidad de turnos agendados por mes y trimestres año 2019 95

Figura 42. Cantidad de pacientes agendados por Establecimientos de Salud 96

Figura 43. Cantidad de turnos agendados por especialidad 97

Figura 44. Cantidad de pacientes que asisten a las agendas por día y establecimiento de salud....98

Figura 45. Médicos con mayor agendamiento por mes, establecimiento de salud y especialidad 99

Glosario

- **Data Warehouse:** Almacén de datos o base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla.
- **Data Mart:** Almacén de datos para un área o tema específico. En otras palabras se trata de un Data Warehouse departamental.
- **BI:** Business Intelligence (Inteligencia de negocio)
- **SGBD:** Sistema de gestión de base de datos
- **ETL:** Extract, Transform and Load (Extraer, transformar y cargar). Refiere a la transformación de los datos.
- **DDL:** Data Definition Language (Lenguaje de definición de datos)
- **DML:** Data Manipulation Language (Lenguaje de manipulación de datos)
- **OLTP:** On-Line Transaction Processing (Procesamiento transaccional en línea)
- **OLAP:** On Line Analytical Processing (Procesamiento analítico en línea)
- **DW:** Referencia a Data Warehouse
- **Multidimensionalidad:** Capacidad que ofrece una herramienta de Inteligencia de Negocios para analizar la información utilizando distintas dimensiones a la vez.
- **Query:** Cadena de consulta, interacción con una base de datos.
- **Desnormalización:** La desnormalización es el proceso inverso del proceso de normalización. La desnormalización funciona agregando datos redundantes o agrupando datos para optimizar el rendimiento.
- **Software Open Source:** Software de código abierto
- **Dashboard:** Tablero de Control
- **Arquitectura:** Diseño de una solución de Inteligencia de Negocios

- **Granularidad:** Consiste en el nivel de detalle de la información al que se decide descender para el análisis de los modelos.
- **Dimensión:** Perspectiva que contextualiza una medida. Consiste en la agrupación de elementos con características comunes, tales como cliente, producto, tiempo, país, etc.
- **Hecho:** Contiene las claves subrogadas de aquellas dimensiones que definen su nivel de detalle y los indicadores.
- **Medida:** Valor, generalmente numérico, que cuantifica la intersección de dimensiones.
- **Modelo:** La representación de una porción de la realidad en sus elementos más pertinentes a la solución del problema.

Resumen

En la actualidad, la información se ha convertido en uno de los activos más valiosos para las empresas. El Distrito de Salud 17d07 está conformado por 18 establecimientos, los cuales generan datos diariamente en la atención de sus pacientes. Estos datos son almacenados en diferentes fuentes a través de distintos sistemas informáticos. Las fuentes de datos son analizadas por el área de Estadística para la generación de reportes de información. Sin embargo, se ha detectado que, al tener datos provenientes de diferentes fuentes, el área de Estadística no analiza la totalidad de los datos, debido a que no realizan la integración de fuentes de datos heterogéneos, ocasionando que no se genere un acertado análisis de información. Para hacer frente a esta problemática, se construyó un almacén de datos especializado para el área de Estadística en el cual se integraron las diferentes fuentes de datos heterogéneos, obteniendo un repositorio de información centralizado, uniforme y confiable, para un adecuado análisis de información y oportuna toma de decisiones utilizando herramientas de integración de datos. Para alcanzar este objetivo se planteó la metodología AD-HOC, la cual consta de 4 fases: Identificación de la situación actual, estudio de viabilidad de la solución, construcción de la solución y validación. Con la implementación del almacén de datos se logró la optimización de consultas facilitando el manejo dinámico de reportes de información. Para la construcción de la solución se utilizó la metodología de Kimball y mediante el proceso ETL se logró mejorar la calidad de los datos y centralizar toda la información.

Palabras clave:

- **INTELIGENCIA DE NEGOCIOS**
- **ALMACÉN DE DATOS**
- **PROCESO ETL**

Abstract

Today, information has become one of the most valuable assets for companies. The Health District 17d07 is made up of 18 establishments, which generate data daily in the care of their patients. These data are stored in different sources through different computer systems. The data sources are analyzed by the Statistics area for the generation of information reports. However, it has been detected that, by having data from different sources, the Statistics area does not analyze all the data, due to the fact that they do not integrate heterogeneous data sources, causing an accurate analysis of data to be generated. information. To deal with this problem, a specialized data warehouse was built for the Statistics area in which the different heterogeneous data sources were integrated, obtaining a centralized, uniform and reliable information repository, for an adequate and timely analysis of information. decision making using data integration tools. To achieve this objective, the AD-HOC methodology was proposed, which consists of 4 phases: Identification of the current situation, feasibility study of the solution, construction of the solution and validation. With the implementation of the data warehouse, query optimization was achieved, facilitating the dynamic management of information reports. For the construction of the solution, the Kimball methodology was used and through the ETL process it was possible to improve the quality of the data and centralize all the information.

Keywords:

- **BUSINESS INTELLIGENCE**
- **DATA MART**
- **ETL PROCESS**

Capítulo I

Introducción

Planteamiento del problema

Hoy en día uno de los principales problemas en toda empresa o institución es la gestión de volúmenes de información, y la forma de explotar dicha información para lograr soporte de las decisiones financieras, administrativas y económicas. Es así que toda empresa actualmente debe mantener un control de la información generada día a día para poder tomar decisiones de una forma óptima (Haro, Pérez, Guzman, & Saquicela, 2014).

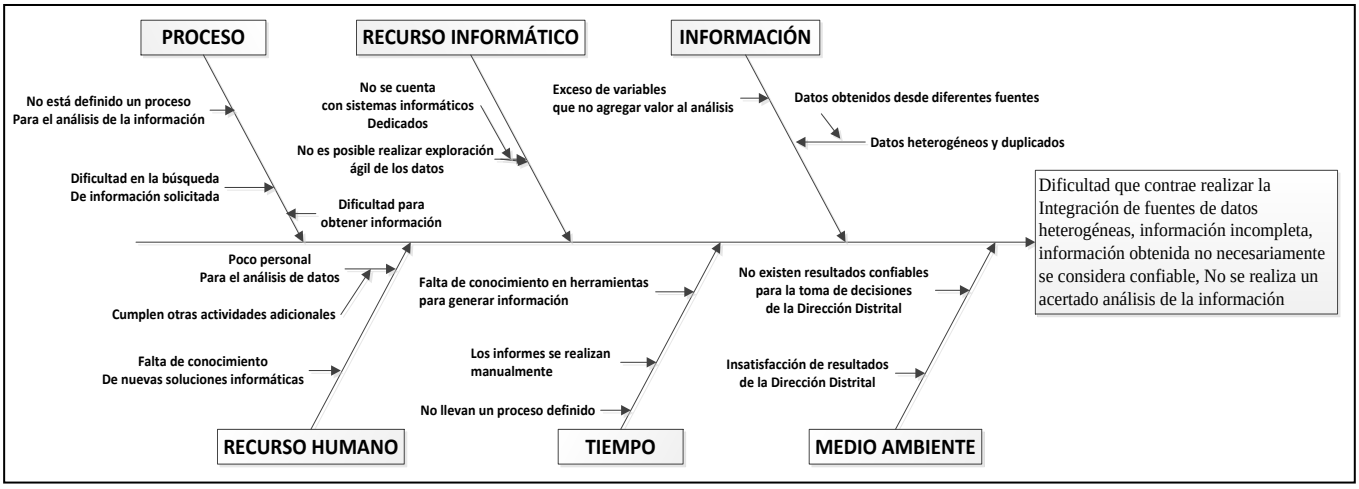
Los 18 establecimientos de salud pertenecientes al Distrito 17d07 del MSP generan datos diariamente mediante la atención de sus pacientes. Estos datos son almacenados en diferentes fuentes y archivos a través de distintos sistemas informáticos.

Actualmente, el área de Estadística Distrital realiza el proceso de análisis y generación de información desde una sola fuente de datos, la misma que es contener el mayor número de registros diarios, este proceso ocasiona que la información obtenida no necesariamente se considere confiable y completa, debido a que no se integran todas las fuentes de datos heterogéneos, por tanto, no se realiza un acertado análisis de la información que ayude a una adecuada toma de decisiones.

Con base en lo expuesto anteriormente se puede analizar la problemática e identificar sus principales causas y efectos. Lo cual se presenta en la Figura 1.

Figura 1.

Diagrama Causa-Efecto



Nota. Diagrama Causa-Efecto para establecer el planteamiento del problema.

Antecedentes

Desde el inicio de la era de la información las empresas necesitan explotar su mayor recurso, la información. La explotación eficiente de la información permite una rápida, acertada y oportuna toma de decisiones bajo el manejo de datos confiables (Yalan, 2013).

La inteligencia de negocios (BI) se utiliza para mejorar las capacidades de toma de decisiones para los procesos de gestión. Con un BI sólido, las empresas pueden respaldar decisiones con algo más que una simple sensación. Satisfacer las necesidades de BI depende de los diseños de almacenamiento de datos y los flujos de trabajo de integración de datos. El almacenamiento de datos enfatiza la captura de datos de diversas fuentes para un análisis y un acceso útil de la información (De Pietro, y otros, Integrating Trajectory Data in the Warehousing Chain: A New Way to Handle the Trajectory ELT Process, 2018).

La integración de datos es el proceso que permite combinar datos heterogéneos de muchas fuentes diferentes en la forma y estructura de una única aplicación. Este proceso

de integración de datos facilita que diferentes tipos de datos, tales como matrices de datos, documentos y tablas, sean fusionados por usuarios, organizaciones y aplicaciones para un uso personal, de procesos de negocio o de funciones (Power Data, 2019).

En 1988, Barry Deylin y Paul Murphy de IBM publicaron un artículo titulado “Una arquitectura para un sistema de negocios e información” que introdujo el concepto de un almacén de datos. Bill Inmon publicó “Construyendo el almacén de datos” en 1992, la cual discutió el diseño e implementación de un gran almacén de datos en toda la empresa, y en 1996, Ralph Kimball introdujo modelos de datos y mercados de datos en su libro “El kit de herramientas de almacenamiento de datos”. Por lo tanto, los almacenes de datos han existido desde hace más de 25 años y son una metodología establecida para la consolidación, organización y procesamiento de datos (Bhushan, 2016).

El propósito de un almacén de datos es proporcionar información para facilitar la toma de decisiones empresariales. El interés está en analizar el estado de la empresa en un momento t (esto puedo incluir datos actuales, así como datos históricos) (Prakash & Prakash, 2018).

El Distrito de Salud 17d07 pertenece al Ministerio de Salud Pública del Ecuador está ubicado en la ciudad de Quito y comprende 18 establecimientos de salud, cubriendo una población aproximada de 400.000 habitantes. Estos establecimientos de salud generan datos diariamente mediante la atención de sus pacientes, los cuales son almacenados en diferentes fuentes de datos a través de distintos sistemas informáticos, estos datos son utilizados por el área de Estadística Distrital para el análisis y generación de información para la toma de decisiones.

Justificación e importancia

El área de Estadística Distrital es la encargada de receptar y analizar las diferentes fuentes de datos proporcionadas por los 18 establecimientos de salud. Estas fuentes se

encuentran en diferentes formatos como .xls, .csv, .sql, entre otros, obtenidos de distintos sistemas informáticos. Todos estos datos son generados en la atención de los pacientes mediante el registro de sus historias clínicas.

En la actualidad, el área de Estadística Distrital no integra todas las fuentes de datos para el análisis y toma de decisiones, debido a la dificultad que genera trabajar con datos heterogéneos. El proceso de análisis de información lo realizan con fuentes de datos que tienen el mayor número de registros, excluyendo otras fuentes de información importantes. Este proceso ocasiona que los reportes de información obtenidos no necesariamente se consideren confiables y completos, por tanto, no se realiza un acertado análisis de información para una adecuada toma de decisiones.

Por ellos es importante la construcción de un almacén de datos en el área de Estadística Distrital ya que permitirá integrar todas las fuentes de datos heterogéneos en un único repositorio, el mismo que servirá de fuente de información centralizada, con datos reales y completos, obteniendo un acertado análisis de información y una adecuada toma de decisiones, garantizando además la elaboración de informes en menor tiempo.

Según lo expuesto en las secciones anteriores, se define que, una solución completa de integración de datos ofrece datos confiables de una variedad de fuentes. Las soluciones de integración de datos ayudan a comprender, limpiar, monitorizar, transformar y entregar datos para que puedan estar seguros de que la fuente de información es confiable, consistente y real.

Alcance y preguntas de investigación

La integración de datos se implementa generalmente en un almacén de datos mediante software especializado que aloja grandes repositorios de datos de recursos internos y externos. Los datos se extraen, se mezclan y se presentan de forma unificada (Power Data, 2015).

El presente proyecto tiene como alcance diseñar y construir un almacén de datos especializado para el área de Estadística Distrital en el cual se integren las diferentes fuentes de datos heterogéneos, obteniendo un repositorio de información centralizado, uniforme y confiable que ayude a un acertado análisis de información para una adecuada toma de decisiones mediante la generación de reportes de información, utilizando herramientas de integración de datos.

Para cumplir con los objetivos específicos deseados de este proyecto, se plantearon las siguientes preguntas de investigación:

- OE1-RQ1: ¿Qué resultados se han obtenido en cuanto a la construcción de un almacén de datos departamental en otras instituciones?
- OE1-RQ2: ¿Qué información se debe considerar para la construcción de un almacén de datos departamental?
- OE2-RQ1: ¿Cuáles son las ventajas de la integración de fuentes de datos heterogéneos para la construcción de un almacén de datos departamental?
- OE2-RQ2: ¿Cuáles son las limitaciones encontradas en el proceso de integración de fuentes de datos heterogéneos que dificultan la construcción de un almacén de datos?
- OE3-RQ1: ¿Qué técnicas de integración de fuentes de datos heterogéneos son las más utilizadas?
- OE3-RQ2: ¿Qué metodologías y herramientas pueden usarse para realizar la construcción de un almacén de datos departamental en una organización?
- OE4-RQ1: ¿Mediante el almacén de datos departamental, es posible facilitar el análisis de información y generación de reportes?
- OE4-RQ2: ¿A través de un almacén de datos departamental, es posible garantizar una adecuada y oportuna toma de decisiones?

Objetivos

Objetivo General

Construir un almacén de datos especializado para el área de Estadística del Distrito de Salud 17d07 en el cual se integren todas las fuentes de datos heterogéneas, obteniendo un repositorio de información centralizado, uniforme y confiable que apoye a una adecuada toma de decisiones mediante la generación de reportes de información, utilizando herramientas de integración de datos.

Objetivos Específicos

OE1: Analizar la información de las diferentes fuentes de datos heterogéneas e identificar los datos requeridos para la construcción del almacén de datos mediante entrevistas con el área de Estadística Distrital.

OE2: Determinar que metodologías y herramientas existen para la integración de fuentes de datos heterogéneas en la construcción de un almacén de datos mediante una revisión preliminar de literatura.

OE3: Diseñar y construir un almacén de datos especializado para el área de Estadística Distrital, mediante el uso de herramientas de integración de datos, aplicando una metodología de data warehouse.

OE4: Evaluar los resultados obtenidos de la construcción del almacén de datos para determinar si existe un adecuado análisis de información y toma de decisiones mediante la generación de reportes de información en el área de Estadística Distrital.

Hipótesis

La construcción de un almacén de datos departamental mediante la integración de las fuentes de datos heterogéneas permitirá al área de Estadística Distrital obtener datos reales, completos y confiables, que ayuden a un acertado análisis de información y una

adecuada toma de decisiones, optimizando además el tiempo de elaboración de reportes de información.

Señalamiento de variables

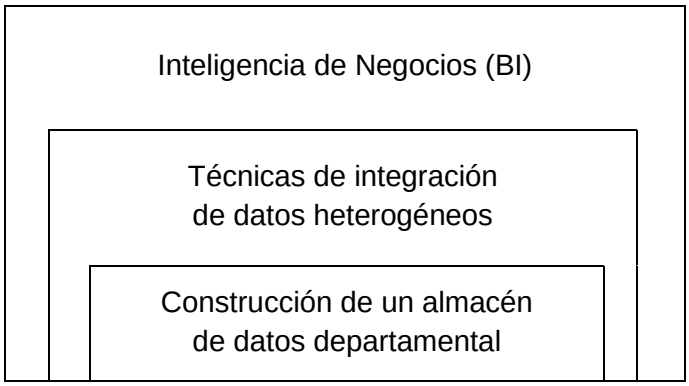
- Variable Independiente: construcción de un almacén de datos departamental

- Variable Dependiente: obtener datos reales, completos y confiables para un adecuado análisis

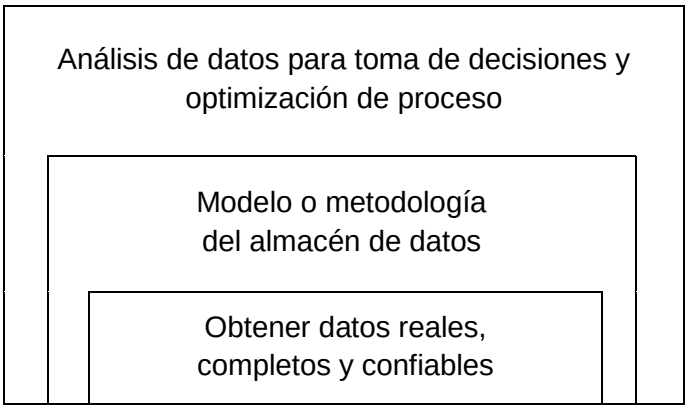
Categorización de las variables de investigación

Con el objetivo de buscar la congruencia en la fundamentación teórica del presente proyecto, se establece la siguiente red de categorías:

- Variables Independientes



- Variables Dependientes



Capítulo II

Marco Teórico

Inteligencia de Negocios (BI)

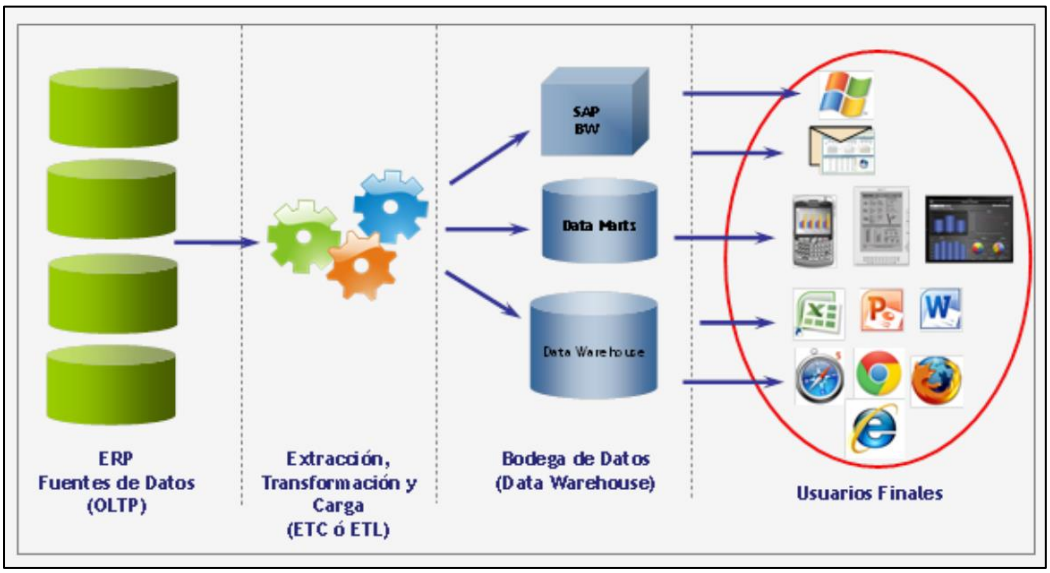
La inteligencia de negocios (Business Intelligence) se utiliza para mejorar las capacidades de toma de decisiones para los procesos de gestión. Con un BI sólido, las empresas pueden respaldar decisiones con algo más que una simple sensación. Satisfacer las necesidades de BI depende de los diseños de almacenamiento de datos y los flujos de trabajo de integración de datos. El almacenamiento de datos enfatiza la captura de datos de diversas fuentes para un análisis y un acceso útil de la información.

La realización del objetivo de BI requiere la transformación de estos datos sin procesar, recopilados en datos analíticos. Gracias a sus características el proceso de extracción, transformación y carga (ETL) presenta el núcleo de la cadena de almacenamiento. Este último extrae los datos de los sistemas fuente y los coloca en un almacén de datos (De Pietro, y otros, Integrating Trajectory Data in the Warehousing Chain: A New Way to Handle the Trajectory ELT Process, 2018).

Una definición más amplia es la que proponen en The data warehouse institute: “Business Intelligence (BI) es un término paraguas que abarca los procesos, las herramientas, y las tecnologías para convertir datos en información, información en conocimiento y planes para conducir de forma eficaz las actividades de los negocios”. BI abarca las tecnologías de data warehousing, los procesos en el ‘back end’, consultas, informes, análisis y las herramientas para mostrar información y los procesos en el ‘fron end’ (LatinoBI, 2013).

Figura 2.

Business Intelligence (BI)



Nota. Solución de Business Intelligence (BI) *(LatinoBI, 2013)*.

En definitiva, una solución BI completa debería permitir valorar aspectos importantes como:

- Observar ¿Qué está ocurriendo?
- Comprender ¿Por qué ocurre?
- Predecir ¿Qué ocurriría?
- Colaborar ¿Qué debería hacer el equipo?
- Decidir ¿Qué camino se debe seguir?

Características de BI

1. La primera fase por la que se atraviesa es la exploración, para lograr comprender que sucede en el negocio y descubrir nuevas relaciones que eran desconocidas.
2. Establecer relaciones entre variables, tendencias, es decir, cuál puede ser la evolución de la variable o patrones.

3. Es un proceso interactivo.
4. La información está almacenada en tablas relacionadas entre ellas.
5. Las tablas tienen registros y cada uno de los registros tiene distintos valores para cada uno de los atributos, mismas que están almacenadas en el Data Warehouse.
6. Permite y soporta un conjunto amplio de capacidades, tales como planificación financiera, presupuesto, previsión, monitoreo en tiempo real y análisis avanzado, que también son impactadas por otras tecnologías y programas: almacenamiento de datos, calidad, integración, gobernabilidad.
7. Comunica los resultados y efectúa los cambios pertinentes en la organización para mejorar su competitividad.
8. Su origen va ligado a proveer acceso directo a la información a los usuarios del negocio para ayudarles en la toma de decisiones, sin intervención de los departamentos de Sistemas de Información (Cano, 2006).

Beneficios de BI

9. Uno de los objetivos básicos de los sistemas de información es que ayuden a la toma de decisiones. Cuando un responsable tiene que tomar una decisión pide o busca información, que le servirá para reducir la incertidumbre. Sin embargo, aunque todos la utilicen, no todos los responsables recogen la misma información: depende de muchos factores, como pueden ser su experiencia, formación, disponibilidad, etc.
10. Beneficios tangibles: Reducción de costos, generación de ingresos, reducción de tiempos para las distintas actividades del negocio.
11. Beneficios intangibles: El hecho de que se tenga disponible la información para la toma de decisiones hará que más usuarios utilicen dicha información para

tomar decisiones y mejorar nuestra posición competitiva. Por ejemplo, optimizar la atención a los clientes, información más actualizada, mayor integración de la información, etc.

12. Beneficios estratégicos: Todos aquellos que facilitan la formulación de la estrategia, es decir, a qué clientes, mercados o con qué productos dirigirnos. Por ejemplo, aumentar el valor de mercado, dar soporte a las estrategias, mayor visibilidad de la gestión, mejorar la toma de decisiones, realizándola de forma más rápida, informada y basada en hechos (Cano, 2006).

