

UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN

Facultad de Ingeniería

Escuela Profesional de Ingeniería en Informática y Sistemas

**ANÁLISIS MULTIDIMENSIONAL DEL RENDIMIENTO ACADÉMICO
DE LA UNJBG UTILIZANDO HERRAMIENTAS
*BUSINESS INTELLIGENCE***

TESIS

Presentada por:

Bach. Jorge Rafael Maquera Parihuana

Para optar el Título Profesional de:

Ingeniero en Informática y Sistemas

Tacna – Perú

2021

UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN

FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS

“ANÁLISIS MULTIDIMENSIONAL DEL RENDIMIENTO ACADÉMICO DE
LA UNJBG UTILIZANDO HERRAMIENTAS BUSINESS INTELLIGENCE”

**TESIS SUSTENTADA Y APROBADA EL 19 DE DICIEMBRE DEL 2019
ESTANDO EL JURADO CALIFICADOR INTEGRADO POR:**

Presidente :



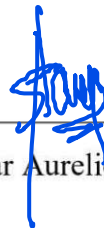
Dr. Jesús Plácido Medina Salas

Secretario :



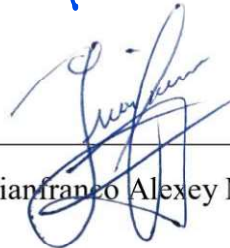
Dr. Edwin Antonio Hinojosa Ramos

Vocal :



MSc. Edgar Aurelio Taya Acosta

Asesor :



Mgtr Gianfranco Alexey Málaga Tejada

UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN
FACULTAD DE INGENIERÍA

JURADO CALIFICADOR Y CALIFICACIÓN DE LA SUSTENACIÓN DE TESIS

TESIS N° _____

TÍTULO PROFESIONAL DE:

Ingeniería en Informática y Sistemas

La Secretaría Académica de la Facultad de Ingeniería, por resolución de Facultad N° 06200-2019-FAIN/UNJBG, designó Jurado para la sustentación oral de la Tesis titulada “ANÁLISIS MULTIDIMENSIONAL DEL RENDIMIENTO ACADÉMICO DE LA UNJBG UTILIZADO HERRAMIENTAS BUSINESS INTELLIGENCE”

El mismo que está conformado por:

Presidente: Dr. Jesús Plácido Medina Salas

Secretario: Dr. Edwin Antonio Hinojosa Ramos

Vocal: MSc. Edgar Antonio Taya Acosta

Para calificar la sustentación de la Tesis en acto público el día 19 de diciembre del 2019. Presentado por el Bachiller Jorge Rafael Maquera Parihuana, de la Escuela Profesional de Ingeniería en Informática y Sistemas.

El Jurado Calificador en forma secreta e individual emitió su opinión sobre el tema de la tesis expuesta y procedió a obtener el promedio que arrojó el calificativo de aprobado con la nota de once (11).

Para ratificar lo detallado firman:



Dr. Jesús Plácido Medina Salas
Presidente



Dr. Edwin Antonio Hinojosa Ramos
Secretario



MSc. Edgar Aurelio Taya Acosta
Vocal

DEDICATORIA

Dedico la presente investigación a mi familia, en especial a mi abuelo que está en el cielo y a Dios, Él es mi guía, mi fuerza, mi fortaleza en cada paso que decido a través de mi vida en este mundo.

AGRADECIMIENTO

A mi padre, a mi madre quienes me brindaron su apoyo incondicional.

A todos quienes me apoyaron para que esta investigación se concrete, sobre todo a quien, su amistad fue entregada de forma sincera, supo motivarme y darme la confianza para seguir adelante en mis metas profesionales y personales. Siempre les estaré agradecido, siempre te agradeceré German Parihuana.

CONTENIDO

	Pág.
DEDICATORIA	iv
AGRADECIMIENTO	v
RESUMEN	xii
ABSTRACT	xiii
INTRODUCCIÓN	1
CAPÍTULO I PLANTEAMIENTO DEL PROBLEMA	4
1.1 Descripción del problema	4
1.1.1 Antecedentes del problema	4
1.1.2 Problemática de investigación	7
1.2 Formulación del problema	8
1.3 Justificación e importancia	8
1.4 Alcances y limitaciones	9
1.4.1 Alcances	9
1.4.2. Limitaciones	9
1.5 Objetivos	10
1.5.1 Objetivos general	10
1.5.2 Objetivos específicos	10
1.6 Hipótesis	11

CAPÍTULO II MARCO TEÓRICO	11
2.1 Antecedentes del estudio	11
2.2 Bases teóricas	13
2.2.1 Business Intelligence	16
2.2.2 Orígenes de los datos	21
2.2.3 Extraer, transformar y cargar	22
2.2.4 MySQL WorkBench	26
2.2.5 Data Warehouse y Data Mart	29
2.2.6 Modelo Dimensional	36
2.2.7 Modelo multidimensionales	40
2.2.8 Software libre vs Software propietario	41
2.2.9 Pentaho Business Intelligence	44
2.2.10 Rendimiento académico	49
2.2.11 Indicadores del rendimiento académico	51
2.2.12 Marco metodológico	52
2.3 Definición de términos	55
CAPÍTULO III MARCO METODOLÓGICO	57
3.1 Diseño experimental o no experimental	57
3.2 Población y muestra	58
3.2.1 Población	58
3.2.2 Muestra	59

3.3 Operacionalización de variables	61
3.4 Técnicas e instrumentos para recolección de datos	62
3.5 Procesamiento y análisis de datos	63
3.5.1 Fuentes de datos	67
3.5.2 Modelo lógico data stage area	69
3.5.3 Modelo físico stage	71
3.5.4 Calidad de datos	75
3.5.5 Procesos ETL diseñados	77
3.5.6 Modelo Físico DM	86
3.5.7 Modelo Lógico DM	97
3.5.8 Diseño OLAP	98
3.5.9 Diseño de Reportes	101
CAPÍTULO IV RESULTADOS Y DISCUSIÓN	112
4.1 Resultados	112
4.2 Discusión	115
CONCLUSIONES	118
RECOMENDACIONES	119
REFERENCIAS BIBLIOGRÁFICAS	120
ANEXOS	125
Anexo 1	126
Anexo 2	127

ÍNDICE DE FIGURAS

	Pág.
Figura 1: Ciclo de vida de las aplicaciones de soporte a la decisión	14
Figura 2: Arquitectura de business intelligence	16
Figura 3 :Cadena de valor del dato	19
Figura 4 :Orígenes de datos.	21
Figura 5 :Tabla data warehouse.	31
Figura 6 :Arquitectura del data warehouse	33
Figura 7 :Ejemplo de modelo estrella.	35
Figura 8 :Ejemplo de modelo estrella.	36
Figura 9 :Ejemplo de modelo copo de nieve	39
Figura 10 :Pentaho schema workbench	46
Figura 11 :Pentaho data integration	46
Figura 12 :Pentaho report designer	49
Figura 13 :Indicadores del rendimiento académico.	51
Figura 14 :Ciclo de vida de la metodología	54
Figura 15 :Archivo transaccional notas	68
Figura 16 :Archivo plano notas	68
Figura 17 :Archivo plano estudiantes.csv	69
Figura 18 :Data stage area notas	70
Figura 19 :Modelo lógico stage	72
Figura 20 : Archivo notas	77
Figura 21 : Pentaho data integration	78
Figura 22 : Pentaho data integration	78
Figura 21 :Proceso dsa estudiantes	79

Figura 22 :Proceso dsa notas	80
Figura 23 :Proceso de la dimensión especialidad	80
Figura 24 :Proceso de la dimensión materia	81
Figura 25: Proceso de la dimensión persona	81
Figura 26 :Proceso de la dimensión estado	82
Figura 27 : Proceso de la dimensión periodo	82
Figura 28 : Proceso de la dimensión sexo	83
Figura 29 :Proceso tabla hechos fain notas	84
Figura 30 :Proceso job notas	85
Figura 31 : Modelado de las dimensiones	91
Figura 32 : Modelo físico DM	92
Figura 33 : Modelo lógico DM	97
Figura 34 : Cubo olap fain	100
Figura 35 : Interfaz saiku analytics	101
Figura 35 : Cuadro estadístico del reporte a	103
Figura 37 : Grafico de barras del reporte a	104
Figura 38 :Cuadro estadístico del reporte b	105
Figura 39 :Grafico de barras del reporte b	106
Figura 40 : Cuadro estadístico del reporte c	107
Figura 41 : Grafico de barras del reporte c	108
Figura 42 :Cuadro estadístico del reporte d	109
Figura 43 : Grafico de barras del reporte d	110
Figura 44 Cuadro estadístico del reporte e	111
figura 45 Grafico de barras del reporte e	112

ÍNDICE DE TABLAS

	Pág.
Tabla 1: Comparativa herramientas de Pentaho	43
Tabla 2: Registros académicos FAIN	58
Tabla 3: Operacionalización de variables	61
Tabla 4: Matriz de categoría de la revisión documental	62
Tabla 5: Tablas del modelo data stage area	73
Tabla 6: Tabla del modelo DSA estudiantes	73
Tabla 7: Tabla del modelo DSA notas	74
Tabla 8: Tiempo de ejecución de los procesos de Integración de Datos	85
Tabla 9: Tipo de datos del Data Warehouse	87
Tabla 10: Llaves primarias del Data Warehouse	88
Tabla 11: Estructura del Data Warehous	93
Tabla 12: Estructura de la dimensión periodo	93
Tabla 13: Estructura de la dimensión materia	94
Tabla 14: Estructura de la dimensión especialidad	94
Tabla 15: Posibles estados del rendimiento académico	95
Tabla 16: Estructura de la dimensión estado	95
Tabla 17: Estructura de la dimensión sexo	95
Tabla 18: Estructura de la dimensión persona	96
Tabla 19: Estructura de la tabla Hechos	96