Componentes Básicos de BI

13. Problemática empresarial a la que se quiere dar respuesta.
14. Un equipo de personas o una persona que lleve a cabo el análisis.
15. Información de los servicios o productos que ofrece la empresa.
16. Información externa de las empresas de la competencia.
17. Una base de datos a la cual se la llama Data Warehouse.
18. Una aplicación de BI que permita trabajar con la información, analizarla y visualizar los resultados (Cano, 2006).

Técnicas de Integración de Datos

La integración de datos es una combinación de tecnología y procesos de negocio que se utilizan para consolidar datos dispares procedentes de diferentes fuentes de datos de forma que se pueda obtener información valiosa para la organización.

Los datos para integrar pueden encontrarse tanto de bases de datos modernas como heredadas, aplicaciones instaladas en ordenadores de sobremesa, comentarios en redes sociales, artículos de blog, sensores de máquinas, y más.

La mayoría de las soluciones de integración de datos modernas están preparadas para integrar esos tipos de datos, así como otras nuevas fuentes de datos que todavía están por aparecer. Además, ayudan a entender, limpiar, monitorizar, transformar y proporcionar datos confiables, consistentes y gobernados en tiempo real (Colombia Digital, 2017).

Los casos de uso para la integración de datos son amplios y varían según las necesidades de la empresa, el volumen y la complejidad de los datos. Por ejemplo:

- Un centro de salud puede necesitar software de integración de datos para consolidar y administrar sus datos de pacientes y empleados de múltiples fuentes en tiempo real.
- Un negocio de compra y venta de vehículos en línea puede necesitarlo para actualizar millones de registros diariamente y reducir el tiempo de incorporación del cliente de meses a horas mediante la asignación de los datos del cliente a la base de datos de la compañía.
- Una oficina de inversiones puede necesitarlo para mapear los datos de dotación de la institución de sistemas de fuentes dispares (incluidos los sistemas internos y los administradores de dinero externos) en un programa de software de seguimiento para el análisis de riesgos.

Para cada caso de uso, se puede construir un proceso para automatizar tareas manuales y agilizar los procesos para la precisión. Y mientras que las necesidades específicas pueden variar, en su núcleo, la integración de datos cubre los procesos de combinación, limpieza y traslado de datos desde la fuente (s) a destino, todo lo cual se puede hacer utilizando diferentes enfoques (Astera Software, 2019).

Entre las principales técnicas de integración de datos se tiene:

ETL (extracción, transformación y carga)

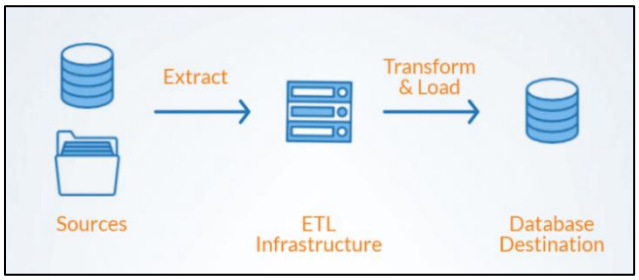
Comúnmente denominado ETL, que trata del proceso de extracción de datos de los sistemas de origen (operacionales) y ponerlos en el data warehouse o data mart.

ETL implica las siguientes tareas:

1. **Extracción de datos:** Es lo primero que hace una herramienta ETL. Se trata de obtener la información de las distintas fuentes de origen, tanto internas como externas. Durante la extracción, se identifica los datos deseados y se extrae de muchas fuentes diferentes, incluyendo los sistemas de bases de datos y aplicaciones. Después de la extracción de datos, tienen que ser transportados físicamente al sistema destino o a un sistema intermedio para su posterior procesamiento y/o transformación.
2. **Transformación de datos:** es el filtrado, limpieza, depuración, homogeneización y agrupación de la información. Incluye la agrupación de los datos de las diferentes fuentes. La transformación se produce mediante el uso de reglas o tablas de consulta o mediante la combinación de los datos con otros datos.
3. **Carga:** es el proceso de escribir los datos en la Data warehouse. La fase de carga es el momento en el cual los datos de la fase anterior (transformación) son cargados en el sistema de destino (Área Tecnológica, 2016). Una vez se encuentra en el almacén de datos, esta pueda ser consultada, compartida o analizada por el personal de la empresa.

Figura 3.

Extracción, Transformación y Carga (ETL)



Nota. Proceso de Extracción, Transformación y Carga (ETL) (Astera Software, 2019)

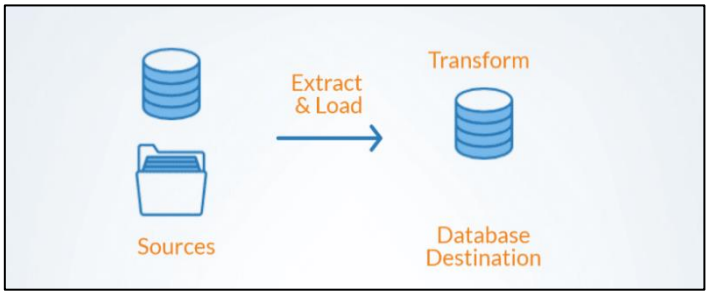
En este enfoque, se extraen los datos, se aplica la lógica de transformación y los datos resultantes se cargan en la base de datos de destino o en el destino del lago de datos. Debido a la amplia disponibilidad de marcos y herramientas que admiten ETL, este enfoque es ideal para las empresas que necesitan integran y procesar grandes volúmenes de datos, aunque el tiempo de procesamiento es mayor para volúmenes más grandes (Astera Software, 2019).

ELT (extracción, carga y transformación)

En esta técnica, los datos extraídos se cargan primero en el destino objetivo y la lógica de transformación se aplica dentro de la base de datos o el almacén de datos. Debido a que la infraestructura ETL se elimina de la ecuación y la transformación ocurre directamente dentro de la base de datos, la potencia total consumida por el sistema y la latencia de los datos se reducen significativamente (Astera Software, 2019).

Figura 4.

Extracción, Carga, Transformación (ELT)



Nota. Proceso de Extracción, Carga, Transformación (ELT) *(Astera Software, 2019)*

ELT implica las siguientes tareas:

4. **Extracción de datos:** Se trata de obtener la información de las distintas fuentes de origen, tanto internas como externas. Durante la extracción, se identifica los datos deseados y se extrae de muchas fuentes diferentes, incluyendo los sistemas de bases de datos y aplicaciones.
5. **Carga:** aquí es donde el ELT se desvía de su pariente cercano, el ETL. En lugar de suministrar todo este volumen de datos en bruto y cargarlos a un servidor de procesamiento provisional para su transformación, el ELT los entrega en su conjunto al punto donde acabarán residiendo. Esto acorta el ciclo entre la extracción y la entrega, pero exige mucho más trabajo previo hasta poder sacar partido a los datos.
6. **Transformación:** la base de datos o el almacén de datos clasifica y normaliza los datos, conserva una parte o la totalidad a mano y accesible para elaborar informes personalizados. Los gastos generales de almacenar esta cantidad de datos son superiores, pero aporta más oportunidades para extraer business intelligence relevante de forma personalizada prácticamente en tiempo real (Talend, 2020).

Data Warehouse

El propósito de un Data warehouse o almacén de datos es proporcionar información para facilitar la toma de decisiones empresariales. El interés está en analizar el estado de la empresa en un momento t (esto puede incluir datos actuales, así como datos históricos) (Prakash & Prakash, 2018).

Un Data warehouse proporciona una visión multidimensional de los datos. Los datos se consideran en términos de hechos y dimensiones, un hecho es los datos básicos que se va a analizar, mientras que las dimensiones son los diversos parámetros a lo largo de los cuales se analizan los hechos (Prakash & Prakash, 2018).

Data Mart

Un data mart es un almacén de datos de pequeño tamaño centrado en un tema específico es decir una base departamental, mientras que un almacén de datos (data warehouse) es para toda la empresa, un centro de datos se construye para abordar las necesidades de análisis específicas de una unidad de negocios. Por lo tanto, se puede definir un data mart como un almacén de datos de pequeño tamaño que contienen un subconjunto del almacén de datos de la empresa o un volumen limitado de datos agregados para las necesidades de análisis específicas de una unidad de negocios, en lugar de las necesidades de toda la empresa. Por lo tanto, una empresa generalmente termina teniendo muchos data marts (Liu & Ozsu, 2018):

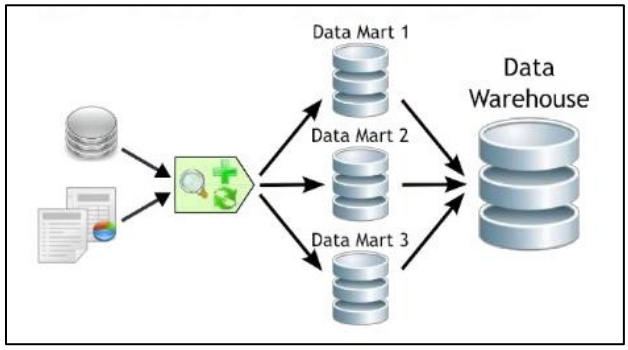
- Si bien el objetivo de un data warehouse es atender las necesidades de toda la empresa, el objetivo de un data mart es abordar las necesidades de una unidad de negocios como un departamento.
- Mientras que los datos de un data warehouse se alimentan de sistemas OLTP (procesamiento de transacciones en línea), el de un data mart se alimenta de la data warehouse de la empresa.

- Si bien la granularidad de una data warehouse está en el nivel de grano más bajo, la de un data mart está en el mismo nivel de grano más bajo o en un nivel ligeramente agregado para un análisis óptimo por parte de los usuarios de la unidad de negocios.
- Si bien la cobertura de un data warehouse es completamente histórica para atender las necesidades de toda la empresa, la de un data mart está limitada a las necesidades específicas de una unidad de negocios.

Un data mart puede ser alimentado desde un data warehouse o integrar por sí mismo un compendio de distintas fuentes de información (TutorialKart, 2018). Ver figura 5.

Figura 5.

Data Mart y Data Warehouse



Nota. Proceso de construcción de un Data Warehouse alimentado por diferentes Data Marts. (Troyanx - Soluciones Informáticas, 2019).

Para crear un Data mart de un área funcional de la empresa será necesario encontrar la estructura óptima para el análisis de su información, estructura que puede estar montada sobre una base de datos OLTP (On Line Transactional Processing), como la propia data warehouse, o sobre una base de datos OLAP (On Line Analytical Processing). La designación de una u otra dependerá de los datos, requisitos y características específicas de cada departamento.

Bases de Datos OLTP y OLAP

OLTP y OLAP son ambos los sistemas de procesamiento en línea. OLTP es un procesamiento transaccional, mientras que OLAP es un sistema de procesamiento analítico. A continuación de detalla cada uno de estos.

OLTP – On Line Transactional Processing

Los sistemas OLTP son bases de datos orientadas al procesamiento de transacciones. Una transacción genera un proceso atómico, y que puede involucrar operaciones de inserción, modificación y borrado de datos. El proceso transaccional es típico de las bases de datos operacionales.

- El acceso a los datos está optimizado por tareas frecuentes de lectura y escritura.
- Los datos se estructuran según el nivel de aplicación (programa de gestión a medida, ERP, CRM implantado, sistema de información departamental…).
- Los formatos de los datos no son necesariamente uniformes en los diferentes departamentos.
- El historial de los datos suele limitarse a los datos actuales o recientes (Sinnexus, 2020).

OLAP – On Line Analytical Processing

Los sistemas OLAP son bases de datos orientadas al procesamiento analítico. Este análisis suele implicar, generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil: tendencias de ventas, patrones de comportamiento de los consumidores, elaboración de informes complejos, etc. OLAP logra su máxima eficiencia y flexibilidad operando sobre bases de datos Multidimensionales. Este sistema es típico de los data marts.

- El acceso a los datos suele ser de sólo lectura. La acción más común es la consulta, con muy pocas inserciones, actualizaciones o eliminaciones.
- Los datos se estructuran según las áreas de negocio, y los formatos de los datos están integrados de manera uniforme en toda la organización.
- Las bases de datos OLAP suelen alimentar de información procedente de los sistemas operacionales existentes, mediante un proceso de extracción, transformación y carga (ETL) (Sinnexus, 2020).

Data warehouse vs Data mart

Tabla 1.

Diferencias entre data warehouse y data mart

	DATA WAREHOUSE	DATAMART
Alcance	Construido para satisfacer las necesidades de información de toda la organización.	Construido para satisfacer las necesidades de un área de negocio específica.
Objetivo	Diseñado para optimizar la integración y la administración de los datos fuente.	Diseñado para optimizar la entrega de la información de soporte de decisiones.
Características de los datos	Administra grandes cantidades de datos históricos a nivel atómico.	Se concentra en administrar resúmenes y/o datos totalizados.
Pertenencia	Pertenece a toda la organización.	Pertenece al área de negocio al cual está orientado.
Administración	Es administrado por la unidad de sistema de la organización.	Es administrado por el personal de sistema de la unidad propietaria del Datamart.

Nota. Diferencias entre arquitectura data warehouse y data mart *(Castillo & Palomino, 2012)*

Construcción de un almacén de datos

A la hora de construir un almacén de datos, los diseñadores deben tener una amplia perspectiva del uso que se espera del almacén.

- No existe un modo de anticipar todas las consultas o análisis posibles durante la fase de diseño.

- Sin embargo, el diseño debería soportar específicamente las consultas ad hoc.

Es necesario seleccionar un esquema adecuado que refleje el uso previsto.

Muchas de las cuestiones que rodean a los sistemas de apoyo para la toma de decisiones, se refieren en primer lugar a las tareas de obtener y preparar los datos.

Los datos deben ser extraídos de diversas fuentes, limpiados, transformados y consolidados en la base de datos de apoyo para la toma de decisiones. Posteriormente, debe ser actualizado periódicamente. Cada una de estas operaciones involucra sus propias consideraciones especiales (Areiza, Pérez, & Rivas, 2016).

En 1988, Barry Devlin y Paul Murphy de IBM publicaron un artículo titulado “Una arquitectura para un sistema de negocio y la información” que introdujo el concepto de un almacén de datos. Bill Inmon publicó “La construcción de almacén de datos” en 1992, que discute el diseño e implementación de un gran almacén de datos de la empresa, y en 1996, Ralph Kimball introdujo dimensiones de modelado de datos y data marts en su libro El kit de herramientas de almacenamiento de datos (Bhushan, Data Lake Integration Design Principles, 2016).

Por lo tanto, los almacenes de datos han existido desde hace más de 25 años y son una metodología establecida para la consolidación de datos, organización y procesamiento. Ha habido una serie de modificaciones y mejoras para el diseño y la implementación de almacenes. Además, hay un gran número de herramientas disponibles para la implementación de almacenes. Hay estructuras avanzadas tales como cubos que pueden ayudar a mejorar el rendimiento de la recuperación de los datos analizados y resumidos, mediante la realización de los cálculos y agregaciones necesarias con antelación.

El mundo que nos rodea está cambiando constantemente y también lo son las fuentes de datos. Alrededor de seis a ocho años atrás, las únicas fuentes de datos fueron la entrada del usuario para las aplicaciones (que utiliza una empresa) o de datos / registros

generados mediante programación. Todas estas fuentes generan datos estructurados que siguieron las reglas específicas, y la gestión de datos fue sencilla. Un almacén de datos era la única opción para consolidar, gestionar y analizar grandes cantidades de datos estructurados (Bhushan, Data Lake Integration Design Principles, 2016).

El amplio uso de los medios sociales, redes profesionales, y otras aplicaciones web generan cantidades masivas de datos semi estructurados y no estructurados que es muy beneficioso analizar. Además, los sensores para una gran variedad de máquinas generan enormes cantidades de datos que deben ser analizados, los almacenes de datos convencionales no son capaces de realizar esta tarea. Posteriormente, es necesario buscar nuevas opciones o plataformas que pueden ayudar en el procesamiento de estos datos (Bhushan, Data Lake Integration Design Principles, 2016).

Las organizaciones de la salud utilizan los sistemas de BI para mejorar la calidad de la atención y la satisfacción del paciente. Las herramientas de inteligencia de negocio ofrecen a los usuarios la capacidad de asociar elementos de datos para el análisis multidimensional de la información en la toma de decisiones estratégicas. Sin embargo, un sistema de BI es simplemente tan bueno como los datos que contiene y la experiencia de los analistas de usarlo. Para entender el beneficio completo de una herramienta de BI, un hospital debe primero mejorar la recolección, extracción, integración y el análisis de la información.

Metodologías: Kimball e Inmon

En la práctica de almacenamiento de datos, hay dos métodos comúnmente utilizados de diseño, es decir, método Inmon y método Kimball. Ambos son muy diferentes en términos de su patrón de diseño. Método Inmon utiliza el enfoque de arriba hacia abajo, lo que significa que el proceso de diseño se realiza mediante un análisis en profundidad sobre la base de las necesidades antes del almacén de datos construido; se puede decir

que el mercado de datos es una unidad trazada desde el almacén de datos. En este método el aspecto técnico es más prominente, ya que conduce al proceso de modelado de datos. El método Kimball tiene diferentes características, usa un enfoque de abajo hacia arriba y está orientado al proceso del negocio, el almacén de datos está construido en base a las necesidades de la organización, de modo que el mercado de datos se construye en primer lugar para que pueda ser utilizado para compilar el almacén de datos (Taufik, Prabasari, Rineksane, Yaya, & Widowati, 2017).

Teniendo en cuenta las diferentes características de cada metodología y el análisis de las necesidades del Área de Estadística de la Dirección Distrital, se optó por la metodología de Ralph Kimball ya que permite implementar almacenes de datos por departamentos o áreas, sin la necesidad de contar previamente con un data warehouse centralizado.

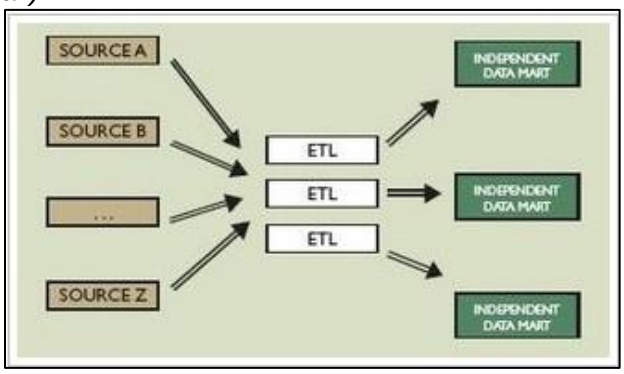
- **Método de Kimball**

  Ralph Kimbal sugiere un enfoque ascendente que utiliza el modelo dimensional para describir la organización de los datos en un Data Warehouse. En lugar de construir una bodega de datos para toda la empresa, Kimball sugiere que se implementen repositorios de datos más pequeños para los principales procesos del negocio.

  El modelo de Kimball deja de lado la necesidad de un data warehouse debido a que la mayoría de los usuarios desea obtener datos detallados, Kimball argumenta que es mejor almacenar los datos en data marts independientes y lógicamente conectados usando dimensiones. Para la optimización de consultas y mejorar la facilidad de uso de data marts, Kimball propone el modelo de datos como esquema estrella (Rosales, 2009).

Figura 6.

Botton-Up (Kimball)

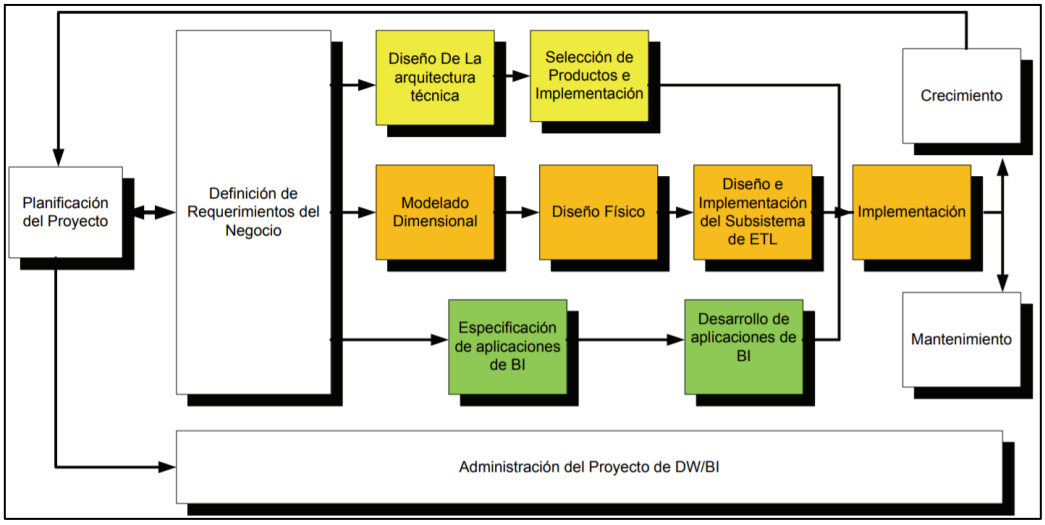


Nota. Botton-Up (Kimball) (Mundo BI, 2012)

Para la construcción de los data mart la metodología de Kimball realiza los pasos descritos en la siguiente figura:

Figura 7.

Proceso Kimball, Business Dimensional



Nota. Fases del Ciclo de Vida Dimensional del Negocio (Bustamante, Macas, & Beatriz, 2019)

Etapas de la metodología de Kimball:

1. **Planificación del proyecto:** busca identificar la definición y el alcance que tiene el proyecto de dwh.
2. **Definición de los requerimientos del negocio:** los diseñadores del dwh deben tener claro cuáles son los factores claves que guían el negocio para determinar efectivamente los requerimientos y traducirlos en consideraciones de diseño apropiadas.
3. **Diseño técnico de la arquitectura:** en esta fase se deben tener en cuenta tres factores: los requerimientos del negocio, los actuales entornos técnicos, y las directrices técnicas y estratégicas futuras planificadas por la compañía.
4. **Selección de productos e instalación:** se evalúa y selecciona cuales son los componentes necesarios específicos de la arquitectura (plataforma de hardware, motor de BD, herramientas ETL, etc).
5. **Modelado dimensional:** se comienza con una matriz donde se determine la dimensionalidad de cada indicador para luego especificar los diferentes grados de detalle dentro de cada concepto del negocio.
6. **Diseño Físico:** se centra en la selección de las estructuras necesarias para soportar el diseño lógico.
7. **Diseño e Implementación de subsistemas ETL:** proceso de extracción, transformación y carga ETL. Estas actividades son altamente críticas ya que tienen que ver con la materia prima del dwh que son los datos.
8. **Especificación de aplicaciones de BI:** en esta tarea se proporciona, a una gran comunidad de usuarios una forma más estructurada y por lo tanto, más fácil, de acceder al almacén de datos. Se proporciona este acceso estructurado a través de lo que se denomina, aplicaciones de inteligencia de negocios.

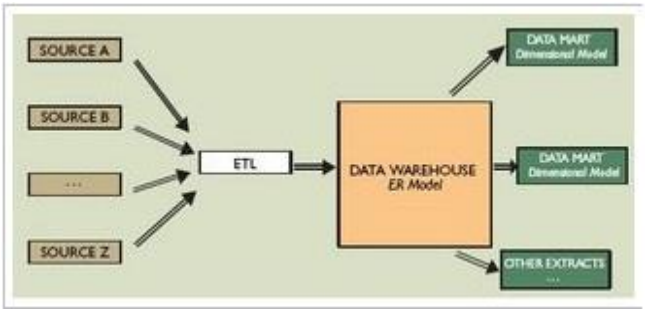
9. **Desarrollo de aplicaciones de BI:** Involucra configuraciones de los metadatos y construcción de reportes específicos.
10. **Implementación:** representa el correcto funcionamiento de la tecnología, los datos y las aplicaciones de usuarios finales accesibles para el usuario del negocio.
11. **Mantenimiento y Crecimiento:** se basa en la necesidad de continuar con las actualizaciones de forma constante para así lograr la evolución de las metas por conseguir.
12. **Gestión del Proyecto:** asegura que todas las actividades del ciclo de vida se lleven a cabo de manera sincronizada (Pacco Palomino, 2013).