RESUMEN

La Universidad Nacional Jorge Basadre Grohmann (UNJBG), cuenta con una gran cantidad de datos de la formación académica de los estudiantes. Actualmente no se cuenta con una herramienta informática enfocada al proceso de toma de decisiones en relación al rendimiento académico, siendo materia de estudio para esta investigación los datos académicos de los alumnos de la Facultad de Ingeniería. Para el desarrollo de esta investigación se aplicó Business Intelligence (BI) orientado al análisis del rendimiento académico. Se implementó un modelo multidimensional, que se utiliza para generar los reportes especializados del rendimiento académico, los que mostraron indicadores del rendimiento académico de los estudiantes. Estos son: el conocimiento que ayuda y apoya a la toma de decisiones, y formación de estrategias. Se concluyó que el modelo multidimensional desarrollado permitió organizar y almacenar los datos relevantes al rendimiento académico de los estudiantes de la Facultad de Ingeniería. Mediante el procesamiento analítico en línea agilizamos la consulta de esta gran cantidad de datos, obteniendo indicadores del rendimiento académico. Proporcionando información que ayuda a la toma de decisiones, teniendo un panorama de la situación del rendimiento académico y así crear estrategias para evitar potenciales fracasos académicos.

Palabras claves: Business Intelligence, Rendimiento académico.

ABSTRACT

The Universidad Nacional Jorge Basadre Grohmann (UNJBG), has a large amount of data on the academic training of students. Currently there is no computer tool focused on the decision-making process in relation to academic performance, being the subject of study for this research the academic data of the students of the Faculty of Engineering. For the development of this research, Business Intelligence (BI) oriented to the analysis of academic performance was applied. A multidimensional model was implemented, which is used to generate specialized reports of academic performance, which showed indicators of academic performance. The reports showed indicators of the academic performance of the students. These are: knowledge that helps and supports decision making, and strategy formation. It was concluded that the multidimensional model developed allowed organizing and storing data relevant to the academic performance of the students of the School of Engineering. By means of online analytical processing, we streamlined the consultation of this large amount of data, obtaining indicators of academic performance. Providing information that helps decision making, having an overview of the academic performance situation and thus creating strategies to avoid potential academic failures.

Keywords: Business Intelligence, Academic performance.

INTRODUCCIÓN

Alcántara (2016) explica que su investigación “surge por la necesidad de no contar con un sistema que solucione las necesidades de tipo informático dentro de la universidad para ayudar en el soporte y proceso para la toma de decisiones en el rendimiento académico del estudiante”. Utiliza para el desarrollo la metodología de Raplh Kimbal, que se recoge como proceso a seguir en el desarrollo de un Data Warehouse. Concluyen que “se logró implementar una solución informática que agiliza la toma de decisiones utilizando para ello herramienta de inteligencia de negocio”.

Antamba (2015) explica que se “observada la carencia de una plataforma tecnológica estructurada y especializada que permita explotar los datos almacenados y provea de canales dinámicos de acceso autónomos a la información”. Planteándose como objetivo “disponer de una herramienta de BI para generar información analítica y estadística sobre el estado del rendimiento académico” utilizando una metodología en cascada “(Antamba, 2015), la cual brinda una secuencia lógica y eficaz en el desarrollo del DWH; además está garantiza que el producto final cumpla todos los requisitos que el usuario final necesita para el análisis estadístico sobre el rendimiento académico”(Antamba, 2015), esta metodología será la usada como para la presente investigación.

La Universidad Nacional Jorge Basadre Grohmann cuenta con una gran cantidad de datos relacionados al rendimiento académico de los estudiantes, que están gestionados por el sistema de información de la Dirección Académica de Actividades y Servicios Académicos (DASA). De la base de datos de este sistema se pueden obtener datos estadísticos, mas no información estratégica del rendimiento académico que pueda ser usada para tomar decisiones por la alta dirección. La información estratégica, como ser reportes especializados que muestren indicadores o patrones, puede ser obtenida usando métodos y herramientas de la BI. La Universidad Nacional Jorge Basadre Grohmann no tiene implementada la BI, por lo que no es posible acceder directamente a los reportes especializados o patrones relacionados al rendimiento estudiantil, que se puede transformar en conocimiento útil para las decisiones estratégicas de la gestión académica de la universidad.

En la presente investigación se realizará un análisis de un modelo multidimensional orientado al rendimiento académico que, usando herramientas de BI, permitirá hacer un análisis, obteniendo información estructurada para ayudar en la toma de decisiones.

La presente investigación contiene en el capítulo I la descripción del problema, se exponen sus antecedentes, su justificación e importancia, así como los

alcances y limitaciones de la investigación; se formulan el objetivo general y específicos; y se trata el tema de la hipótesis.

En el capítulo II se exponen los antecedentes de la investigación, el marco teórico donde se explica los procesos que involucran aplicar BI, y se definen los términos que se usarán en el presente trabajo.

En el capítulo III se definen el tipo y diseño de la investigación, la población y muestra, la operacionalización de la variable, la técnica e instrumentos para la recolección de datos, y se realiza el procesamiento y análisis de datos.

En el capítulo IV se muestran los resultados de la investigación, y se realiza la discusión de estos.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

1.1 Descripción del problema

1.1.1 Antecedentes del problema

Antamba (2015) en su tesis “Desarrollo de una solución de Business Intelligence orientada al análisis del rendimiento académico y la planificación de los cursos de la carrera de Ingeniería Informática en la Facultad de Ingeniería, Ciencias Físicas y Matemática”, explica que “El sistema académico universitario brinda información de notas, materias y otros, sin embargo no tienen una herramienta que permita analizar e interpretar la información” (Antamba, 2015). Su propósito general es disponer de una herramienta de BI para generar información analítica en la Facultad. Concluyendo que la herramienta de BI desarrollada permite organizar y conocer el rendimiento académico estudiantes dando a cada destinatario la información que necesita, mediante la de consulta reportes.

López (2007) en su tesis “Análisis, diseño e implementación de una *data mart* para la Dirección Financiera y Recursos Humanos de la Escuela Politécnica del Ejército para una toma de decisiones efectiva” se orienta a brindar una herramienta de apoyo en la tecnología BI de Oracle que sea capaz de solventar la información solicitada, logrando formar cubos dimensionales de decisión. Utilizó la metodología *Rapid Warehousing*. Este método es iterativo, y está basado en el desarrollo incremental del proyecto de Data Warehouse. Implementó una herramienta apoyada en la tecnología BI, la cual es capaz de solventar información solicitada en la Dirección Financiera y Recursos Humanos, permitiendo a los directivos tener información precisa y confiable en apoyo a la toma de decisión efectiva.

Castro y Huaynate (2016) en su tesis “Implementación de una solución de inteligencia de negocio orientada a la mejora de proceso de formación académica estudiantes en la UCH”, implementó una solución de BI para la mejora de proceso de formación estudiantil, mejorando el rendimiento académico mediante el análisis de la solución de BI, obtención de reportes para la toma de decisiones. Concluyó que los resultados obtenidos en los reportes son los indicadores que ayudarán a observar un panorama del rendimiento académico y se podrán tomar mejores acciones para la creación de estrategias.

Pascal, Servetto, Lobo y Luna (2017) en su tesis “Aplicación de Business Intelligence para la toma de decisiones en instituciones universitarias. implementación de boletines estadísticos en la Universidad Nacional de Lomas de Zamora”, expusieron que en las universidades es necesario incorporar tecnologías para el apoyo a la toma de decisiones. El objetivo es disponer de información relevante y confiable para realizar seguimiento de las funciones principales de la institución. Concluyeron que se provee a las autoridades un instrumento para la autoevaluación y la toma de decisiones basadas en el conocimiento de la situación real de la institución.

Díaz, Osorio, Amadeo, y Romero (2013), en su artículo científico “Aplicando estrategias y tecnologías de inteligencia de negocio en sistemas de gestión educativa”, analiza el perfil del estudiante de la Facultad de Informática de la Universidad Nacional de La Plata, carreras y planes de estudio, y la actividad de los docentes en relación al desenvolvimiento de sus materias. Permitió realizar acciones para resolver distintas situaciones identificadas a partir de datos confiables y fiables. Como resultado, se ha afianzado la materia optativa “Tecnologías aplicadas a BI”, también se ha realizados convenios de capacitación.

1.1.2 Problemática de investigación

La Universidad Nacional Jorge Basadre Grohmann cuenta en la actualidad con 34 carreras profesionales; a lo largo de estos años ha podido acumular una gran cantidad de datos, con los cuales, se puede analizar los siguientes puntos.

Con respecto al rendimiento académico de los alumnos en las diferentes carreras, ciclos, turnos, semestres o años académicos; al culminar el semestre académico y registrar los resultados obtenidos por los estudiantes, vemos que algunas veces no son los resultados que se esperaban, partiendo de dos de los tres objetivos de la Universidad que tienen como premisa “Mejorar la calidad de la formación profesional universitarios”(Universidad Nacional Jorge Basadre Grohmann, 2018) y “Fortalecer la investigación científica, tecnológica y humanística de la comunidad académica”(Universidad Nacional Jorge Basadre Grohmann, 2018).

Desde que la Universidad Nacional Jorge Basadre Grohmann inició su labor, han existido casos en los cuales los alumnos han desertado de la carrera profesional por diferentes motivos, y algunos de ellos son las expectativas del estudiante con respecto al nivel académico, existiendo dentro de las carreras profesionales un alto grado de deserción de estudiantes que, con un adecuado análisis, pueden generar conocimiento específico de la situación del rendimiento académico a lo largo del tiempo.

1.2 Formulación del problema

¿Es factible aplicar un análisis multidimensional del rendimiento académico de la Universidad Nacional Jorge Basadre Grohmann utilizando herramientas Business Intelligence?