- **Método de Inmon**

El padre del almacén de datos, Inmon utiliza un enfoque de arriba hacia abajo. Según su modelo, el diseño de un almacén de datos comienza con la estructura general. Primero, se configura todo el modelo de datos estandarizado, y después, los data marts.

Figura 8.

Top-Down (Inmon)



Nota. Metodología de trabajo Top-Down (Inmon), se centra primero en una visión global de la empresa, para ir dividiéndola en pequeños sets de datos departamentales. *(Mundo BI, 2012)*.

Bill Inmon ve la necesidad de transferir la información de los diferentes OLTP (Sistemas Transaccionales) de las organizaciones a un lugar centralizado donde los datos puedan ser utilizados para el análisis. Insiste además en que ha de tener las siguientes características:

- Orientado a temas: Los datos en la base de datos están organizados de manera que todos los elementos de datos relativos al mismo evento u objeto del mundo real queden unidos entre sí.
- Integrado: La base de datos contiene los datos de todos los sistemas operacionales de la organización, y dichos datos deben ser consistentes.
- No volátil: La información no se modifica ni se elimina, una vez almacenado un dato, éste se convierte en información de solo lectura, y se mantiene para futuras consultas.
- Variante en el tiempo: Los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones.

La información ha de estar a los máximos niveles de detalle. Los dwh departamentales o data marts son tratados como subconjuntos de este Dw corporativo, que son construidos para cubrir las necesidades individuales de análisis de cada departamento, y siempre a partir de este Dw Central (Espinosa, 2010).

Cuadro comparativo metodología Kimball vs metodología Inmon

Tabla 2.

Cuadro comparativo Kimball vs Inmon

CARACTERÍSTICAS	INMON	KIMBALL
Estructura de la arquitectura	Un Data warehouse que abarque toda la empresa y que se alimente de las bases de datos de los departamentos	Modelamiento de un Data mart por proceso de negocio.
Enfoque	Top-down	Bottom up
Complejidad	Alta	Baja
Accesibilidad de usuario final	Bajo	Alto
Alcance	Toda la empresa	Departamentos individuales
Tiempo de entrega	Requiere más tiempo de desarrollo	Tiempo de desarrollo inferior
Presupuesto	Coste inicial alto	Coste inicial bajo
Expertise	Equipo con especialización alta	Equipo con especialización media

Nota. La metodología aplicada para este proyecto está basado en el ciclo de vida de Kimball.

Análisis OLAP

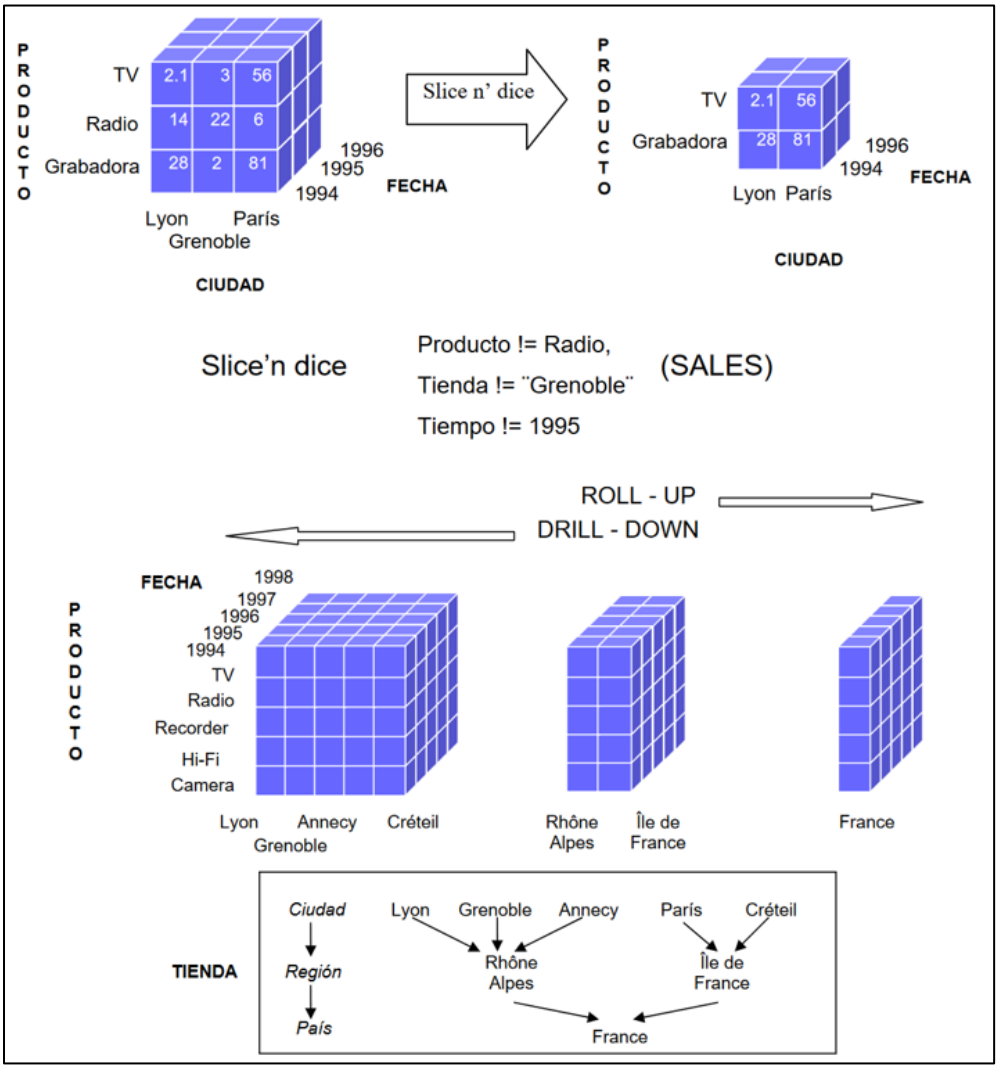
Al realizar un análisis del Data Mart, se toma en cuenta la manipulación o explotación de la información almacenada, la forma en que el usuario consultará el sistema y la parte de información a la que este tendrá acceso, de tal manera que dicho análisis dependerá directamente de las necesidades de los usuarios y del tipo de decisiones que se deseen tomar a través del uso del data mart construido. Este tipo de aplicaciones OLAP utiliza estructuras multidimensionales para proporcionar un acceso rápido a los datos con el fin de analizarlos. Los datos de origen de OLAP se almacenan habitualmente en almacenes de datos en una base de datos relacional.

Las técnicas OLAP son ampliamente utilizadas para este tipo de tareas, a través del uso de sus operadores se lleva a cabo la explotación de la información almacenada. A

continuación se detallan los operadores con los que se puede realizar un proceso de análisis: Ver figura 9.

Figura 9.

Operadores OLAP



- Slice: Extrae un sub-cubo de las celdas verificando restricciones a lo largo de una dimensión.
- Dice: Extrae un sub-cubo de las celdas verificando restricciones a lo largo de varias dimensiones.

- Roll-up: Abstrae detalles, navega entre las jerarquías disminuyendo el nivel de detalle.
- Drill-down: Aumenta el detalle de los datos, navega entre las jerarquías buscando detalles no visualizados, permite bajar a los niveles más atómicos de nuestro esquema multidimensional.
- Pivot: Permite diferencias las visualizaciones a través de cambios de columnas por filas o alterando ejes de las tablas.
- Rank: permite ordenar los datos de una dimensión de acuerdo con la medida corriente (Puerta, 2016).

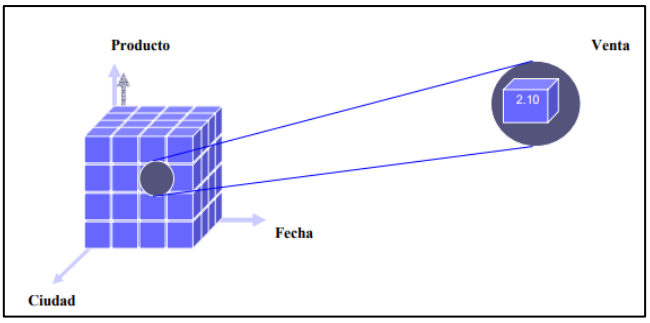
Modelado Multidimensional

Inmon propone un modelado a partir del modelo del negocio, pero Kimball propone usar un modelo multidimensional, basado en “hechos” y “dimensiones”.

Las bases de datos multidimensionales permiten tener el acceso al almacenamiento de datos de un data warehouse o data mart, donde se aprecia las dimensiones, medidas, indicadores y hechos del mismo (Domínguez Martínez, 2008). Ver figura 10.

Figura 10.

Modelo Multidimensional



En general, la estructura básica de un Data Warehouse o Data Mart para el Modelo Multidimensional está definida por dos elementos: esquemas y tablas.

Tablas del Modelo Multidimensional

Hay dos tipos básicos de tablas en el Modelo Multidimensional:

- **Tabla Fact o de Hechos**

  Es la tabla central en un esquema dimensional. Es en ella donde se almacenan las mediciones numéricas del negocio. Estas medidas se hacen sobre el grano, o unidad básica de la tabla.

  El grano o granularidad de la tabla queda determinada por el nivel de detalle que se almacenará en la tabla. Por ejemplo, para el caso de producto, mercado y tiempo, el grano puede ser la cantidad de madera vendida ´mensualmente´. El grano revierte las unidades atómicas en el esquema dimensional.

- **Tablas Look-up o Dimensionales**

  Estas tablas son las que se conectan a la tabla fact, son las que alimentan a la tabla fact. Una tabla dimensional almacena un conjunto de valores que están relacionados a una dimensión particular. Las tablas dimensionales no contienen hechos, en su lugar los valores en las tablas dimensionales son los elementos que determinan la estructura de las dimensiones. Así entonces, en ellas existe el detalle de los valores de la dimensión respectiva.

Para decidir si un campo de datos es un atributo o un hecho se analiza la variación de la medida a través del tiempo. Si varía continuamente implicaría tomarlo como un hecho, caso contrario será un atributo.

Los atributos dimensionales son un rol determinante en un data warehouse o data mart. Ellos son la fuente de todas las necesidades que debieran cubrirse. Esto significa que la base de datos será tan buena como lo sean los atributos dimensionales, mientras más descriptivos, manejables y de buena calidad, mejor será el data warehouse o data mart (Domínguez Martínez, 2008).

Esquemas del Modelo Multidimensional

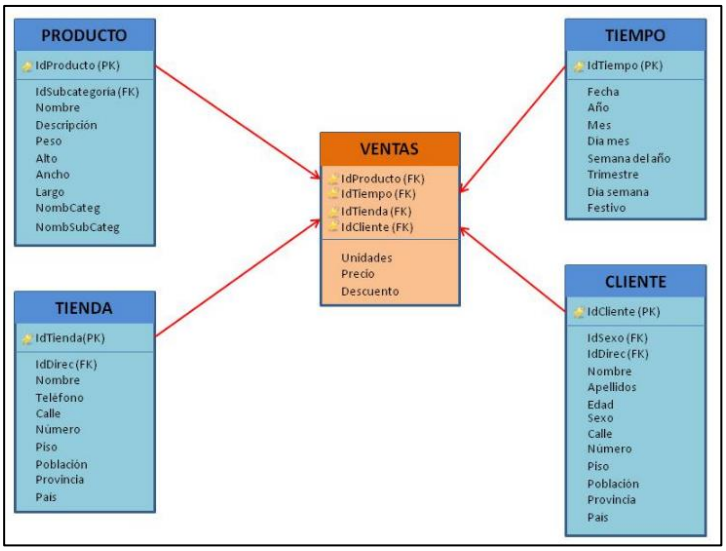
La colección de tablas en el Data Warehouse se conoce como Esquemas. Los esquemas caen dentro de dos categorías básicas: esquemas estrellas y esquemas snowflake o copo de nieve.

- **Esquema Estrella**

  Es el más sencillo en estructura. Consta de una tabla central de “Hechos” y varias “Dimensiones”, incluida una dimensión “Tiempo”. Lo característico de una arquitectura estrella es que sólo existe una tabla de dimensiones para cada dimensión. Esto quiere decir que la única tabla que tiene relación con otra es la de hechos, lo que significa que toda la información relacionada con una dimensión debe estar en una sola tabla (biverano, 2011).

  Figura 11.

  Esquema Estrella



Nota. Esquema Estrella *(mundodb.es, 2013)*

Se basa en una de hechos central que representa las medidas y que está enlazada a las tablas de dimensiones relacionadas que son las categorías

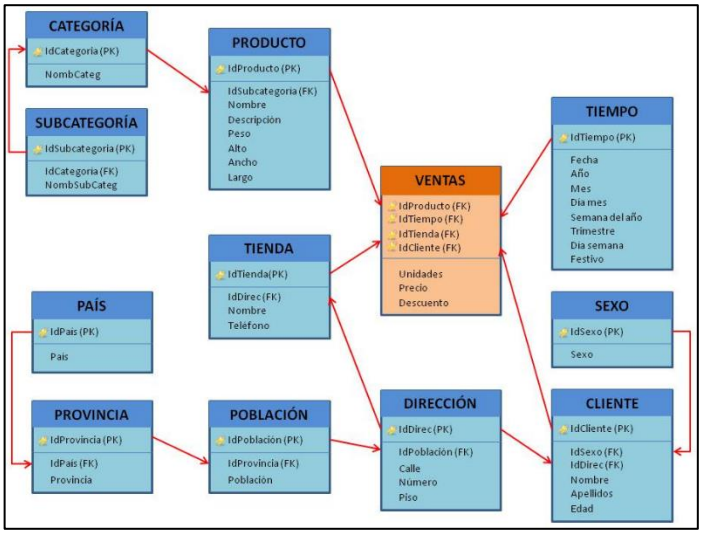
descriptivas de las medidas (Evaluando Software, 2016). Los queries multidimensionales son hechos haciendo joins entre las tablas de hechos y de dimensiones.

- **Esquema Copo de Nieve**

  Tiene el mismo concepto que el modelo estrella pero a su vez se enlaza a otras tablas dimensionales. El uso de estos esquemas o modelos simplifica la comprensión de los datos y maximiza el desempeño de las peticiones (querys) de la base de datos ahorrando espacio de almacenamiento (Evaluando Software, 2016). Los esquemas de copos de nieve contienen una tabla de hechos central sin normalizar para el tema y numerosas tablas de dimensión para la información descriptiva sobre las dimensiones del tema. El modelo fue concebido para facilitar el mantenimiento de las dimensiones, sin embargo esto hace que se vinculen más tablas, haciendo la extracción de datos más difícil.

Figura 12.

Esquema Copo de Nieve



Nota. Esquema Copo de Nieve *(mundodb.es, 2013)*

Sistemas Gestores de Bases de Datos

Un sistema gestor de base de datos (SGBD) pueden verse como una capa intermedia que integra el DML (Data Manipulation Language) y DLL (Data Definition Language) para facilitar la gestión de tuplas y consultas generalmente basadas en un estándar de un lenguaje para ejecución de consultas (SQL). Un SGBD debe facilitar las siguientes tareas:

- Definición y creación de las bases de datos.
- Manipulación de los datos realizando consultas, inserciones y actualizaciones.
- Acceso controlado a los datos mediante mecanismos de seguridad de acceso a los usuarios.
- Mantener la integración de los datos.
- Controlar la concurrencia a la base de datos.
- Mecanismos de copias de respaldo y recuperación para restablecer la información en caso de fallos de sistema.

Los gestores de bases de datos más utilizados son los siguientes:

PostgreSQL

Es un sistema de gestión de base de datos de libre distribución, publicado bajo la licencia BSD. Por ser un proyecto de código abierto, el desarrollo de PostgreSQL no es manejado por una empresa y/o persona, sino que es dirigido por una comunidad de desarrolladores que trabajan de forma desinteresada, altruista, libre y/o apoyada por organizaciones comerciales. La comunidad PostgreSQL se denominada el PGDG (PostgreSQL Global Development Group).

Sus principales características son: Alta concurrencia, ahorros considerables de costos de operación, estabilidad y confiabilidad (Iruela, 2018).

Oracle

Es la base de datos relacional que tiene una mayor fiabilidad y la más utilizada. Su desarrollo comenzó en 1977 y es propiedad de Oracle Corporation. Se construyó para poder acceder de forma directa a los objetos mediante el lenguaje de consulta SQL, es una arquitectura de tipo escalable y que se usa con mucha frecuencia en el campo empresarial. Tiene su propio componente de red, que hace posible que pueda existir una comunicación mediante las redes. Su ejecución se realiza en la mayoría de las plataformas, entre las cuales se puede citar a Windows, Linux, Unix, Mac OS, etc. Su arquitectura, se divide entre lógica y física. Esto hace que exista una flexibilidad mayor entre las redes de datos y una mayor robustez en la estructura de los mismos (CoRegistros, 2020).

MySQL

Es un sistema gestor de bases de datos relacional por excelencia. Es un SGDB multihilo y multiusuario utilizado en la gran parte de las páginas web actuales. Además es el más usado en aplicaciones creadas como software libre.

Se ofrece bajo GNU GPL aunque también es posible adquirir una licencia para empresas que quieran incorporarlo en productos privativos. Las principales ventajas de este Sistema Gestor de Bases de datos son:

- Facilidad de uso y gran rendimiento.
- Facilidad para instalar y configurar.
- Soporte multiplataforma
- Soporte SSL.

La principal desventaja es la escalabilidad, es decir, no trabaja de manera eficiente con bases de datos muy grandes que superan un determinado tamaño (INESEM, 2020).

Herramientas para procesos ETL

Las herramientas de ETL han existido durante más de 30 años y a lo largo de este tiempo han surgido diferentes tipos de herramientas a medida que la tecnología ha ido evolucionando. Estas pueden ser catalogadas en dos grandes categorías, las herramientas ´Enterprise´ y las ´Open Source´.

Hay varias compañías de software que se especializan exclusivamente en vender soluciones ETL, como Informática, IBM, Oracle y Microsoft mientras que, por otro lado, destacan también herramientas ETL de código abierto.

Las herramientas ETL son los instrumentos principales que permiten construir un data warehouse o data mart. Sin embargo, no siempre es sencillo saber cómo elegir la herramienta correcta y que mejor se adapte a nuestros objetivos.

ETL (extract, transform and load) es el proceso que permite extraer datos de fuentes heterogéneas y con distintos formatos en un único lugar; además, los datos se validan, se limpian y se aplican las transformaciones necesarias para que puedan ser analizados de forma sencilla; finalmente los datos se cargan en una base de datos, data warehouse o data mart, donde se encuentran listos para ser explotados, según nuestros objetivos de negocio.

Un proceso ETL puede llegar a ser muy complejo, también teniendo en cuenta el elevado tamaño de los datos para extraer, transformar y cargar. Por lo cual, las herramientas ETL juegan un papel fundamental ya que son la base para cualquier estrategia de análisis de datos y de inteligencia de negocio (MediaPro, 2018). Entre las herramientas ETL más utilizadas se tiene:

Informática PowerCenter

Ha sido la herramienta de integración de datos mejor valorada por la compañía Gartner. Prácticamente, tiene conectores para todo tipo de fuente de datos. La integración

de datos se realiza a modo punto a punto con un modelo distribuido. Se integra perfectamente con PowerCenter y proporciona datos operacionales disponibles de forma instantánea y escalable (IMF International Business School, 2019). Informática es menos maduro que otros productos para fuentes semiestructuradas y no estructuradas.

Las principales características de Informática PowerCenter son:

- Ofrece datos en tiempo real de manera precisa, permitiendo así los análisis oportunos y sin pérdida de tiempo.
- Dispone de visualizaciones de datos de última generación. Sus gráficas completas y amigables facilitan la gestión y la gobernanza de los metadatos.
- Provee de autoservicio en las áreas de negocio. Así, los tomadores de decisiones pueden acceder a información fiable y certera.

IBM InfoSphere DataStage

Herramienta desarrollada por IBM que a modo de workflow permite realizar todo el proceso ETL completo en múltiples sistemas, admite la administración extendida de metadatos y la conectividad empresarial con herramientas Big Data, así como herramientas en la nube. Es una herramienta que facilita la integración de datos en los modos de procesamiento por lotes (batch) o bien para escenarios SOA (IMF International Business School, 2019). Entre las principales características se tiene:

- Implementa reglas de validación de datos.
- Es útil para procesar y transformar grandes cantidades de datos.
- Puede manejar transformaciones complejas y administrar múltiples procesos de integración.
- Puede funcionar en batch, en tiempo real o como un servidor web (MediaPro, 2018).

Pentaho Data Integration

Solución tecnológica conocida como Kettel, es una de las herramientas ETL open source más potentes y versátiles a la hora de diseñar los procesos de integración a la medida de las necesidades de la empresa, fundamentalmente con el objetivo de construir y explotar su data warehouse. Pentaho Data Integration puede de manera muy simple tomar datos de una fuente de archivos locales y remotos, bases de datos, repositorios y aplicar un procesamiento a dichos datos como filtros, condiciones, cálculos, consultas y almacenar los resultados en un destino como archivos, bases de datos o repositorios. Entre sus principales características se encuentran:

- Open source.
- Entorno gráfico de desarrollo
- Migración de datos entre aplicaciones o bases de datos.
- Exportar datos desde bases de datos o archivos planos (también pudiendo volcar información desde archivos json a través de conectores NOSQL).
- Limpieza de datos (IMF International Business School, 2019).
- Multiplataforma
- Incluye cuatro herramientas: Spoon, para diseñar transformaciones ETL. Pan, para ejecutar transformaciones diseñadas con spoon. Chef, para crear trabajos. Kitchen, para ejecutar trabajos.
- La solución tiene una versión comunitaria de uso gratuito.

Cuadro comparativo entre Informática PowerCenter, IBM DataStage y Pentho DI

Tabla 3.

Cuadro comparativo herramientas ETL

	INFORMÁTICA POWERCENTER	IBM DATASTAGE	PENTAHO DATA INTEGRATION
Característica	Integración de datos, ETL	Integración de datos, ETL	Integración de datos, ETL
Licencia gratuita	Es mucho más caro, casi tres veces más caro que la mayoría de las otras soluciones.	El precio varía según el uso y no es tan costoso como algunas soluciones empresariales de la competencia.	La solución tiene una versión comunitaria de uso gratuito
Funciones	Amplios antecedentes y experiencia en el mercado ETL con su capacidad para escalar el rendimiento en el manejo de volúmenes de datos muy grandes en entornos complejos y heterogéneos.	Amplios antecedentes y experiencia en el mercado ETL con su capacidad para escalar el rendimiento en el manejo de volúmenes de datos muy grandes en entornos complejos y heterogéneos.	Software open source que cuenta con herramientas de Big Data y IoT.
Soporte	El soporte técnico es excelente. Es una de las razones por las que realmente gusta. Cuando se compara el soporte de IBM y el soporte de Informática, Informática es mucho mejor	No tiene buen soporte con respecto a Informática.	No tiene buen soporte con respecto a Informática.
Interfaz	El rendimiento y el diseño de Informática han sido muy valiosos.	La interfaz necesita mejoras. Es realmente demasiado técnica. Ese es el principal problema.	No es fácil de usar.
Documentación	La documentación de la herramienta para desarrolladores se puede mejorar con una explicación más clara de cada utilidad, acompañada de ejemplos relevantes, para que los desarrolladores puedan crear programas con facilidad	La documentación y la ayuda en la aplicación para esta solución deben mejorarse, especialmente para las nuevas funciones	No hay demasiada documentación fiable

Nota. La herramienta utilizada para el proceso ETL del proyecto es Pentaho Data Integration.