Problemas específicos

- a) ¿Es factible implementar un modelo multidimensional a partir de los archivos transaccionales de rendimiento académico?
- b) ¿Es factible determinar las dimensiones que permitan generar indicadores para la toma de decisiones?
- c) ¿Es factible elegir una herramienta BI que permita obtener reportes especializados?

1.3 Justificación e importancia

Este estudio está dirigido a apoyar el proceso de formación estudiantil a través del conocimiento generado a partir de los datos del rendimiento académico. Las herramientas de BI, a través de un análisis multidimensional, permiten generar reportes que representan la información crítica.

La información procesada, presentada en estos reportes será el conocimiento que permitirá evaluar, interpretar e inferir una solución oportuna y sólida de mejora dentro del proceso de formación, por parte de las autoridades pertinentes de la Universidad Nacional Jorge Basadre Grohmann.

1.4 Alcances y limitaciones

1.4.1 Alcances

Esta investigación servirá para indicar en donde existen dificultades en el rendimiento académico de los alumnos, filtrando la información de los estudiantes. La solución será dirigida al personal de la Dirección Académica y Registros Académicos.

Con el apoyo de esta implementación, la Universidad Nacional Jorge Basadre Grohmann podrá acceder a reportes especializados por periodos, estados, materias, carreras, cursos, etc., de la Facultad de Ingeniería (FAIN) y de esta manera se podrá agilizar el análisis, planificación y mejoramiento continuo del proceso de formación del estudiante.

1.4.2. Limitaciones

Esta implementación estará dirigida solamente a la FAIN. Será utilizada por los responsables de las siguientes áreas: Dirección Académica de la FAIN, y registros académicos.

Facilitará acceso a información que permitirá tomar decisiones estratégicas que permitan apoyar en solucionar problemas sobre el rendimiento académico, pero no será una solución directa en la mejora de la formación estudiantil.

No será una solución aplicable a toda la Universidad en general, sino una solución para la mejora continua en la formación universitaria.

1.5 Objetivos

1.5.1 Objetivos general

Analizar un modelo multidimensional del rendimiento académico en la Universidad Nacional Jorge Basadre Grohmann utilizando Business Intelligence.

1.5.2 Objetivos específicos

- a) Implementar un modelo multidimensional a partir de los archivos transaccionales de rendimiento académico.
- b) Determinar las dimensiones que permitan generar indicadores para la toma de decisiones.
- c) Elegir una herramienta BI que permita obtener reportes especializados.

1.6 Hipótesis

No aplica. No en todas las investigaciones cuantitativas se plantean hipótesis. El hecho de que formulemos o no hipótesis depende de un factor esencial: el alcance inicial del estudio. La presente investigación presentará como una descripción el análisis de un modelo multidimensional que se llevará a cabo, sin que se pretenda pronosticar valor alguno o relacionar algunos conceptos o variables. Por lo indicado, esta investigación tendrá un alcance descriptivo. (Hernández Sampieri, Fernández Collado, y Baptista Lucio, 2014, p. 92)

Pollo (2008) en su tesis maestra "Análisis de precisión de técnicas de agregación en contextos experimentales poco maduros", indica que "los métodos de agregación deben ser considerados métodos de estimación de parámetros y no de contraste de hipótesis, a pesar de que sus resultados se utilizan para determinar si un tratamiento es mejor que otro", siendo sustento para que su investigación no incluya una hipótesis. En la misma línea, Málaga (2016) en su tesis maestra "Modelo de Gestión de Incidentes Basado en ITIL v.3", hace referencia a (Hernández Sampieri et al., 2014, p. 104) sustentando que "para una investigación descriptiva, en la que no se pronostica hecho o dato alguno, no es posible plantear una hipótesis."

CAPÍTULO II

MARCO TEÓRICO

2.1 Antecedentes del estudio

Usando una arquitectura de Data Warehouse (DWH) o Data Mart (DM), es posible almacenar información histórica. Para realizar un análisis predictivo se usaron redes neuronales en el trabajo de Zambrano, Rojas, Carvajal, y Acuña (2011), para realizar “el análisis del desempeño académico de estudiantes”.

Cobo, Gurmendi, Menéndez, y others (2016), usaron un “modelo de análisis que permitió implementar herramientas de Business Intelligence en el Sistema Universitario Argentino” para brindar más y mejores servicios, y generar datos en mayor calidad y cantidad. Pentaho fue la herramienta que permitió la integración, construcción de la estructura multidimensional para luego poder realizar su respectivo análisis y reportes de dicho modelo.

Zambrano (2011), nos dicen que “es dable indicar que al obtener resúmenes y reporte usando Data Warehouse, producto del análisis histórico de los datos, es posible crear una base sólida de información para la arquitectura RNA y la predicción de comportamiento futuro”, pues como se muestra en el trabajo es posible predecir los indicadores de gestión que se obtiene del DWH.

A diferencia del trabajo se analizará un modelo multidimensional Data Mart basándonos en el Procesamiento Analítico en Línea (OLAP), que es una solución utilizada en la BI, cuyo objetivo es agilizar la consulta de gran cantidad de datos, pasando de un enfoque general a lo particular y viceversa, es decir, de estar viendo un contexto general del desempeño académico de FAIN, a pasar a alguna división particular como un reporte por curso de alguna de las carreras de la facultad, con una simple selección de variables de entrada.

2.2 Bases teóricas

En un tiempo atrás las corporaciones se desarrollaban mediante sus departamentos tecnológicos informáticos para elaborar reportes estándares y condicionados respecto al *hardware*. Esto empieza en el periodo que se manipulaban los ordenadores *mainframes* y minicomputadoras cuando los usuarios no contaban con acceso directo a los ordenadores.

Se empleaban los sistemas principalmente para transacciones de negocios y el número de informes era limitado por estos tipos de dispositivos, además existía congestión de procesos por lo que los usuarios debían esperar días o semanas para la generación de reportes.

“Se desarrollaron al cabo de un tiempo los sistemas de información ejecutiva que apoyaron de soporte a las labores de ejecutivos y administradores. Con la aparición de la computadora personal, enlaces de computadora, avance de los procesos y aplicaciones informáticas las herramientas BI permitieron a los usuarios crear rutinas propias y así generar reportes personalizados.” (Moss, 2013)

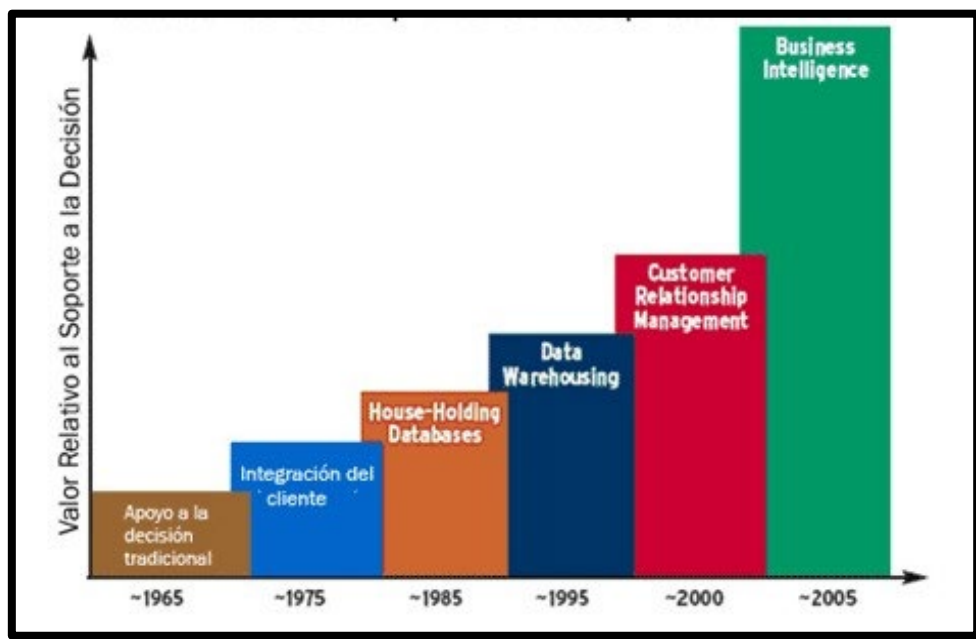


Figura 1: Ciclo de vida de las aplicaciones de soporte a la decisión

Fuente: Moss, 2013

- 1969: Origen del concepto de base de datos.
- 1970: Surgen los sistemas de gestión bases de datos, modelos relacionales y las primeras aplicaciones empresariales (*SAP, JD EData Warehouse ards, Siebel, PeopleSoft*). Es posible el “data entry” en los sistemas, maximizando la información disponible. Aún no es posible brindar un acceso rápido y sencillo a la información. Los informes son estáticos y sumamente orientados a la información transaccional.
- 1980: Creación del concepto Data Warehouse (Kimball, Inmon), y desarrollo de los primeros sistemas de *reporting*. Aún es complicado y funcionalmente deficiente. Existían relativamente potentes sistemas de bases de datos, pero no había aplicaciones que facilitasen su explotación.
- 1989: Introducción del término Business Intelligence.
- 1990: Business Intelligence 1.0. Avance de múltiples aplicaciones BI. Estos proveedores resultaban costosos, pero facilitaron el acceso a la información.
- 2000: Business Intelligence 2.0. Solidificación de las aplicaciones BI en unas pocas plataformas Oracle, SAP, IBM, Microsoft. A parte de la información estructurada, se empieza a considerar otro tipo de información y documentos no estructurados.