Herramientas de soporte a la toma de decisiones

La Inteligencia de Negocios es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios (Sinnexus, 2020).

Un Sistema de Soporte a la Decisión (DSS) es una herramienta Business Intelligence enfocada al análisis de los datos de una organización.

Las herramientas de inteligencia de negocios pueden estar categorizadas según el área de la industria en donde vayan a ser utilizadas y la cantidad de usuarios (individual, empresa pequeña u organización). A continuación, se describe algunas de las herramientas más utilizadas:

Power BI

Es una herramienta desarrollada por Microsoft, actualmente, lidera el cuadrante Gartner del sector, y es por ello que es la herramienta mejor considerada para el desarrollo de este tipo de proyectos.

Power BI es una solución empresarial que permite la visualización de datos y compartir información con toda la organización, o insertarla en su aplicación o sitio web. Proporciona una vista única de los datos más críticos de su negocio, y así, poder supervisar el estado de su empresa mediante un panel activo e informes interactivos. Se puede acceder a la información desde cualquier lugar.

Con Power BI se puede generar cuadros de mando potentes porque es una herramienta muy intuitiva y sencilla de manejar por usuarios que no están familiarizados con el análisis de datos, así que podrán sacar partido a su información desde el primer momento (Hiberus TI, 2016). A continuación, se describe las principales características de la herramienta:

- Arranque rápido: con Power BI podrás publicar y acceder a tus paneles e informes rápidamente y en pocos pasos.
- Fácil manejo: no será necesario que tengas conocimientos avanzados para crear tus propios cuadros de mandos e informes. Simplemente construye, edita, publica y comparte para colaborar con quien lo necesites.
- Centraliza información: unifica datos desde diferentes orígenes en un mismo informe y cuadro de mandos, sin tener que cambiar de aplicaciones.
- Capacidad de integración: Aprovecha la capacidad de integración de orígenes de datos tan diferentes como los que te puede proporcionar Salesforce, MailChimp, SAP BW, SAP HANA, MySQL, SQL Server, Teradata, Oracle, Google, Twitter, Facebook, entre otros.
- Utiliza tu propio lenguaje: Haz uso de la potente funcionalidad Q&A para obtener respuestas a la información que buscas en Power BI. Esta funcionalidad lo que va a permitir es realizar una pregunta en lenguaje natural y recibir una respuesta inmediata.

Tableau

Es una solución completa de Business Intelligence, actualmente, ubicado en segundo lugar según Gartner y que permite a las personas encargadas de tomar decisiones disponer de toda la información necesaria en tiempo real al alcance de sus manos.

Es una herramienta de visualización de datos interactiva, es decir, el usuario tiene la posibilidad de interactuar con los datos: comparar, filtrar, conectar unas variables con otras, etc. Además, los informes y dashboards que se pueden crear con la herramienta son muy visuales lo que facilita la comprensión rápida de los datos. A continuación, se describe las principales características de la herramienta:

- Permite la generación de análisis avanzados en tiempos récord, generando visualizaciones y demostraciones impactantes sin necesidad de realizar desarrollos complejos con una configuración muy flexible, ya que puede funcionar bajo un servidor, de forma local en el equipo de los usuarios o en la nube.
- Destaca por su facilidad para integrar datos de diferentes orígenes y su sencillez de uso, que permite realizar un análisis ágil y rápido en un entorno colaborativo.
- El acceso a los datos es instantáneo, independientemente de los orígenes de información que disponga la empresa. Como software líder en plataformas de Bussiness Intelligence, Tableau permite integrar toda la información de la empresa en un único modelo, de forma que se puede analizar al mismo tiempo datos de sistemas ERP como JDE o SAP, base de datos Oracle, SQL Server, hojas de cálculo, etc.

Qlik

A pesar de sus 15 años consecutivos como líder y ser uno de los primeros proveedores en ofrecer una experiencia de usuario final con capacidad de implementación ¨multinube¨ y mantener una clara visión orientada hacia Machine Learning (ML) y el crecimiento basado en la Inteligencia Artificial (IA), no ha crecido al mismo ritmo de sus competidores directos. Es una herramienta que permite la visualización y análisis de datos, brinda soluciones a las necesidades del negocio como apoyo para la toma de decisiones. A continuación, se describe las principales características de la herramienta:

- Auto-Servicio: Cualquier usuario crear sus propias visualizaciones de datos, sus cuadros de mando integral, al tiempo que ofrece a TI la confianza de estar diseñando unas librerías seguras y consistentes y unos datos bien gobernados.

- Multifuente: Qlik Sense se conecta con múltiples fuentes de datos, incluyendo entradas de datos en tiempo real, a fin de proporcionar unas vistas aún más exhaustivas, sin comprometer el rendimiento de las aplicaciones.

Cuadro comparativo entre Power BI, Tableau y Qlik

Tabla 4.

Cuadro comparativo herramientas de visualización de datos

	POWER BI	**TABLEAU**	**QLIK**
Ubicación Gartner 2020	Cuadrante de líderes (1er puesto)	Cuadrante de líderes (2do puesto)	Cuadrante de líderes (3er puesto)
Licencia gratuita	Posee Power BI Desktop que es una aplicación gratuita.	Posee Tableau Desktop, una aplicación gratuita con duración de 14 días	Posee Qlik sense desktop, una aplicación gratuita con duración de 30 días
Definición	Una solución de análisis empresarial, permite a las organizaciones convertir cada conjunto de datos en una oportunidad y tomar decisiones basadas en datos en varios aspectos del negocio. Ofrece a los usuarios una gran cantidad de herramientas de visualización de datos para ver los datos de nuevas formas, descubrir información y crear informes.	Una plataforma de análisis moderna, permite a las organizaciones aprovechar el poder de los datos y transformar la forma en que analizan y usan los datos. Con capacidades de autoservicio únicas que permiten a los usuarios crear, explorar y ver datos, permite poderosos análisis de arrastrar y soltar, lo que permite a los equipos tomar decisiones precisas y oportunas basadas en datos.	Una plataforma de análisis y gestión de datos de un extremo a otro, permite a las organizaciones aprovechar los descubrimientos de datos para respaldar la toma de decisiones comerciales críticas.
Implementación	La interfaz fácil de usar de Power BI es lo que realmente impulsa la visualización interactiva y rica.	Más fácil de usar; Los usuarios no técnicos implementan fácilmente la herramienta.	No es fácil de usar, necesita mayor expertise
Costo	Menos costoso	Comparativamente más caro	A un precio razonable
Perspectivas del futuro	Las actualizaciones y mejoras continuas del producto son espectaculares, y nunca se había visto una evolución así en un producto de estas características.	Camino ideal pasando de ser un producto desafiante a un líder por la vía directa.	Se nota el proceso de evolución técnica hacia el cloud que ha supuesto que no avance mucho en funcionalidad adicional en los últimos años

Nota. La herramienta utilizada para la visualización de datos del proyecto es Power BI.

Capítulo III

Construcción de la Solución y Visualización de los Resultados

El proyecto de investigación se desarrolla bajo la metodología de Ralph Kimball, la cual está compuesta por las siguientes fases: Planificación del proyecto, definición de requerimientos del negocio, modelado dimensional, diseño e implementación del subsistema de ETL, selección del producto, implementación y desarrollo del cubo, especificación de aplicación de BI y el diseño de la arquitectura técnica.

Planificación del proyecto

Definición del proyecto

Se determinó la necesidad de construir un datamart como solución de BI para el departamento de Estadística del Distrito de Salud 17d07 con el fin de evitar que la información que generan los distintos establecimientos de salud se encuentren dispersas e incompletas y de esta manera evitar que las decisiones se tomen en base a información que no está actualizada e integrada.

Objetivos y alcance del proyecto

Los objetivos y alcance del proyecto se detallan en el capítulo I del presente documento, donde además se define el planteamiento del problema y se propone la solución.

La solución consiste en construir un almacén de datos especializado para el área de Estadística del Distrito de Salud 17d07 utilizando herramientas de integración de datos, logrando así integrar todas las fuentes de datos heterogéneos en un único repositorio centralizado, que sirva de fuente de información completa y real, ayudando a un acertado análisis de información y una adecuada toma de decisiones, optimizando además el tiempo de elaboración de reportes.

Situación actual

Actualmente el área de Estadística no integra ni analiza la totalidad de los datos que se generan en las diferentes fuentes de sistemas de información, debido a que no poseen un almacén de datos centralizado que les permita integrar fuentes de datos heterogéneos. El proceso de análisis de información lo manejan con bases de datos independientes que poseen la misma estructura, excluyendo otras fuentes de información con datos heterogéneos.

Estrategia de implementación

La construcción del data mart como solución de BI se realizará utilizando las siguientes herramientas: Pentaho Data Integration (spoon) para los procesos de ETL y Power BI Desktop para la creación de cubos olap y dashboard, así como visualización y análisis de datos en múltiples dimensiones. Para alojar el Data Mart se utilizará como motor de base de datos MySQL. Todas estas herramientas mencionadas cumplen con las especificaciones técnicas para la realización del proyecto.

Selección de la metodología de desarrollo

Para el desarrollo de este proyecto se ha seleccionado la metodología de Kimball, debido a que se requiere implementar un data mart con un enfoque bottom up (de abajo hacia arriba). Este enfoque se aplica bien al proyecto propuesto porque se puede construir un almacén de datos especializado para un área específica. La metodología de Kimball permitirá construir una arquitectura que se adapte fácilmente al cambio ya que a medida que avanza el tiempo, se presentan nuevos requerimientos, por lo que es necesario dejar abierta la posibilidad de implementar nuevos data marts.

Tiempo e Inversión

El proyecto propuesto tiene una duración de 6 meses y se estima los siguientes costos de inversión en hardware y software:

Tabla 5.

Costo de hardware

Hardware	Costo
PC Dell Inspiron 15 Procesador: Intel Core i7 Diso Duro: 1tb RAM: 8gb	$860,00
Total	$860,00

Tabla 6.

Costo de software

Software	Costo
MySQL	$0,00
Pentaho Data Integration - spoon	$0,00
Power BI Desktop	$0,00
Total	$0,00

Tabla 7.

Recurso humano

Recurso Humano	Valor Mensual	Valor Total
Desarrollador	$1.200,00	$7.200,00
Total		$7.200,00

Tabla 8.

Otros gastos

Otros Gastos	Costo
Servicio de internet	$120,00
Energía eléctrica	$60,00
Transporte	$60,00
Total	$240,00

Roles

En la tabla 9 se detallan los roles de los involucrados en la construcción del proyecto:

Tabla 9.

Roles

Roles	Área	Responsabilidad
Responsable de área	Estadística y Análisis de la información	Brindar información sobre los requerimientos del área y entrega de información histórica
Analista	Estadística y Análisis de la información	Brindar información sobre los requerimientos del área
Responsable de área	Tecnologías de la Información y Comunicaciones (TICs)	Construir almacén de datos especializado para el área de Estadística

Definición de requerimientos del negocio

En la definición de los requerimientos del negocio se hace de vital importancia acudir a las personas que trabajan en el día a día con los datos.

Luego de varias reuniones (entrevistas) con el personal involucrado en el proyecto se pudo apreciar y entender los procesos que actualmente realiza el área de Estadística Distrital para la extracción, análisis y entrega de información. Existen 18 establecimientos de salud pertenecientes al distrito 17d07, cada uno de estos establecimientos generan datos diariamente mediante la atención de sus pacientes, los datos son registrados en diferentes tipos de archivos (xls, sql). El área de Estadística Distrital es la encargada de recolectar estos datos para el análisis y entrega de información a la Dirección para posterior toma de decisiones.

A continuación, se detallan los requerimientos funcionales y no funcionales para la construcción del almacén de datos departamental.

Requerimientos Funcionales

El desarrollo del presente proyecto se basó en las necesidades de información del área de Estadísticas del Distrito de Salud. Esta área está encargada de recopilar todos los datos que se generan en los diferentes establecimientos de salud, analizarlos y transformarlos en información relevante para la toma de decisiones por parte de la dirección distrital. A continuación, se detallan los requerimientos funcionales:

Tabla 10.

Requerimientos funcionales

Nro.	Requerimientos Funcionales
1	Crear un data mart para el área de Estadística en donde se almacene información relevante de las distintas fuentes de datos, teniendo una sola fuente de datos centralizada.
2	Extraer los datos de las siguientes fuentes de información. - Base de datos Sistema SGMAS Guama - Base de datos Sistema SGMAS Uni - Archivo .xls PRAS - Archivo .xls RDACAA
3	Unificar los datos de las fuentes de archivos .xls ((PRAS y RDACAA), con los siguientes campos: - hc_digital, establecimiento, atemed_lug_ate, atemed_fec_ini, atemed_con_diag, pcte_nom, pcte_tip_iden_pcte_ide, pcte_sexo, pcte_fec_nac, pcte_nacionalidad, pcte_autid_etn, pcte_seg, pcte_des_prov, pcte_des_cant, pcte_des_parr, pcte_grp_pri, prof_nombrecompleto, prof_esp_ate, concat_cie10_con_descripcion.
4	En la base de datos unificada (pras y rdacaa) agregar nuevo campo donde se identifique el origen de los datos, sea este pras o rdacaa.
5	En fuente de datos RDACAA realizar la concatenación de nombres de paciente y profesional.
6	En fuente de datos PRAS realizar la concatenación de atención médica cie 10 y descripción cie 10.
7	Estandarizar los valores de los campos PCTE_SEXO y PCTE_AUTID_ETN entre las fuentes de datos de PRAS y RDACAA.
8	En fuentes de datos SGMAS, PRAS y RDACAA asignar a los valores vacíos como n/d.
9	Estandarizar los valores de los campos SEXO y NACIONALIDAD de la fuente de datos SGMAS-Paciente
10	Concatenar nombres y apellidos del paciente de la fuentes de datos SGMAS-Paciente
11	Estandarizar los valores de los campos NOMBRE_GRUPCULTURAL de la fuente de datos SGMAS-Grupo Cultural
12	Identificar en nuevas tabla los tipos de sexo y auto identificación étnica de los pacientes.
13	En fuente de datos SGMAS-Personal_médico separar el campo nombre_medico en establecimiento_salud y nombre_medico.

Nro.	Requerimientos Funcionales
14	En fuente de datos SGMAS-Guama-personal_medico añadir constante que identifique que el establecimiento de salud es Guamaní.
15	Tener la capacidad de generar reportes por fecha, identificando días, meses, años, trimestre.
16	Unificar los turnos de las fuentes de datos de SGMAS-Guama y SGMAS-Uni.
17	Obtener datos referentes a provincia, cantón y parroquia de la fuente de datos de SGMAS para posteriores análisis, cuando el sistema se integre a nivel nacional.
18	Tener la capacidad de generación de reportes utilizando los datos proporcionados a través de cubos de información
19	Tener la capacidad de almacenamiento histórico para el análisis de la información.
20	Total de pacientes registrados año 2019 en las fuentes de datos PRAS, RDACAA y SGMAS, mostrando las fuentes de datos a los que pertenecen.
21	Cantidad de pacientes registrados por tipo de sexo y origen de datos.
22	Pacientes registrados por grupo étnico
23	Cantidad de turnos agendados por mes y trimestres año 2019
24	Cantidad de pacientes agendados por Establecimiento de Salud
25	Cantidad de turnos agendados por especialidad
26	Cantidad de pacientes que asisten a las agendas por día y establecimiento de salud.

Requerimientos No Funcionales

Tabla 11. Requerimientos no funcionales

Requerimientos no funcionales

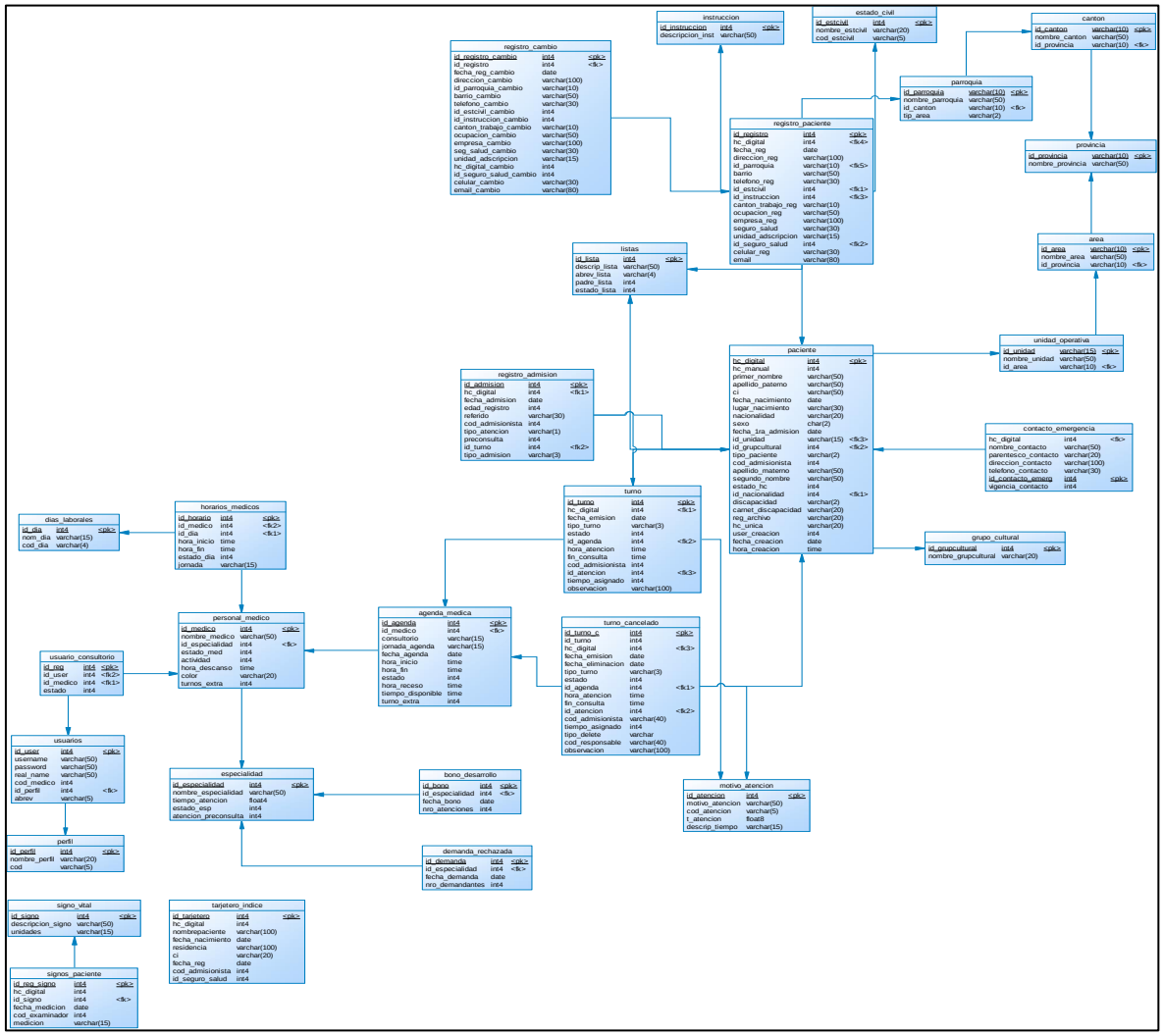
Nro.	Requerimientos No Funcionales
1	La implementación del Data Mart se realizará con herramientas free como son: Pentaho Data Integration (spoon), Base de Datos MySQL, Power BI Desktop.
2	El Data Mart se construirá sobre una base de datos MySQL.
3	La herramienta de explotación debe permitir la creación de reportes personalizados por el usuario final
5	La extracción de la data se realizará mensualmente.
6	La interfaz de BI debe ser fácil de usar con la finalidad de que los usuarios no especializados puedan hacer uso de los reportes.
7	Las funcionalidades del data mart y la carga de datos solo deben ser accesibles para los usuarios del área de Estadística y Tecnología.
9	Se considera que la generación de cubos olap y dashboards se realizarán con el software Power BI Desktop.

Modelo Físico de la base de datos transaccional sgmas en PostgreSQL

A continuación, se muestra el modelo físico generado de la base de datos del Sistema de Gestión Médico SGMAS, la misma que está compuesta por una tabla principal denominada ‘Turno’, que contiene toda la información y ficha de los agendamientos de los pacientes.

Figura 13.

Modelo Físico Base de Datos Transaccional SGMAS



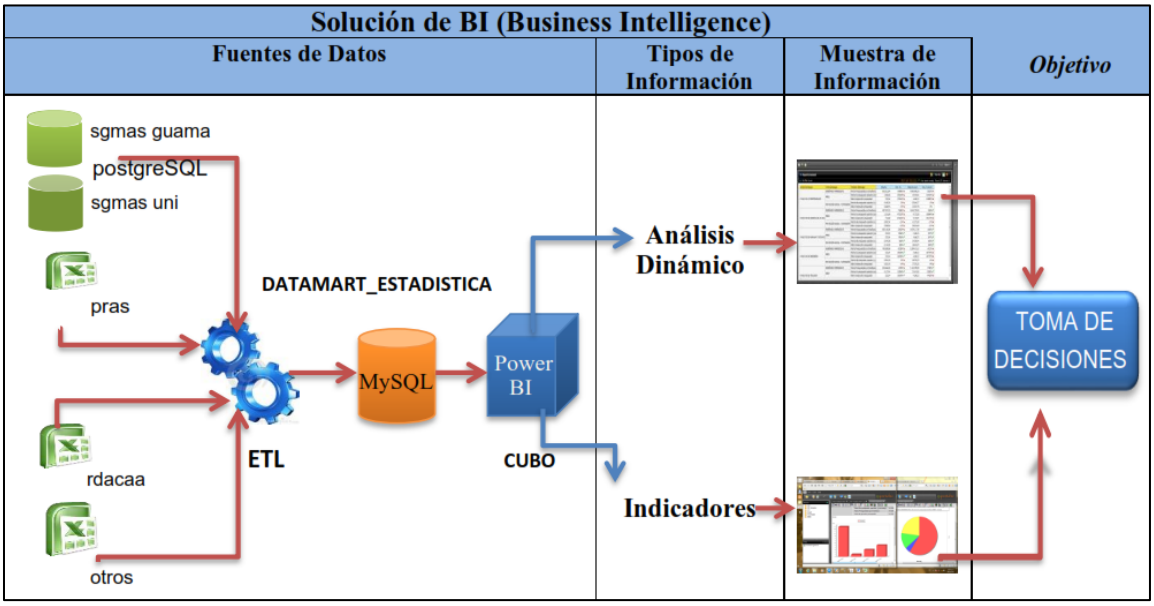
Nota. Modelo Físico del Sistema de Gestión Médico SGMAS cuya base de datos se encuentra en PostgreSQL.

Diseño técnico de la arquitectura

Para el diseño de la arquitectura técnica del Data Mart se debe tener en cuenta el esquema técnico donde se posicionará el Data Mart dentro de la empresa, los diferentes tipos de entradas de datos, como se puede ver en la figura 14.

Figura 14.

Arquitectura técnica Data Mart



Nota. Arquitectura técnica de la solución de BI utilizada en el proyecto.

Para el proceso de extracción de la data se cuenta con 2 bases de datos PostgreSQL y 2 fuentes de archivos tipo Excel. Mediante el proceso ETL estas fuentes son centralizadas en un único repositorio localizado en MySQL. Para la visualización de reportes de información se utiliza la herramienta Power BI.

Selección de productos e instalación

Para el desarrollo del proyecto de implementación del Data Mart se trabajó con las herramientas detalladas en la siguiente tabla:

Tabla 12.