2.2.1 Business Intelligence

La BI es un conjunto de procesos y métodos que permiten transformar y consolidar los datos disponibles de distintos tipos de fuentes de una organización en información significativa (conocimiento útil) al servicio de los intereses de una organización y expuesta en escenarios, pronósticos e informes dinámicos de acuerdo a las peticiones del usuario y mediante esto establecer decisiones y resoluciones que establezcan mejora en el desarrollo estratégico, táctico, operativo y así tener un óptimo desempeño en las funciones y procesos administrativos. (Buigues, 2015)

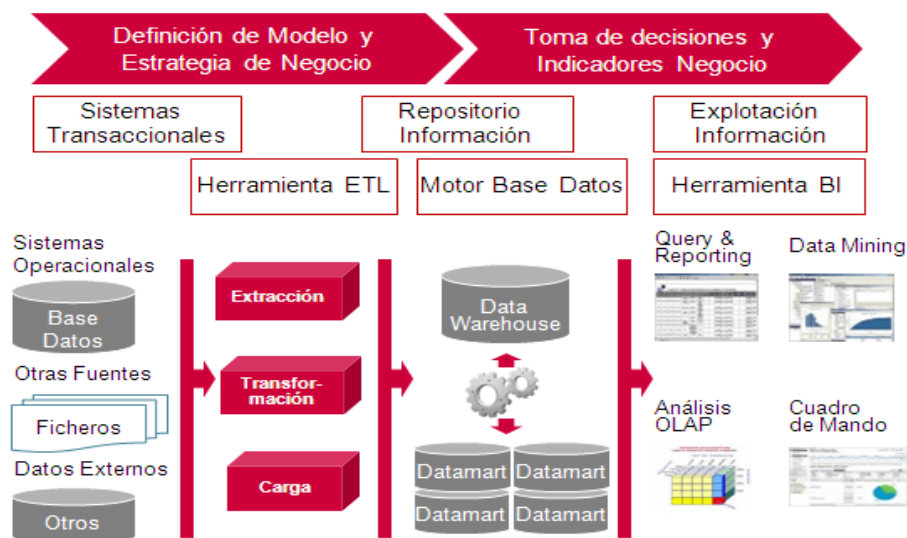


Figura 2: Arquitectura de Business Intelligence
Fuente: DIARIUM, 2016

2.2.1.1 Principios de BI

- Accesibilidad a la información: La herramienta en conjunto con las técnicas facilita a los usuarios el acceso directo y autónomo mediante una “vista” apropiada de los datos deseados de toda aquella información analítica independiente de la procedencia de esta.
- Apoyo a la toma de decisiones: el resultado de interpretar y analizar información adecuadamente presentada de acuerdo las peticiones del usuario, a fin de tomar mejores y rápidas decisiones mediante la identificación de problemas u oportunidades en los procesos.
- Orientación al usuario final: llegar alcanzar la autonomía entre los conocimientos técnicos de los usuarios y su capacidad para utilizar estas herramientas.

2.2.1.2 Uso de la Inteligencia de Negocios

- Adquirir información de las organizaciones de manera oportuna y precisa.
- Evaluar el desempeño organizacional.
- Predicción de eventos futuros.
- Descubrimiento de conocimiento oculto.
- Ejecución y monitorización de planes operacionales.
- Analizar condiciones de mercado y entornos de mercado.
- Conocer las tendencias y previsiones.

- Proyección de escenarios ocultos.
- Centraliza datos dispersos.
- Obtener métricas clave del negocio.
- Análisis multidimensionales.

2.2.1.3 Cadena de valor

Dato: Es la materia prima con los que cualquier organización dispone. También se conoce como la representación de una realidad discreta, no orientativos, circunscrita a un contexto reducido, de carácter operacional, radica en la adquisición de los datos de partida, a través de su recepción, estructuración y registro en los sistemas operacionales pertinentes.

Información: Es el conjunto estructurado de datos procesados. Resultantes de la incorporación, acoplamiento y contextualización de muchos datos con un objetivo determinado. Produciendo conocimientos o inteligencia, y cambia el estado del individuo que percibe dicho mensaje, ocasionando efecto sobre sus juicios de valor y sus comportamientos.

Conocimiento: Es el producto de inferir y procesar información debidamente presentada, buscando el descubrimiento o la verificación de una realidad compleja antes desconocida o solo supuesta. Basado en la interpretación de información empírica y cuantificable. Esta etapa se ocupa de diseñar, construir

y mantener los canales de distribución de información que permitan y faciliten la interpretación y el análisis por parte de los usuarios finales.

Decisión: Es el resultado de la cadena de valor, en donde se practica una o varias acciones, normalmente para cambiar algunas de las realidades constatadas mediante el conocimiento en la fase anterior, y en esto radica su propuesta de valor: en la transformación sistemática de datos en decisiones informadas, pertinentes y orientadas a la mejora, a cualquier nivel, y en cualquier ámbito.