Selección de productos e instalación

Producto	Característica	Uso	Hardware
Pentaho Data Integration 8.3 - spoon	Diseñador gráfico de transformaciones y trabajos del sistema de ETL´s	Diseño y elaboración del proceso de extracción, transformación y carga	Windows 10 12 GB RAM DD 1TB Core i7
PostgreSQL 9.5.22	Servidor de base de datos	Base de dato transaccional SGMAS GUAMA	
PostgreSQL 9.5.22	Servidor de base de datos	Base de dato transaccional SGMAS UNI	
MySQL 15.1	Servidor de base de datos	data mart	
MySQL 15.1	Servidor de base de datos	rdacaa y pras	
Power BI Desktop 2.87	Información multidimensional del negocio	Herramienta de visualización de datos y creación de cubos olap	

Nota. Detalle de las herramientas utilizadas para la construcción del almacén de datos.

Modelado Dimensional

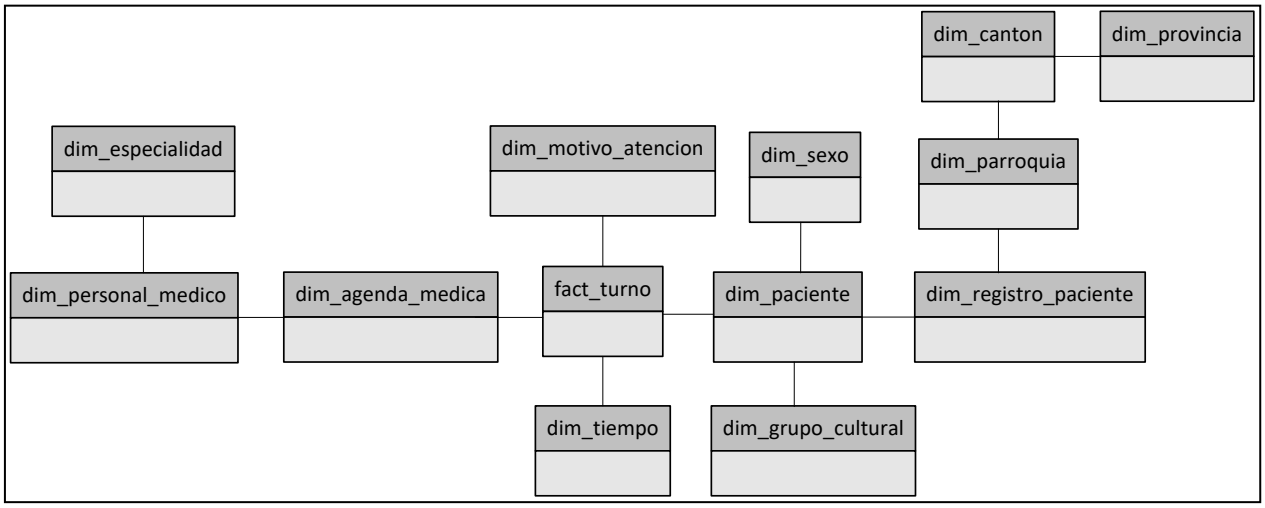
Se eligió el modelo estrella ya que su estructura es simple y hace que la extracción de datos sea más rápida, consiste en una tabla de hechos y una o varias dimensiones. Su diseño es sencillo de mantener y actualizar.

Diseño del Modelo Estrella

El diseño empleado para el modelo dimensional se visualiza en la siguiente figura:

Figura 15.

Modelo Estrella Data Mart



Nota. El modelo dimensional tipo estrella empleado para la construcción del almacén de datos consta de una tabla de hechos denominada ‘fact_turno’ que contiene los campos claves de cada dimensión.

Dimensiones

Las dimensiones que conforman el Data Mart se muestran en la siguiente tabla:

Tabla 13.

Dimensiones

Nro.	Dimensiones	Descripción
1	dim_agenda_medica	almacena los médicos que se encuentran agendados
2	dim_canton	almacena todos los cantones del Ecuador
3	dim_especialidad	almacena todas las especialidades con las que cuenta el Distrito de Salud
4	dim_grupo_cultural	almacena todas las auto identificaciones étnicas de los pacientes
5	dim_motivo_atencion	almacena todos los motivos por los que el paciente asiste a la atención médica

Nro.	Dimensiones	Descripción
6	dim_paciente	almacena todos los pacientes que fueron registrados en el proceso de admisión
7	dim_parroquia	almacena todas las parroquias del Ecuador
8	dim_personal_medico	almacena los datos de todo el personal médico que labora en el Distrito de Salud
9	dim_provincia	almacena todas las provincias del Ecuador
10	dim_registro_paciente	almacena información adicional del paciente
11	dim_sexo	almacena los tipos de sexo de los pacientes registrados
12	dim_tiempo	almacena la fecha desde el año 2018 divido por días, meses, años.

Atributos de las Dimensiones

En las siguientes tablas se describen los atributos de las dimensiones que conforman el Data Mart:

Tabla 14.

dim_agenda_medica

1. dim_agenda_medica		
Field	**Type**	**Descripción**
id_agenda	int(10)	clave primaria
id_dim_age_med	int(10)	id incremental
fecha_agenda	date	fecha de agenda del médico
turno_extra	int(10)	número de turnos extra del médico
nombre_medico	varchar(50)	nombre del médico
id_medico	int(10)	clave foránea de la dim personal medico

Tabla 15.

dim_canton

2. dim_canton		
Field	**Type**	**Descripción**
id_canton	varchar(100)	clave primaria
id_dim_canton	int(10)	id incremental
nombre_canton	varchar(50)	nombre del cantón
id_provincia	varchar(10)	clave foránea de la dim provincia

Tabla 16.

dim_especialidad

3. dim_especialidad		
Field	**Type**	**Descripción**
id_especialidad	int(10)	clave primaria
id_dim_especialidad	int(10)	id incremental
nombre_especialidad	varchar(50)	nombre de la especialidad
tiempo_atencion	float	tiempo de atención por especialidad

Tabla 17.

dim_grupo_cultural

4. dim_grupo_cultural		
Field	**Type**	**Descripción**
id_dim_gcultural	int(10)	clave primaria
PCTE_AUTID_ETN	varchar(100)	auto identificación étnica del paciente

Tabla 18.

dim_motivo_atencion

5. dim_motivo_atencion		
Field	**Type**	**Descripción**
id_atencion	int(10)	clave primaria
id_dim_mot_ate	int(10)	id incremental
motivo_atencion	varchar(50)	motivo de la atención del paciente
t_atencion	float	tiempo de atención por motivo

Tabla 19.

dim_paciente

6. dim_paciente		
Field	**Type**	**Descripción**
hc_digital	int(10)	clave primaria
id_dim_paciente	int(10)	id incremental
PCTE_IDE	varchar(30)	identificación del paciente

Field	Type	Descripción
PCTE_NOM	varchar(100)	nombre del paciente
PCTE_FEC_NAC	date	fecha de nacimiento del paciente
PCTE_NACIONALIDAD	varchar(50)	nacionalidad del paciente
origen_datos	varchar(50)	origen de datos (pras o rdacaa)
establecimiento_reg	varchar(50)	Establecimiento del registro del pac.
id_dim_sexo	int(10)	clave foránea de la dim sexo
id_dim_gcultural	int(10)	clave foránea de la dim gcultural

Tabla 20.

dim_parroquia

7. dim_parroquia		
Field	**Type**	**Descripción**
id_parroquia	varchar(10)	clave primaria
id_dim_parroquia	int(10)	id incremental
nombre_parroquia	varchar(50)	nombre de la parroquia
tip_area	varchar(2)	tipo de área urbana o rural
id_canton	varchar(100)	clave foránea de la dim canton

Tabla 21.

dim_provincia

9. dim_provincia		
Field	**Type**	**Descripción**
id_provincia	varchar(10)	clave primaria
id_dim_provincia	int(10)	id incremental
nombre_provincia	varchar(50)	nombre de la provincia

Tabla 22.

dim_personal_medico

8. dim_personal_medico		
Field	**Type**	**Descripción**
id_medico	int(10)	clave primaria
id_dim_personal_med	int(10)	id incremental

Field	Type	Descripción
nombre_medico	varchar(50)	nombre del médico
establecimiento_salud	varchar(100)	nombre del establecimiento de salud
turnos_extra	int(10)	turnos extras del personal médico
nombre_especialidad	varchar(50)	nombre de la especialidad
tiempo_atencion	float	tiempo de atención de la especialidad
id_especialidad	int(10)	clave foránea de la dim especialidad

Tabla 23.

dim_registro_paciente

10. dim_registro_paciente		
Field	**Type**	**Descripción**
id_registro	int(10)	clave primaria
id_dim_reg_pac	int(10)	id incremental
fecha_reg	date	fecha de registro del paciente
direccion_reg	varchar(100)	dirección de registro del paciente
id_parroquia	varchar(10)	clave foránea de la dim parroquia
hc_digital	int(10)	clave foránea de la dim paciente

Tabla 24.

dim_sexo

11. dim_sexo		
Field	**Type**	**Descripción**
id_dim_sexo	int(10)	clave primaria
PCTE_SEXO	varchar(30)	tipos de sexo del paciente

Tabla 25.

dim_tiempo

12. dim_tiempo		
Field	**Type**	**Descripción**
time_id	int(10)	clave primaria
the_date	date	la fecha (aa-mm-dd)

Field	Type	Descripción
the_day	varchar(30)	el día
the_month	varchar(30)	el mes
the_year	int(11)	el año
month_of_year	int(11)	mes del año
day_of_month	int(11)	día del mes
quarter	int(11)	trimestre
day_of_week	int(11)	día de la semana
day_of_year	int(11)	día del año

Hechos o fact

La tabla de hechos o fact que conforma el Data Mart se muestra en la siguiente tabla. Está conformada por las claves foráneas de las dimensiones y métricas que se desean analizar.

Tabla 26.

Tabla de hechos o fact

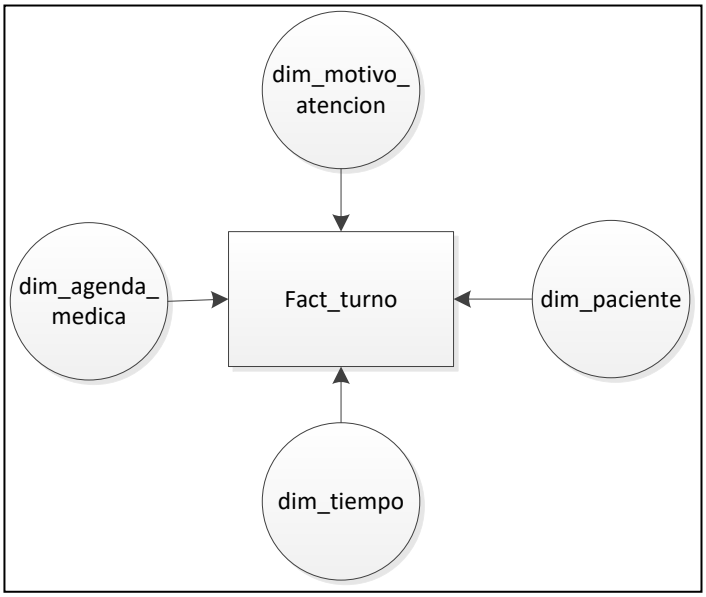
DESCRIBE fact_turno		
Field	**Type**	**Descripción**
id_fact_turno	int(10)	id incremental
FECHA	varchar(45)	fecha turno del paciente
hora_atencion	time	hora de atención
fin_consulta	time	hora fin de consulta
tiempo_asignado	int(4)	tiempo asignado de la consulta
tipo_turno	varchar(3)	tipo de turno
time_id	int(10)	clave foránea de dim tiempo
id_agenda	int(10)	clave foránea de dim agenda médica
hc_digital	int(10)	clave foránea de dim paciente
id_atencion	int(10)	clave foránea de dim motivo atención

Modelo gráfico de alto nivel

Para concluir con el proceso dimensional se realiza un gráfico denominado modelo dimensional de alto nivel (o gráfico de burbujas, Bubble chart), como lo menciona Kimball en el ciclo de vida de la metodología, enfocado en ‘fact turno’ que es la tabla de hecho de la cual se obtendrá las diferentes medidas en función de los turnos de los pacientes.

Figura 16.

Modelo bubble del indicador turno



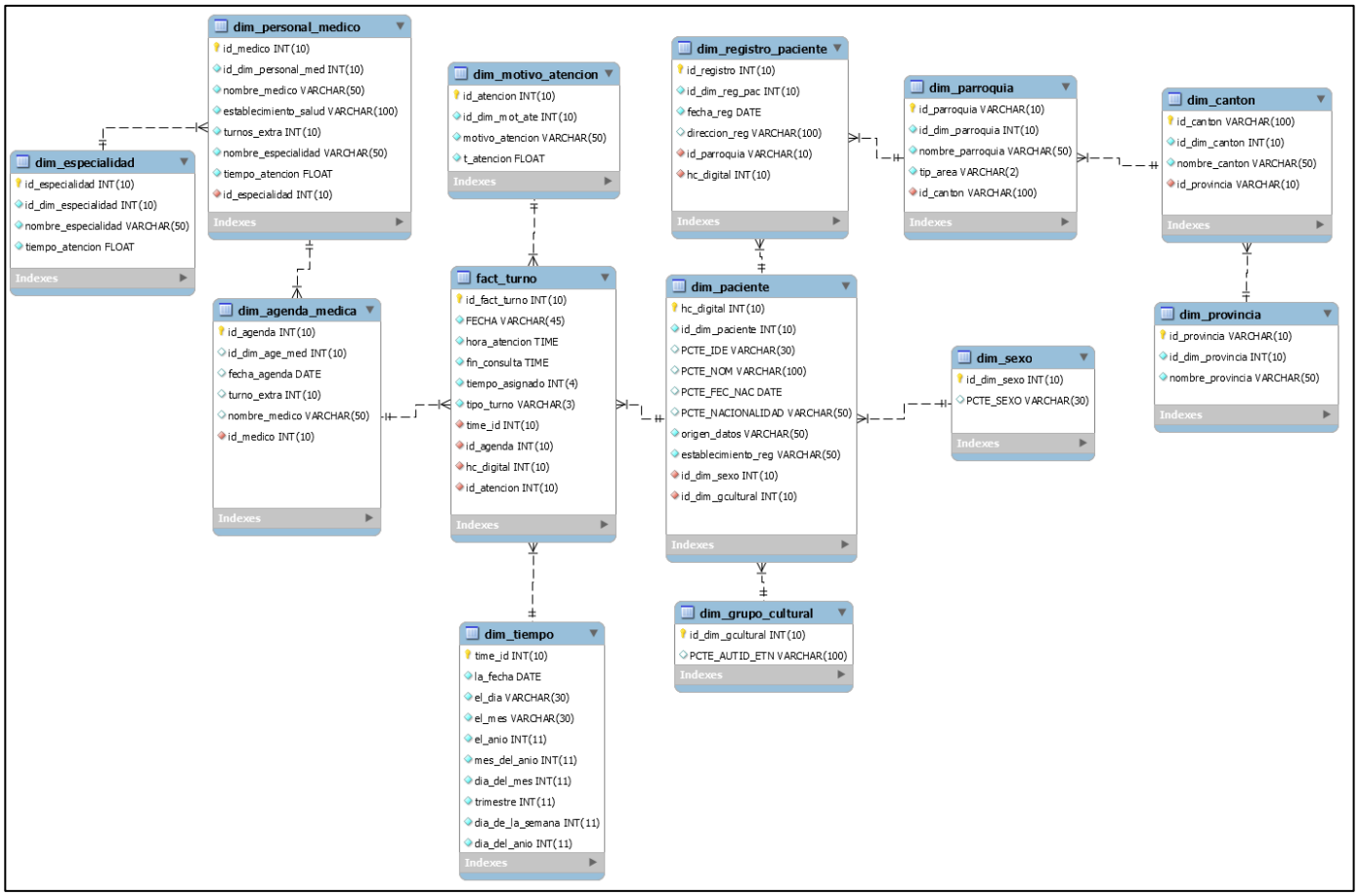
Nota. Modelo gráfico de alto nivel mediante la metodología de Kimball.

Diseño Físico

A continuación, se muestra el diseño físico estrella en base a los requerimientos y dimensiones detallas anteriormente.

Figura 17.

Diseño físico Data Mart



Nota. Diseño físico Data Mart con los atributos de cada tabla, conformada por una tabla de hechos.

El diseño físico del almacén de datos muestra cada una de las tablas con sus atributos y está conformada por la tabla de hechos denominada ‘fact_turno’ que contiene las claves primarias de las principales dimensiones.

Diseño e Implementación

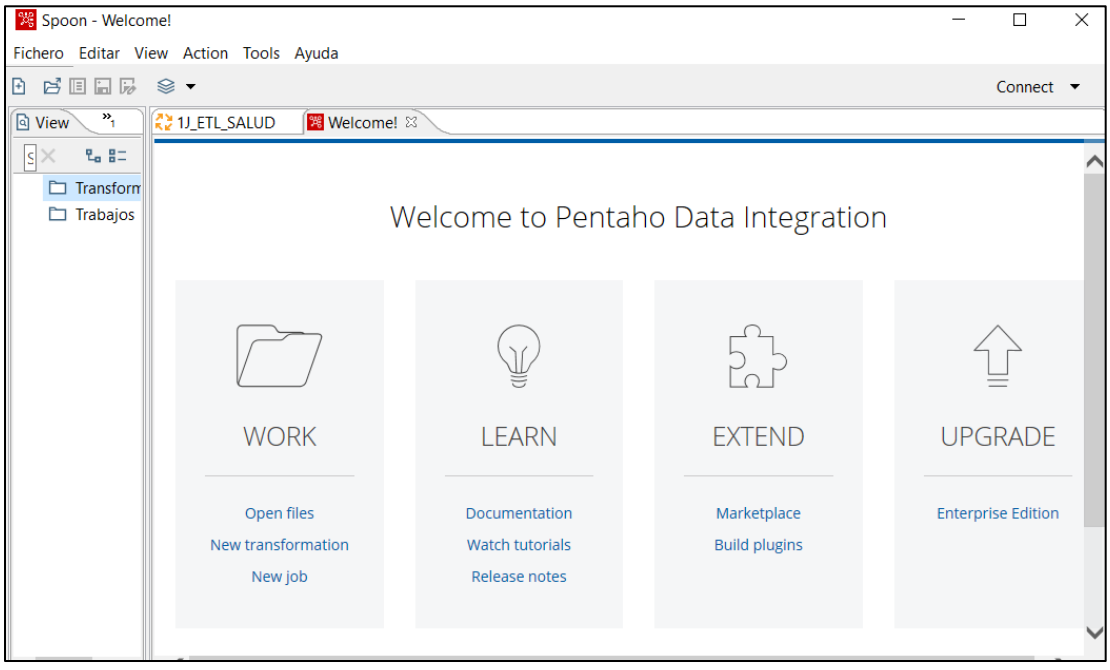
El proceso de extracción, transformación y carga (ETL) es conocida como la parte más extensa en el desarrollo de un proyecto de Inteligencia de Negocios.

Proceso de extracción, transformación y carga (ETL)

El proceso ETL del proyecto se realizó con la herramienta denominada **Spoon** que proviene de pentaho data integration. La herramienta spoon permite en primera instancia realizar una extracción de la información las diferentes fuentes de datos (sgmas guama, sgmas uni, pras, rdacaa), seguido de una transformación y limpieza de los datos y por último la carga de la nueva información en una base de datos destino de tipo Data Mart.

Figura 18.

Pantalla de entrada de spoon



Nota. Herramienta Spoon para la creación del proceso ETL.

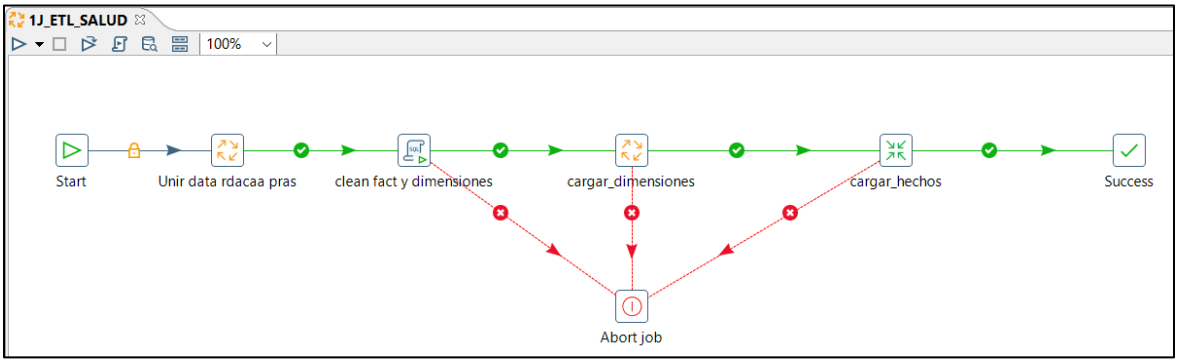
Para la creación del proceso ETL se realizaron los siguientes pasos:

Creación del trabajo ETL_SALUD

La creación de este job en spoon engloba todo el proceso automático del ETL, desde la extracción de los datos de las diferentes fuentes, transformación y limpieza, hasta la carga de los datos en el Data Mart en MySQL, como se puede ver en la figura 19:

Figura 19.

Job - ETL_SALUD



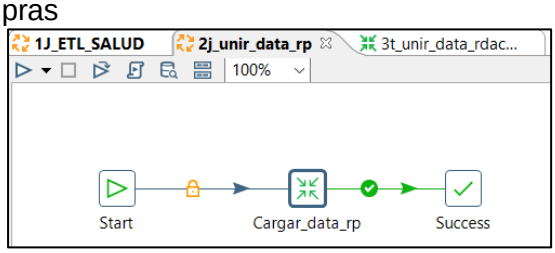
Nota. Proceso ETL global para la construcción del Data Mart.

Unir data rdacaa pras

En este trabajo se realiza la concatenación de los archivos pras.xls y rdacaa.xls en una sola base de datos cargada en MySQL denominada ‘rdacaa_pras’, la misma que tiene los campos definidos en los requerimientos funcionales. Ver figura 20.

Figura 20.

Job - unir data rdacaa pras

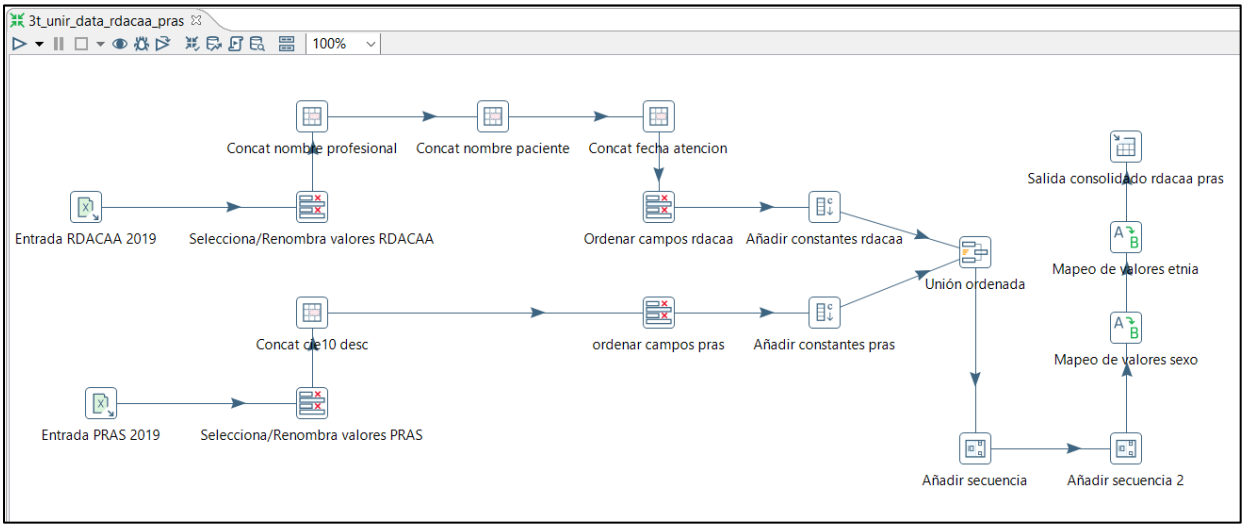


Nota. Job para unir data de sistemas rdacaa y pras

Este trabajo posee una transformación denominada ‘Cargar_data_rp’ la cual cumple la función descrita en el trabajo. En este proceso se realizan diferentes tareas como: concatenación de nombres, concatenación de fechas, ordenamiento de campos, agregación de constantes, mapeo de valores, entre otras tareas. Ver figura 21.

Figura 21.

Transformación - Cargar data rp



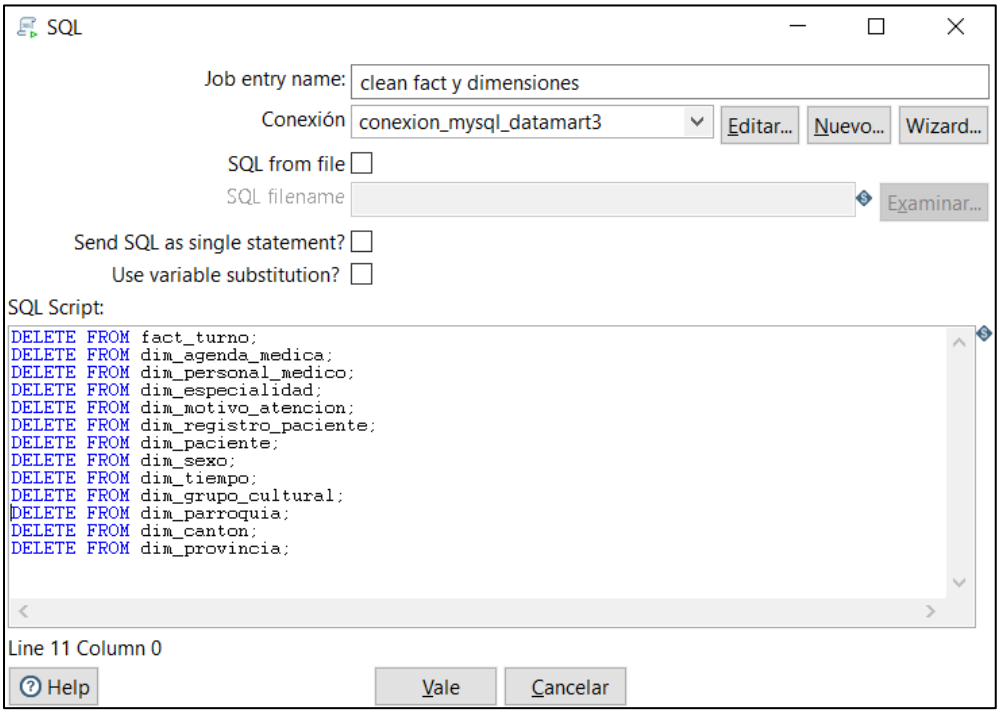
Nota. Transformación – Carga de datos unificados de sistemas rdacaa y pras en tabla de MySQL.

Clean fact y dimensiones

Limpieza de datos de las tablas de dimensiones y hechos que fueron cargadas en el Data Mart. Este paso se realiza cuando existe alguna actualización en las fuentes de datos. Ver figura 22.

Figura 22.

SQL - Clean fact y dimensiones



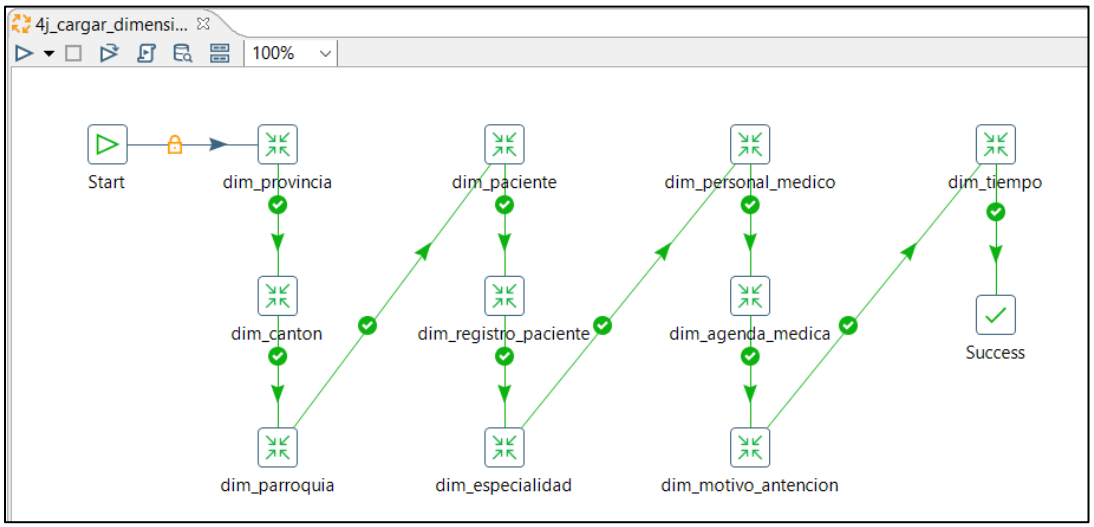
Nota. Proceso de limpieza de datos SQL - Clean fact y dimensiones

Cargar dimensiones

En este trabajo se realiza la carga de cada una de las dimensiones del Data Mart. Con el componente ‘Start’ se inicia el trabajo hasta llegar al componente ‘Success’ que indica que el proceso concluyo exitosamente. Ver figura 23.

Figura 23.

Job – Cargar dimensiones



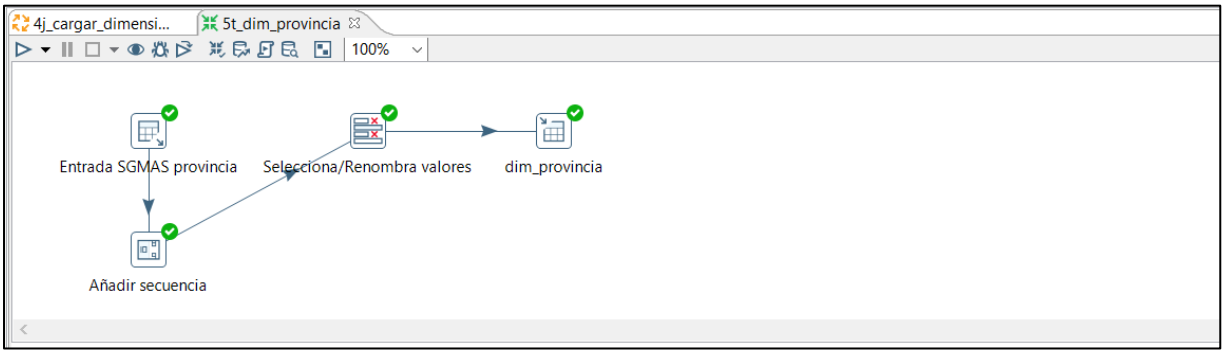
Nota. Job – Proceso de carga de dimensiones

Cada una de estas dimensiones es una transformación. A continuación, se muestran todas las transformaciones que forman el trabajo ‘Cargar dimensiones’.

- **dim_provincia:** utiliza como entrada la base de datos SGMAS con la tabla ‘provincia’ y mediante el proceso ETL se carga las provincias del Ecuador en la tabla dim_provincia del Data Mart. Ver figura 24.

Figura 24.

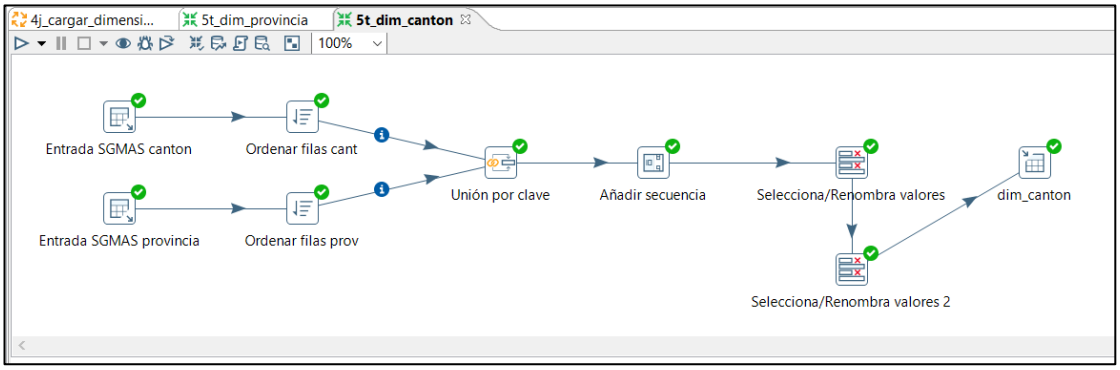
Transformación – dim provincia



- **dim_canton:** utiliza como entrada la base de datos SGMAS con la tabla ‘cantón’ y mediante el proceso ETL se carga los cantones de cada provincia del Ecuador en la tabla dim_canton del Data Mart. Ver figura 25.

Figura 25.

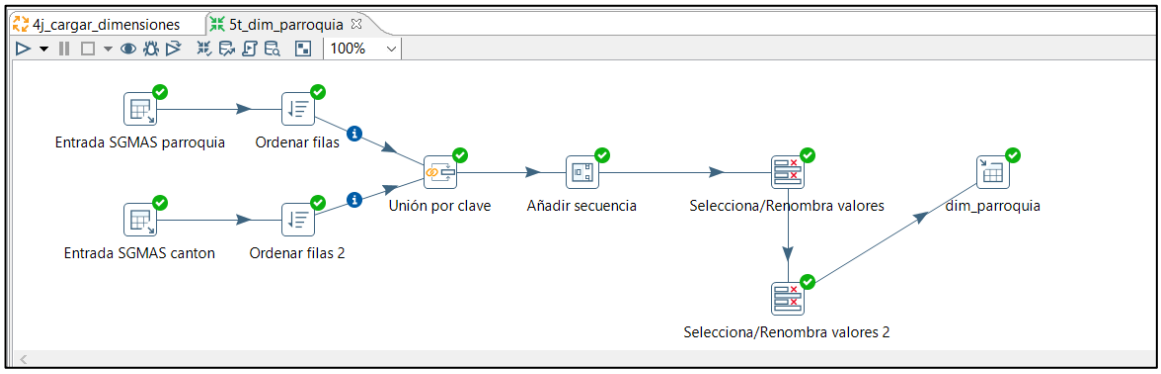
Transformación – dim cantón



- **dim_parroquia:** utiliza como entrada la base de datos SGMAS con la tabla ‘parroquia’ y mediante el proceso ETL se carga las parroquias de cada cantón del Ecuador en la tabla dim_parroquia del Data Mart. Ver figura 26.

Figura 26.

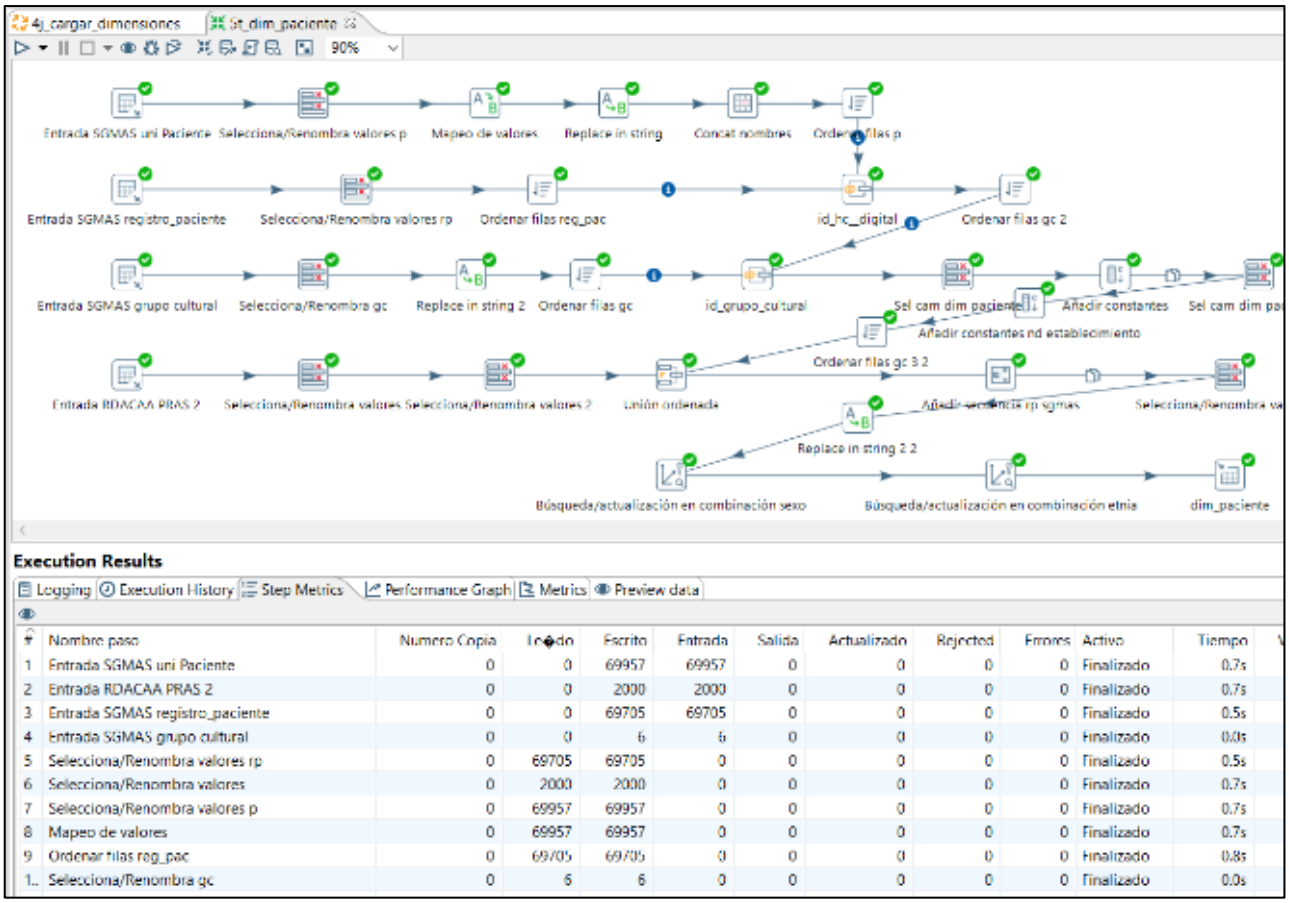
Transformación – dim parroquia



- **dim_paciente:** utiliza como entrada la base de datos ‘SGMAS uni’ con las tablas ‘paciente’, ‘registro_paciente’, ‘grupo_cultural’ y la base de datos unificada ‘RDACAA PRAS’ y mediante el proceso ETL se cargan todos los pacientes registrados en ambas bases en la dim_paciente del Data Mart. Mediante este proceso ETL también se obtiene en tablas diferentes del Data Mart la identificación del grupo cultural y el tipo de sexo de los pacientes. Ver figura 27.

Figura 27.

Transformación – dim paciente



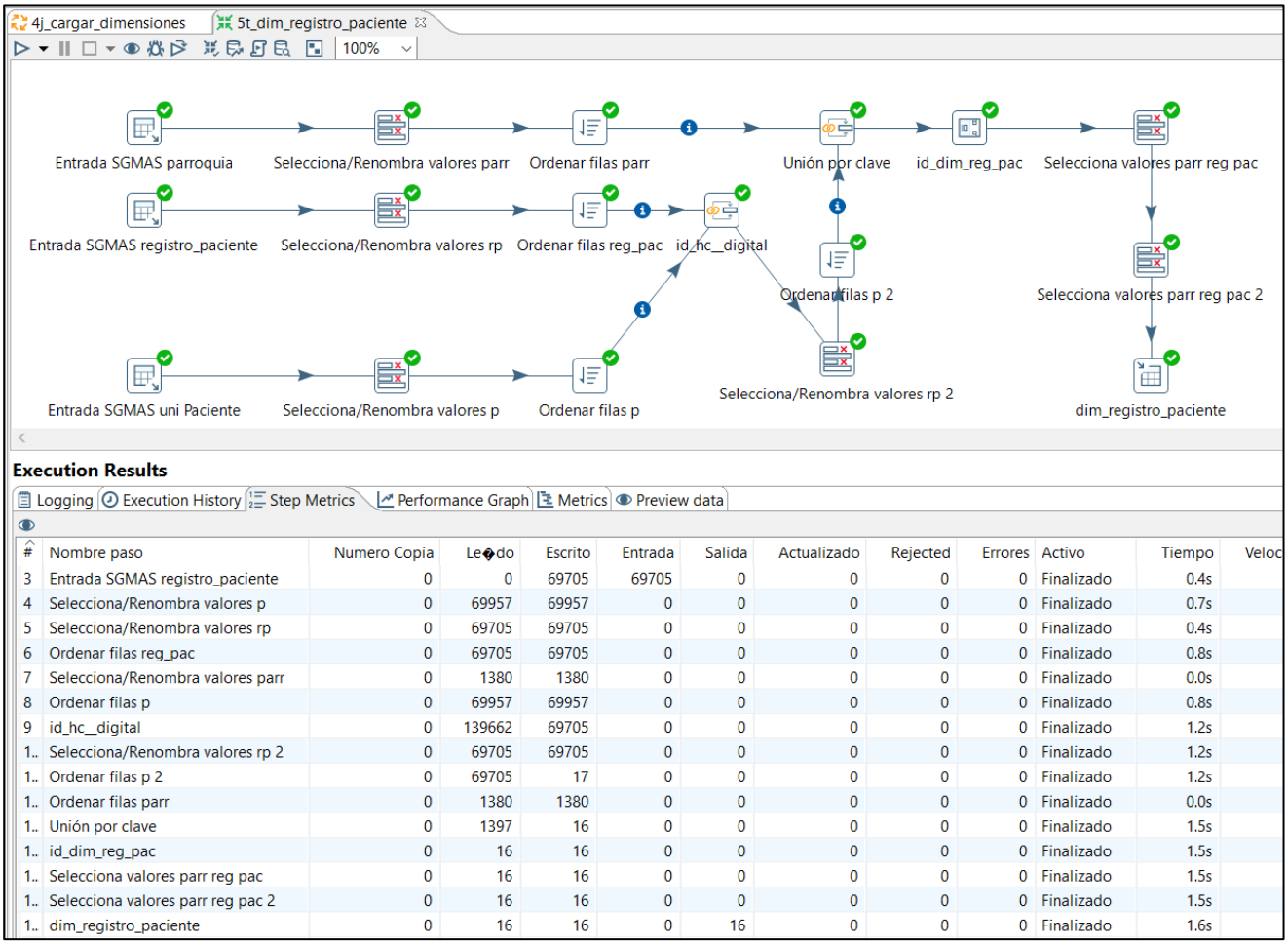
#	Nombre paso	Numero Copia	Leído	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo
1	Entrada SGMAS uni Paciente	0	0	69957	69957	0	0	0	0	Finalizado	0.7s
2	Entrada RDACAA PRAS 2	0	0	2000	2000	0	0	0	0	Finalizado	0.7s
3	Entrada SGMAS registro_paciente	0	0	69705	69705	0	0	0	0	Finalizado	0.5s
4	Entrada SGMAS grupo cultural	0	0	6	6	0	0	0	0	Finalizado	0.0s
5	Selecciona/Renombra valores rp	0	69705	69705	0	0	0	0	0	Finalizado	0.5s
6	Selecciona/Renombra valores	0	2000	2000	0	0	0	0	0	Finalizado	0.7s
7	Selecciona/Renombra valores p	0	69957	69957	0	0	0	0	0	Finalizado	0.7s
8	Mapeo de valores	0	69957	69957	0	0	0	0	0	Finalizado	0.7s
9	Ordenar filas reg_pac	0	69705	69705	0	0	0	0	0	Finalizado	0.8s
1..	Selecciona/Renombra gc	0	6	6	0	0	0	0	0	Finalizado	0.0s

Nota. Transformación – dim paciente. En este proceso se unifican la data de los sistemas rdacca y pras con la data del Sistema SGMAS.

- **dim_registro_paciente:** utiliza como entrada la base de datos SGMAS con las tablas ‘parroquia’, ‘registro_paciente’, ‘paciente’ y mediante el proceso ETL se cargan los pacientes registrados únicamente de la base de datos ‘SGMAS’ en la tabla ‘dim_registro_paciente’ del Data Mart. Ver figura 28.

Figura 28.

Transformación – dim registro paciente



Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

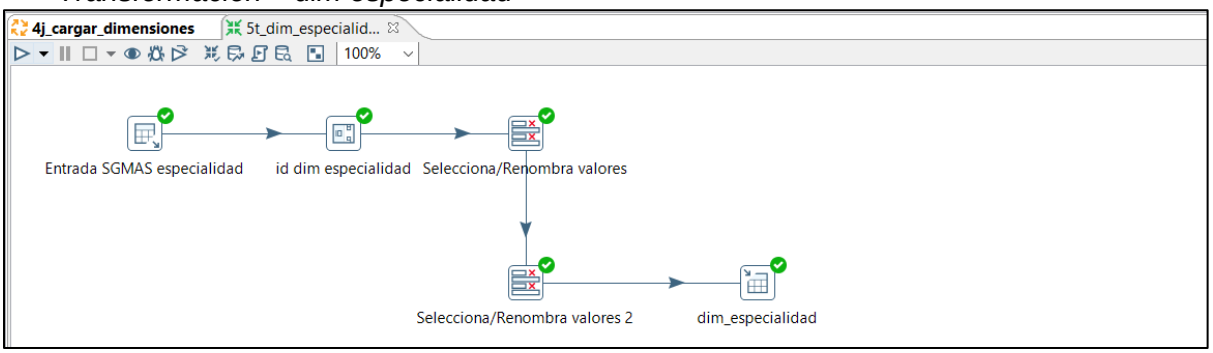
#	Nombre paso	Numero Copia	Le◆do	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Veloc
3	Entrada SGMAS registro_paciente	0	0	69705	69705	0	0	0	0	Finalizado	0.4s	
4	Selecciona/Renombra valores p	0	69957	69957	0	0	0	0	0	Finalizado	0.7s	
5	Selecciona/Renombra valores rp	0	69705	69705	0	0	0	0	0	Finalizado	0.4s	
6	Ordenar filas reg_pac	0	69705	69705	0	0	0	0	0	Finalizado	0.8s	
7	Selecciona/Renombra valores parr	0	1380	1380	0	0	0	0	0	Finalizado	0.0s	
8	Ordenar filas p	0	69957	69957	0	0	0	0	0	Finalizado	0.8s	
9	id_hc__digital	0	139662	69705	0	0	0	0	0	Finalizado	1.2s	
1..	Selecciona/Renombra valores rp 2	0	69705	69705	0	0	0	0	0	Finalizado	1.2s	
1..	Ordenar filas p 2	0	69705	17	0	0	0	0	0	Finalizado	1.2s	
1..	Ordenar filas parr	0	1380	1380	0	0	0	0	0	Finalizado	0.0s	
1..	Unión por clave	0	1397	16	0	0	0	0	0	Finalizado	1.5s	
1..	id_dim_reg_pac	0	16	16	0	0	0	0	0	Finalizado	1.5s	
1..	Selecciona valores parr reg pac	0	16	16	0	0	0	0	0	Finalizado	1.5s	
1..	Selecciona valores parr reg pac 2	0	16	16	0	0	0	0	0	Finalizado	1.5s	
1..	dim_registro_paciente	0	16	16	0	16	0	0	0	Finalizado	1.6s	

Nota. Proceso de transformación para la carga de la dimensión dim registro paciente.

- **dim_especialidad:** utiliza como entrada la base de datos SGMAS con la tabla ‘especialidad’ y mediante el proceso ETL se carga las especialidades con las que cuenta el Distrito de Salud en la tabla dim_especialidad del Data Mart. Ver figura 29.

Figura 29.

Transformación – dim especialidad

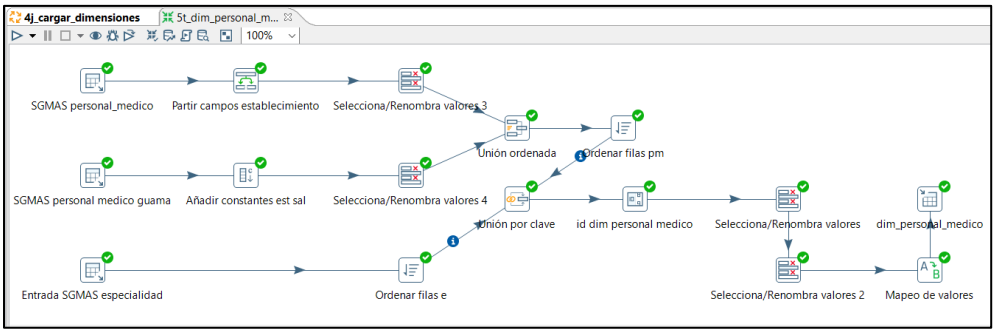


Nota. Proceso de transformación para la carga de la dimensión dim especialidad.

- **dim_personal_medico:** utiliza como entrada la base de datos SGMAS con las tablas ‘personal_medico’, ‘personal_medico_guama’, ‘especialidad’ y mediante el proceso ETL se cargan los datos del personal médico identificando a que establecimiento de salud pertenecen. La carga se hace en la tabla ‘dim_personal_medico’ del Data Mart.

Figura 30.

Transformación – dim personal médico

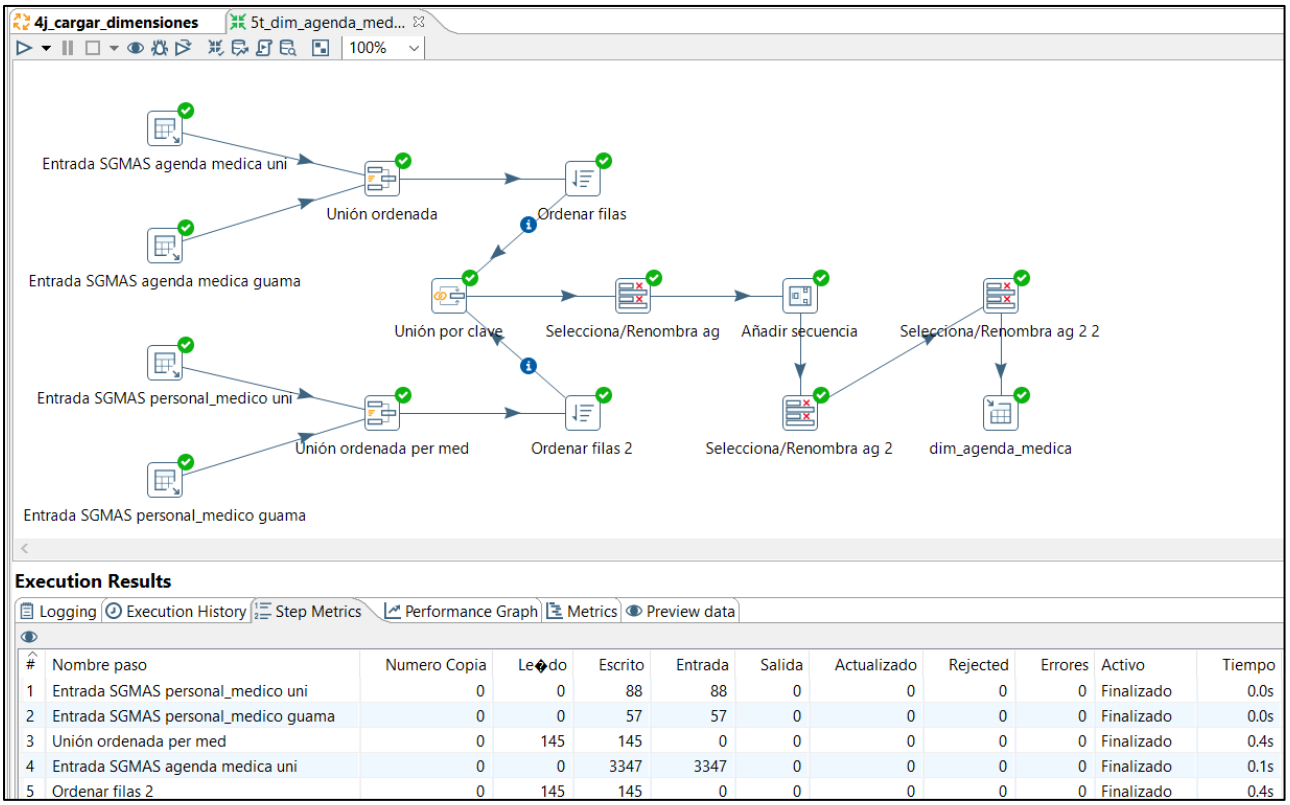


Nota. Proceso de transformación para la carga de la dimensión dim personal médico.

- **dim_agenda_medica:** utiliza como entrada las bases de datos SGMAS uni y SGMAS guama con las tablas ‘agenda medica uni’, ‘agenda medica guama’, ‘personal médico uni’, ‘personal médico guama’ y mediante el proceso ETL se cargan los datos unificados de agenda médica y personal médico de ambas bases de datos. La carga de datos se hace en la tabla ‘dim_agenda_medica’ del Data Mart. Ver figura 31.

Figura 31.

Transformación – dim agenda médica



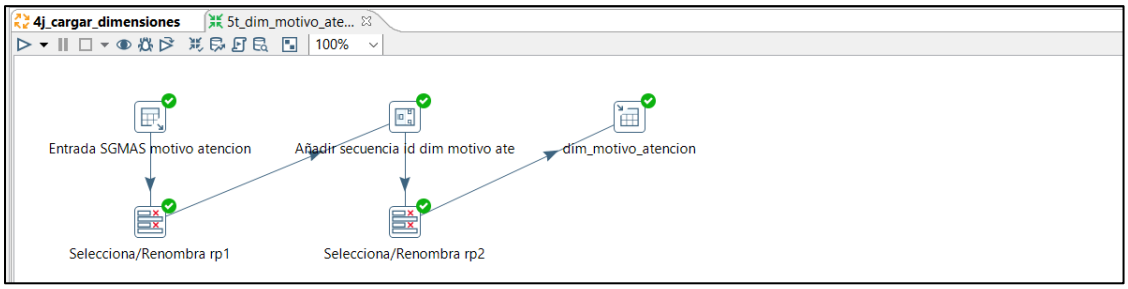
#	Nombre paso	Numero Copia	Le◆do	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo
1	Entrada SGMAS personal_medico uni	0	0	88	88	0	0	0	0	Finalizado	0.0s
2	Entrada SGMAS personal_medico guama	0	0	57	57	0	0	0	0	Finalizado	0.0s
3	Unión ordenada per med	0	145	145	0	0	0	0	0	Finalizado	0.4s
4	Entrada SGMAS agenda medica uni	0	0	3347	3347	0	0	0	0	Finalizado	0.1s
5	Ordenar filas 2	0	145	145	0	0	0	0	0	Finalizado	0.4s

Nota. Proceso de transformación para la carga de la dimensión dim agenda médica.

- **dim_motivo_atencion:** utiliza como entrada la base de datos SGMAS con la tabla ‘motivo atención’ y mediante el proceso ETL se cargan los diferentes tipos de atenciones en la tabla dim_motivo_atencion del Data Mart. Ver figura 32.

Figura 32.

Transformación – motivo atención

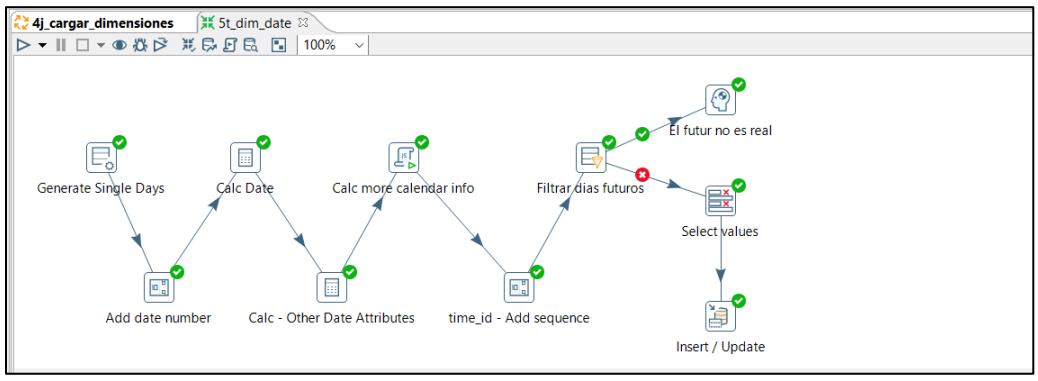


Nota. Proceso de transformación para la carga de la dimensión dim motivo atención.

- **dim_tiempo:** generador de fechas desde una fecha inicio hasta una fecha final. Aquí se extrae información como: la fecha, el día, el mes, el año, mes del año, día del mes, trimestre, día de la semana, y día del año, que servirán para obtener reportes más detallados por fecha. La carga de estos datos se realiza en la tabla dim_tiempo del Data Mart. Ver figura 33.

Figura 33.

Transformación – dim tiempo



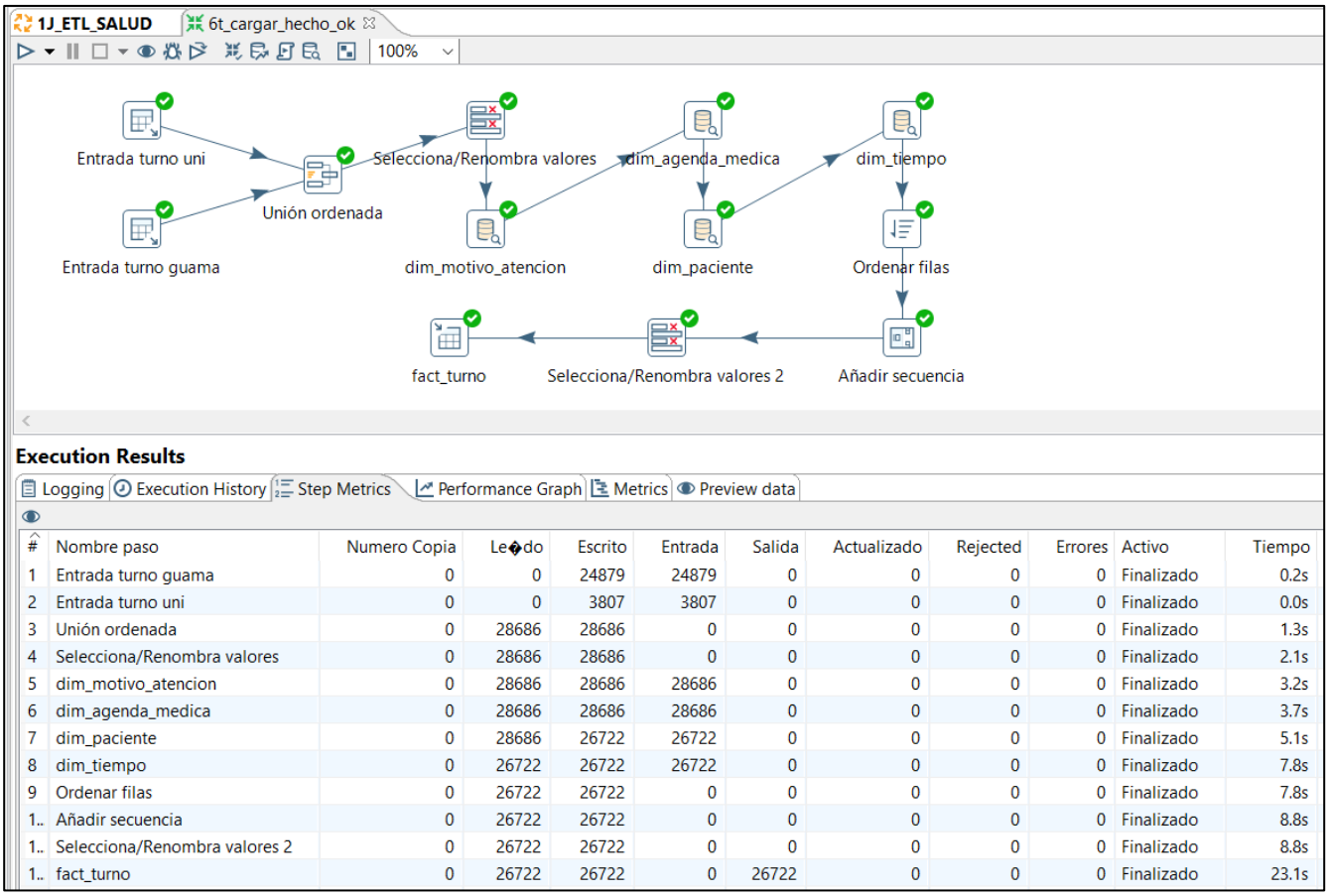
Nota. Proceso de transformación para la carga de la dimensión dim tiempo.

Cargar hecho

Mediante este trabajo se realiza la carga de la tabla de hecho del Data Mart. Utiliza como entradas las bases de datos ‘SGMAS Uni’ y ‘SGMAS guama’ con la tabla ‘turno’ unificada. Mediante el proceso ETL se creará la tabla de hecho ‘fact_turno’ la cual contendrá las claves primarias de las tablas de dimensiones. Ver figura 34.

Figura 34.

Transformación – fact turno



#	Nombre paso	Numero Copia	Le◆do	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo
1	Entrada turno guama	0	0	24879	24879	0	0	0	0	Finalizado	0.2s
2	Entrada turno uni	0	0	3807	3807	0	0	0	0	Finalizado	0.0s
3	Unión ordenada	0	28686	28686	0	0	0	0	0	Finalizado	1.3s
4	Selecciona/Renombra valores	0	28686	28686	0	0	0	0	0	Finalizado	2.1s
5	dim_motivo_atencion	0	28686	28686	28686	0	0	0	0	Finalizado	3.2s
6	dim_agenda_medica	0	28686	28686	28686	0	0	0	0	Finalizado	3.7s
7	dim_paciente	0	28686	26722	26722	0	0	0	0	Finalizado	5.1s
8	dim_tiempo	0	26722	26722	26722	0	0	0	0	Finalizado	7.8s
9	Ordenar filas	0	26722	26722	0	0	0	0	0	Finalizado	7.8s
1..	Añadir secuencia	0	26722	26722	0	0	0	0	0	Finalizado	8.8s
1..	Selecciona/Renombra valores 2	0	26722	26722	0	0	0	0	0	Finalizado	8.8s
1..	fact_turno	0	26722	26722	0	26722	0	0	0	Finalizado	23.1s

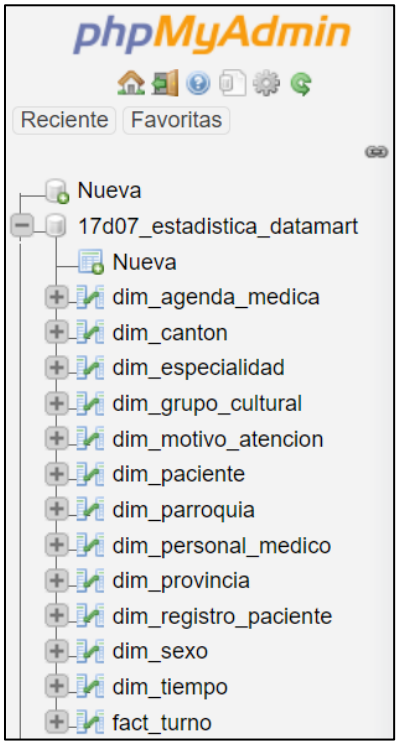
Nota. Proceso de carga de la tabla de hechos, la cual está conformada por las claves primarias de las tablas de dimensiones.

Resultado de carga de dimensiones y hecho

Al ejecutar con éxito el proceso ETL descrito anteriormente, se logró poblar las tablas de dimensiones y la tabla de hecho en el Data Mart denominado ‘17d07_estadistica_datamart’, el mismo que se encuentra alojado en una base de datos MySQL. En la figura 35 se muestran las tablas de dimensiones y tabla de hecho del Data Mart, en la figura 36 se muestra los datos cargados de una de las dimensiones, en este caso es la ‘dim_paciente’ (todas las dimensiones se encuentran con sus respectivos datos), y en la figura 37 se muestran los datos cargados en la tabla de hechos ‘fact_turno’.

Figura 35.

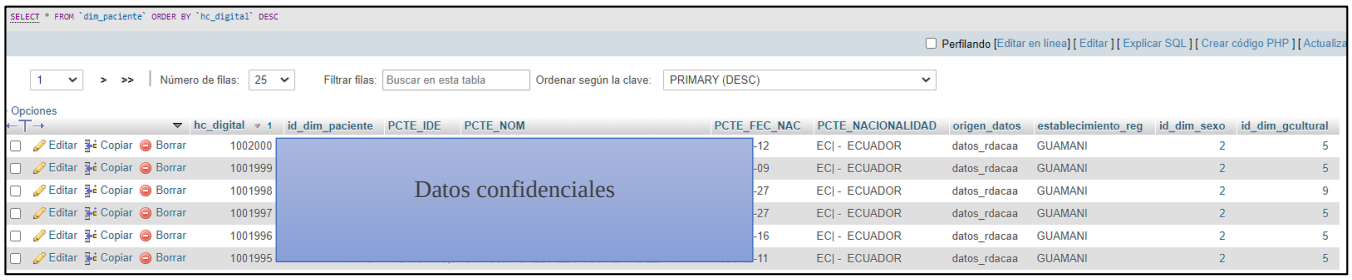
Tablas de dimensiones y tabla de hecho del Data Mart



Nota. Visualización de las tablas de dimensiones y tabla de hecho que conforman el Data Mart.

Figura 36.

Datos cargados dim paciente



Nota. Visualización de los datos de la dimensión ‘dim_paciente’ cargados mediante el proceso ETL.

Figura 37.

Datos cargados fact turno

Nota. Visualización de los datos de la tabla de hechos ‘fact_turno’ cargados mediante el proceso ETL.

Especificación de Aplicaciones de BI

Las aplicaciones de BI proporcionan información útil a los usuarios, e incluyen un amplio espectro de tipos de reportes y herramientas de análisis, que van desde informes simples de formato estático a sofisticadas aplicaciones analíticas que usan algoritmos complejos.

Los reportes proporcionan a los usuarios un conjunto básico de información acerca de lo que está sucediendo en un área determinada de la empresa.

Para este proyecto se desarrolló reportes con la herramienta Power BI Desktop, la cual es una plataforma de visualización de datos potente utilizada en el área de Inteligencia de Negocios, que permite generar informes empresariales, cuyo contenido se puede extraer de diferentes bases de datos con distintos formatos.

Desarrollo de Aplicaciones de BI

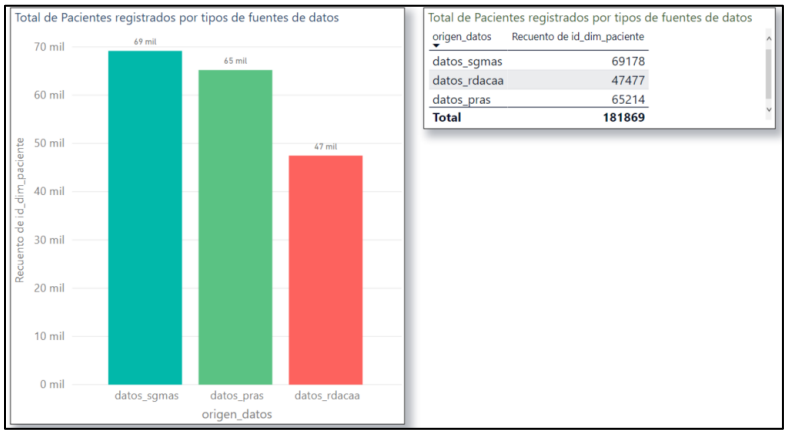
Este paso consiste en el desarrollo de las aplicaciones BI, como cubos OLAP, reportes, dashboard, etc. La creación de estas aplicaciones debe estar en directa relación con los requerimientos establecidos. Se considera tanto el diseño como la creación y prueba de las aplicaciones (Soto Olivares, 2011).

Reportes basados en los requerimientos funcionales

A continuación, se muestran los reportes obtenidos del Data Mart en la herramienta Power BI Desktop según los requerimientos funcionales de la Institución.

Figura 38.

Total de pacientes registrados por tipos de fuentes de datos.

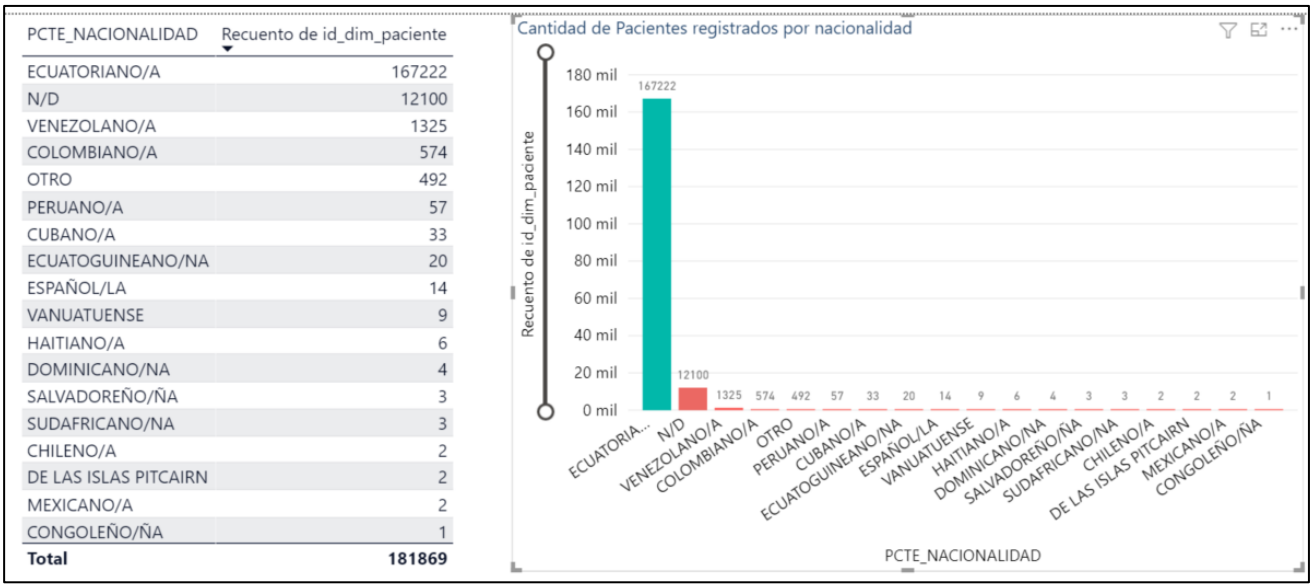


Total de pacientes registrados año 2019 en las fuentes de datos PRAS, RDACAA y SGMAS, mostrando el origen de las fuentes de datos a las que pertenecen.

Se analizó la cantidad de datos obtenidos de los diferentes sistemas informáticos que maneja la Institución, y se determinó que a nivel general de todos los Establecimientos de Salud siguen utilizando el sistema rdacaa cuya funcionalidad es obsoleta. Esto se debe a que los usuarios tienen poca adaptación al cambio y han venido trabajando con dicho sistema varios años.

Figura 39.

Cantidad de pacientes registrados por nacionalidad

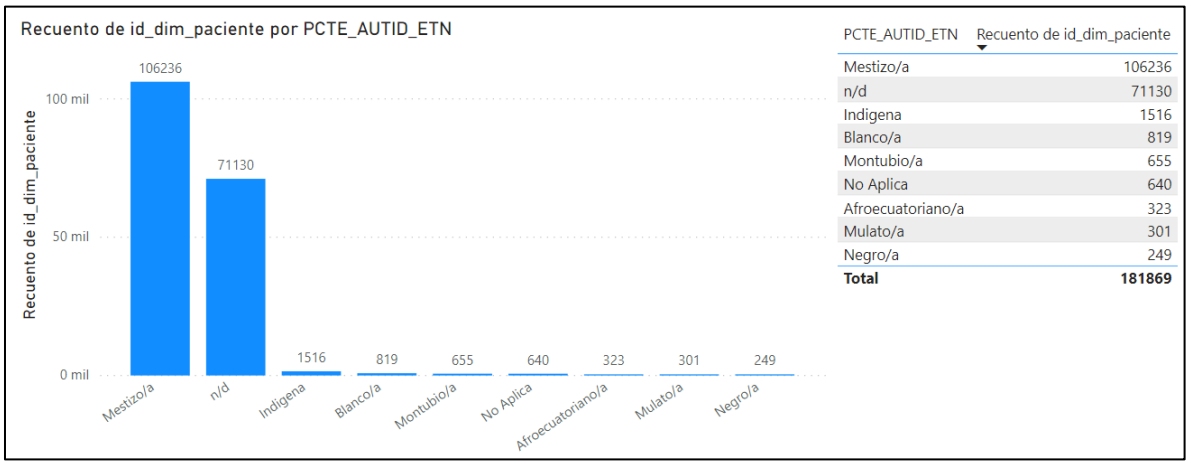


PCTE_NACIONALIDAD	Recuento de id_dim_paciente
ECUATORIANO/A	167222
N/D	12100
VENEZOLANO/A	1325
COLOMBIANO/A	574
OTRO	492
PERUANO/A	57
CUBANO/A	33
ECUATOGUINEANO/NA	20
ESPAÑOL/LA	14
VANUATUENSE	9
HAITIANO/A	6
DOMINICANO/NA	4
SALVADOREÑO/ÑA	3
SUDAFRICANO/NA	3
CHILENO/A	2
DE LAS ISLAS PITCAIRN	2
MEXICANO/A	2
CONGOLEÑO/ÑA	1
Total	**181869**

Se analizó la nacionalidad de los pacientes que más acuden a las atenciones médicas, esto con la finalidad de realizar un seguimiento personalizado de la historia clínica de los pacientes extranjeros. Hay que mencionar además que existe un buen número de pacientes no definidos (N/D), esto se debe a que los médicos no están llenando el campo de la nacionalidad del paciente al momento de llenar la ficha.

Figura 40.

Cantidad de Pacientes registrados por grupo étnico



PCTE_AUTID_ETN	Recuento de id_dim_paciente
Mestizo/a	106236
n/d	71130
Indigena	1516
Blanco/a	819
Montubio/a	655
No Aplica	640
Afroecuatoriano/a	323
Mulato/a	301
Negro/a	249
Total	**181869**

Se identificó la etnia de los pacientes para poder dar un tratamiento adecuado según las costumbres o alimentación que tiene cada grupo étnico, y a su vez para que el personal médico conozca sobre medicina ancestral. Hay que mencionar además que existen datos inexistentes identificados como ‘n/d’ debido a que el médico no ingresa correctamente la información, y otros datos identificados como ‘No Aplica’ cuando el paciente no proporciona dicha información.

Figura 41.

Cantidad de turnos agendados por mes y trimestres año 2019.

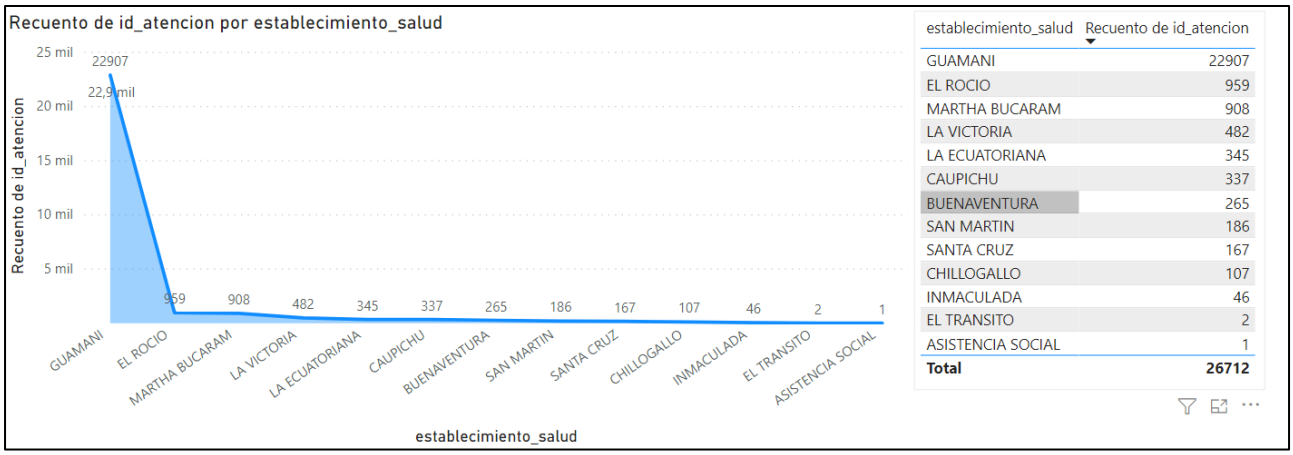
the_month	Recuento de id_agenda
julio	3999
abril	3642
marzo	3572
mayo	3553
junio	3389
febrero	3052
enero	3009
agosto	2496
Total	**26712**

quarter	Recuento de id_agenda
2	10584
1	9633
3	6495
Total	**26712**

Estos datos ayudan a identificar cuáles son los meses de mayor afluencia a consultas médicas para tomar decisiones respecto a la distribución del personal médico con proyección a los siguientes meses.

Figura 42.

Cantidad de pacientes agendados por establecimientos de salud



establecimiento_salud	Recuento de id_atencion
GUAMANI	22907
EL ROCIO	959
MARTHA BUCARAM	908
LA VICTORIA	482
LA ECUATORIANA	345
CAUPICHU	337
BUENAVENTURA	265
SAN MARTIN	186
SANTA CRUZ	167
CHILLOGALLO	107
INMACULADA	46
EL TRANSITO	2
ASISTENCIA SOCIAL	1
Total	**26712**

Con estos resultados se dio a conocer a los responsables de los establecimientos de salud que no se encuentran registrando los agendamientos en los Sistemas Informáticos, los reportes con menor agendamiento son en Asistencia Social, El Tránsito y La Inmaculada. Guamaní al ser un establecimiento tipo C (tipo Hospital) con mejor gestión por parte de las autoridades, maneja un mejor agendamiento en los Sistemas.

Figura 43.

Cantidad de turnos agendados por especialidad

Cantidad de Turnos Agendados por Especialidad

nombre_especialidad	Recuento de id_agenda
PSICOLOGÍA	806
PEDIATRÍA	3052
ODONTOLOGÍA	2936
OBSTETRICIA GENERAL	2359
NUTRICIÓN	1178
MEDICINA GENERAL	1120
MEDICINA FAMILIAR	8653
GINECO OBSTETRICIA	3662
FISIATRÍA	2944
ACUPUNTURA	2
Total	**26712**

Estos resultados permiten tomar decisiones respecto a la necesidad de evaluar el incremento de profesionales especialistas en las áreas más demandadas como: Medicina familiar, Gineco obstetricia y Pediatría.

Figura 44.

Cantidad de pacientes que asisten a las agendas por día y establecimiento de salud

the_day	ASISTENCIA SOCIAL	BUENAVENTURA	CAUPICHU	CHILLOGALLO	EL ROCIO	EL TRANSITO	GUAMANI	INMACULADA	LA ECUATORIANA	LA VICTORIA	MARTHA BUCARAM	SAN MARTIN	SANTA CRUZ	Total
lunes		117	122	36	253	2	5283	6	86	101	403	58	11	**6478**
martes		56	78	40	118		5132	12	72	124	195	30	55	**5912**
miércoles	1	70	62	12	146		4674	14	100	97	97	64	50	**5387**
jueves		21	75	17	59		4210	6	83	103	146	30	51	**4801**
viernes		1		2	151		3007	8	4	57	67	4		**3301**
sábado					51		601							**652**
domingo					181									**181**
Total	**1**	**265**	**337**	**107**	**959**	**2**	**22907**	**46**	**345**	**482**	**908**	**186**	**167**	**26712**

Se obtuvo la cantidad de agendamientos por establecimientos de salud para realizar un análisis o seguimiento al establecimiento e identificar las causas que originan el reporte de bajo agendamiento.

Figura 45.

Médicos con mayor agendamiento por mes, establecimiento de salud y especialidad

nombre_medico	establecimiento_salud	nombre_especialidad	Recuento de id_agenda	the_month
VICTOR RIVADENEIRA	GUAMANI	PEDIATRÍA	510	enero
VERONICA EGAS MF	GUAMANI	MEDICINA FAMILIAR	504	mayo
VICTOR RIVADENEIRA	GUAMANI	PEDIATRÍA	400	marzo
VICTOR RIVADENEIRA	GUAMANI	PEDIATRÍA	362	abril
SUAREZ GONZAGA NANCY PAOLA	EL ROCIO	MEDICINA FAMILIAR	348	julio
VICTOR RIVADENEIRA	GUAMANI	PEDIATRÍA	312	junio
VICTOR RIVADENEIRA	GUAMANI	PEDIATRÍA	297	mayo
SHAKIRA BELTRAN MF	GUAMANI	MEDICINA FAMILIAR	288	febrero
VERONICA EGAS MF	GUAMANI	MEDICINA FAMILIAR	288	enero
VERONICA EGAS MF	GUAMANI	MEDICINA FAMILIAR	282	abril
VICTOR RIVADENEIRA	GUAMANI	PEDIATRÍA	230	febrero
VICTOR RIVADENEIRA	GUAMANI	PEDIATRÍA	214	julio
VERONICA EGAS MF	GUAMANI	MEDICINA FAMILIAR	208	febrero
SHAKIRA BELTRAN MF	GUAMANI	MEDICINA FAMILIAR	188	enero
SOFIA DURAN MONTENEGRO	LA VICTORIA	MEDICINA FAMILIAR	181	agosto
SUAREZ GONZAGA NANCY PAOLA	EL ROCIO	MEDICINA FAMILIAR	136	agosto
VERONICA EGAS MF	GUAMANI	MEDICINA FAMILIAR	98	marzo
VALDIVIESO ROGEL ANA ELIZABETH	MARTHA BUCARAM	MEDICINA FAMILIAR	92	agosto
SOFIA DURAN MONTENEGRO	LA VICTORIA	MEDICINA FAMILIAR	81	julio
TOAZA PATIÑO ALDO VICENTE	CAUPICHU	PSICOLOGÍA	76	agosto
TENE RUEDA AMANDA	CAUPICHU	MEDICINA FAMILIAR	74	agosto
VILLAMARIN TAPIA ROSA MATILDE	MARTHA BUCARAM	OBSTETRICIA GENERAL	64	agosto
TOPA PILA ANGEL FABIAN	CHILLOGALLO	MEDICINA FAMILIAR	62	agosto

Con estos resultados analizamos a los médicos que tienen mayor agendamiento de pacientes para evaluar su desempeño laboral por cada mes, y así obtener un promedio de agendamiento mensual por establecimientos y especialidades.

Implementación

El hardware para esta implementación ya se encuentra instalado en uno de los servidores que posee la Institución, se considera que es necesario repotenciar el servidor para que soporte la carga de datos de las diferentes fuentes de información. La herramienta que se va a utilizar para el análisis y reportes de información es Power BI Desktop y para el proceso ETL se utilizará Pentaho Data Integration con spoon. A continuación, se muestra el hardware y software a utilizar para la implementación del Data Mart. Ver tabla 25.

Tabla 27.

Tecnología disponible para la implementación

Tipo	Producto	Característica	Uso
Software	Pentaho Data Integration 8.3 spoon	Diseñador gráfico de transformaciones y trabajos del sistema de ETL´s	Diseño y elaboración del proceso de extracción, transformación y carga
Software	PostgreSQL 9.5.22	Servidor de base de datos	Bases de datos transaccionales
Software	MySQL 15.1	Servidor de base de datos	Data Mart
Software	Power BI Desktop 2020.3.1	Herramienta de visualización de datos y creación de cubos olap	Reportes de información
Hardware	PC Desktop	Win10, 12 gb ram, disco duro 1tb, core i7	Servidor para implementación del Data Mart

Nota. Características de las herramientas utilizadas para la implementación del almacén de datos.

Mantenimiento y Crecimiento

Como parte de la metodología que se está aplicando y su ciclo de vida dimensional del negocio, el Data Mart va necesariamente a evolucionar y crecer con el tiempo. Se pretende implementar el Data Mart para el área de Estadística a nivel de Pichincha con todos los Distritos de Salud. Una vez concluido con la implementación del Data Mart se procede a realizar las pruebas funcionales, de integridad y carga.

Gestión del Proyecto

Asegura que las actividades del ciclo de vida dimensional del negocio se cumplan y se lleven de forma sincronizada. Entre sus actividades principales se encuentra la monitorización del estado del proyecto y el acoplamiento entre los requerimientos del negocio.

Capítulo IV

Conclusiones y Recomendaciones

Conclusiones

- Con la implementación del Data Mart se ha cumplido el objetivo general planteado ya que se ha logrado integrar las fuentes de datos heterogéneas en un repositorio de información centralizado, generando así un adecuado análisis de información y mejor toma de decisiones.
- Mediante el proceso de ETL se logró mejorar la calidad de los datos teniendo en consideración los requerimientos funcionales, sin embargo queda data por limpiar y transformar a nivel que vaya creciendo el proyecto.
- Utilizar una metodología de desarrollo de Data Warehouse como Kimball, es de gran ayuda debido a que aporta conocimiento y una secuencia de pasos que facilitó la realización del proyecto.
- Se puede decir que el desarrollo de un sistema de almacén de datos es complejo, debido a que abarca varios recursos como usuarios del negocio, desarrolladores, directivos, personal de tecnología, recursos económicos y físicos de la Institución, pues se debe contar tanto con hardware y software adecuado que cumpla con las necesidades del proyecto.
- Con la implementación del Data Mart se ha logrado la optimización de consultas multidimensionales facilitando así el manejo dinámico de los reportes de información.
- El uso de una herramienta como Power BI para la elaboración de reportes permite a los usuarios finales un manejo intuitivo y sencillo para generar reportes y análisis acorde a las necesidades del negocio.

Recomendaciones

- Tener en cuenta que el proyecto no finalizó ahí, sino que es necesario seguir alimentando el modelo generado, de tal manera que pueda ir mejorando el proceso de toma de decisiones en la Institución.
- Dedicar el tiempo necesario para conocer el negocio y sus requerimientos antes de iniciar con la implementación del Data Mart, esto ayudará a agilizar el trabajo al momento de su construcción y ejecución.
- Para implementar el proceso de ETL, se recomienda trabajar con una cierta cantidad de datos, no con la totalidad debido a la demora que puede generar el proceso.
- Es importante recolectar la información de los requerimientos directamente con los usuarios involucrados, ya que si es a través de intermediarios dicha información puede resultar mal entendida.
- Realizar capacitación al personal del área de Estadística y el área de Tecnología en el manejo de la aplicación de Power BI y Pentaho Data Integration respectivamente, con la finalidad de dar respuesta oportuna a los reportes y requerimientos solicitados.

Bibliografía

INESEM. (2020). *Los gestores de bases de datos más usados en la actualidad*. Obtenido de https://revistadigital.inesem.es/informatica-y-tics/los-gestores-de-bases-de-datos-mas-usados/

Área Tecnológica. (2016). *Data warehouse*. Obtenido de https://www.areatecnologia.com/informatica/data-warehouse.html

Areiza, E., Pérez, D., & Rivas, J. (2016). *Sistema de ayuda a la toma de decisiones.* Caracas.

Astera Software. (19 de 02 de 2019). *Integración de datos*. Obtenido de https://www.astera.com/es/type/blog/data-integration-tools-for-businesses/

Astera Software. (19 de 02 de 2019). *Integración de datos: qué es y cómo elegir la herramienta adecuada para su negocio*. Obtenido de https://www.astera.com/es/type/blog/data-integration-tools-for-businesses/?no_redirect=true

Bhushan, L. (2016). Data Lake Integration Design Principles. *Apress*, 38.

biverano. (04 de 09 de 2011). *Modelo Estrella y Modelo Copo de Nieve*. Obtenido de http://biverano2011.blogspot.com/2011/09/modelo-estrella-y-modelo-copo-de-nieve.html

Bustamante, X., Macas, M., & Beatriz, C. (2019). Data Warehouse: Análisis Multidimensional de BAFICI utilizando Power Pivot. *Espacios*, 24.

Cano, J. L. (2006). *Business Intelligence: Competir con información .* España .

Castillo, J., & Palomino, L. (2012). *Implementación de un Datamart como una solución de Inteligencia de Negocios para el área de logística de T-Impulso.* San Marcos: Revista de Investigación de Sistemas e Informática.

Colombia Digital. (24 de 07 de 2017). *Entendiendo la integración de datos y sus principales desafíos*. Obtenido de https://colombiadigital.net/actualidad/articulos-informativos/item/9824-entendiendo-la-integracion-de-datos-y-sus-principales-desafios.html

CoRegistros. (2020). *Las mejores bases de datos*. Obtenido de https://www.coregistros.com/2017/04/11/mejores-bases-de-datos/

De Pietro, G., Gallo, L., Howlett, R., Jain, L., Azaiez, N., & Akaichi, J. (2018). Integrating Trajectory Data in the Warehousing Chain: A New Way to Handle the Trajectory ELT Process. *Springer International Publishing*, 9.

De Pietro, G., Gallo, L., Howlett, R., Jain, L., Azaiez, N., & Akaichi, J. (2018). Integrating Trajectory Data in the Warehousing Chain: A New Way to Handle the Trajectory ELT Process. *Springer International Publishing*, 9.

Domínguez Martínez, J. (2008). *Diseño de un modelo multidimensional de data mart del área de capacitación en el INEGI.* Aguascalientes.

Espinosa, R. (19 de 04 de 2010). *El Rincon del BI*. Obtenido de https://churriwifi.wordpress.com/2010/04/19/15-2-ampliacion-conceptos-del-modelado-dimensional/

Evaluando Software. (31 de 10 de 2016). *Datawarehouse y Datamart*. Obtenido de https://www.evaluandosoftware.com/que-es-un-data-warehouse/

Haro, V., Pérez, W., Guzman, L., & Saquicela, V. (2014). *Diseño e implementación de un sistema de soporte de decisiones para el Centro de Documentación Regional "Juan Bautista Vázquez".* Cuenca.

Hiberus TI. (2016). *¿Qué es Microsoft Power BI?* Obtenido de https://www.hiberus.com/crecemos-contigo/data-storytelling-con-microsoft-power-bi-por-que-es-necesario/

Idea Consultoría Informática. (2017). *Las 10 características clave de Power BI que debes conocer*. Obtenido de https://www.ideaconsulting.es/caracteristicas-power-bi/

IMF International Business School. (2019). *Las 10 mejores herramientas de integración de datos*. Obtenido de https://blogs.imf-formacion.com/blog/tecnologia/10-herramientas-integracion-datos-201907/

Iruela, J. (8 de 4 de 2018). *Revista Digital INESEM*. Obtenido de https://revistadigital.inesem.es/informatica-y-tics/los-gestores-de-bases-de-datos-mas-usados/

LatinoBI. (2013). *The Dataearehouse Institute*. Obtenido de https://www.latino-bi.com/espanol/fundamentos-bi/introduccion-al-bi.php

Liu, L., & Ozsu, T. (2018). Encyclopedia of Database Systems. *Springer New York*, 538.

MediaPro. (2018). *Datastage: qué es, cómo funciona y cómo puede ayudar a tu empresa*. Obtenido de https://blog.mdcloud.es/datastage-que-es-como-funciona-y-como-puede-ayudar-a-tu-empresa/

MediaPro. (2018). *Herramientas ETL: comparativa y principales categorías*. Obtenido de https://blog.mdcloud.es/herramientas-etl-comparativa-y-principales-categorias/

Mundo BI. (22 de 04 de 2012). *Inmon y Kimball*. Obtenido de http://mundobi.com.ar/?p=614

mundodb.es. (09 de 11 de 2013). *Diseño Data warehouse: hechos y dimensiones; modelo estrella y copo de nieve*. Obtenido de http://mundodb.es/diseno-data-warehouse-hechos-y-dimensiones-modelo-estrella-vs-copo-de-nieve

Neteris Consulting. (2020). *Software Tableau*. Obtenido de https://neteris.com/software/tableau-software-visualizacion-datos

Pacco Palomino, R. (2013). *Sistema de Gestión Financiera basado en Sistemas de Información Ejecutiva y modelo Kimball.* Lima.

Power Data. (2015). *Integración de datos: Concepto e importancia en la empresa actual*. Obtenido de https://www.powerdata.es/integracion-de-datos

Power Data. (2019). *Integración de datos: Concepto e importancia en la empresa actual*. Obtenido de https://www.powerdata.es/integracion-de-datos

Prakash, N., & Prakash, D. (2018). Requirements Engineering for Data Warehousing. *Springer Singapore*, 32.

Puerta, A. (2016). *Business Intelligence y las Tecnologías de la Información.*

Rosales, C. (2009). *Análisis, diseño e implementación de un datamart para el soporte de toma de decisiones y evaluación de las estrategias sanitarias en las direcciones de salud.* Lima.

Siag Consulting. (06 de 09 de 2017). *Qlik Sense en tu empresa*. Obtenido de https://siagconsulting.es/qlik-sense-empresa/

Sinnexus. (2020). *Bases de datos OLTP y OLAP*. Obtenido de https://www.sinnexus.com/business_intelligence/olap_vs_oltp.aspx

Sinnexus. (2020). *Business Intelligence*. Obtenido de https://www.sinnexus.com/business_intelligence/

Soto Olivares, J. (2011). *Solución de Inteligencia de Negocios para una PYME.* Valparaíso.

Talend. (2020). *Diferencias entre ETL y ELT*. Obtenido de https://es.talend.com/resources/elt-vs-etl/

Taufik, T., Prabasari, I., Rineksane, I., Yaya, R., & Widowati, R. (2017). Developing Academic Executive Information System Uses Kimball Methodology: Case Study in an Indonesia Higher Education System. *Universitas Muhammadiyah Yogyakarta*.

Troyanx - Soluciones Informáticas. (2019). *Data Mart*. Obtenido de http://troyanx.com/Hefesto/data_mart.html

TutorialKart. (2018). *What is Data mart ?* Obtenido de https://www.tutorialkart.com/data-warehouse/what-is-data-mart/

Vargas, G. (2018). *Some data analytics elements.* France.

Yalan, J. (2013). *Implementación de un Datamart como una solución de Inteligencia de Negocios para el área de logística de T-impulso.* Lima.