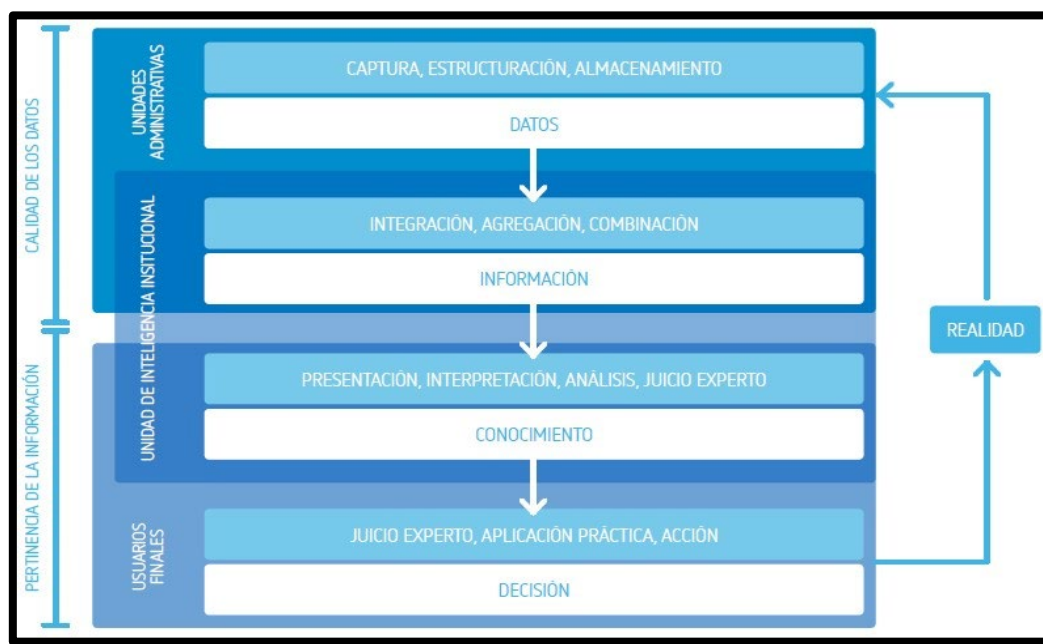


Figura 3 :Cadena de valor del dato
Fuente: Davenport, 1998

2.2.1.4 Elementos de Inteligencia de Negocios

Multidimensional: La información proviene de diversas fuentes tales como: hojas de cálculo, bases de datos, la herramienta de BI debe poder reunir información completa, aunque se encuentre separadas para resumirla y tenga un alcance de análisis profundo, con bases sólidas y datos actualizados.

Data mining: Hace una adecuada clasificación de acuerdo a las actividades competentes y observar tendencias, comportamientos o cambios.

Agentes: Los softwares analizan los datos sin necesidad del raciocinio, pues están programadas para interpretar datos.

Data Warehouse: Como respuesta de las TIC'S para descentralizar la toma de decisiones, colocando está en las áreas necesarias de la información.

2.2.2 Orígenes de los datos

Las distintas fuentes de información con las que podemos suministrar un DWH (Figura 4):

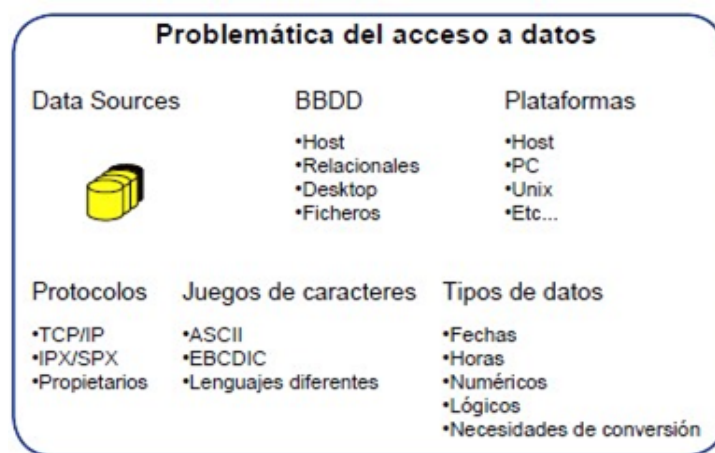


Figura 4 :Orígenes de datos.

Fuente: Business Intelligence: Competir con información (2007).

- Planificación de recursos empresariales.
- Base de Datos: *Open Source* (Mysql, PostgreSQL, LucidDB). Comercial (Oracle, SQLServer, IBMDB2).
- Fuente de información externa.

La calidad de los datos debe ser la máxima para no generar errores en el DWH, la responsabilidad de la calidad corresponde a cada uno de los propietarios de los procesos y de las aplicaciones que los soportan.

Según Giner (2007) las características que debería cumplir los datos para tener una buena calidad (p. 102):

- Precisión: Representar con precisión una realidad o una fuente de datos que se pueda verificar.
- Integridad: Mantener constantemente la estructura de los datos y la relaciones a través de las entidades y los atributos.
- Coherencia: Los elementos de los datos son entendidos y comprendidos.
- Totalidad: Se cuenta con todos los datos necesarios.
- Validez: Sus valores son aceptables en los rangos definidos por el negocio.
- Disponibilidad: Están disponibles cuando se necesitan.
- Accesibilidad: Se puede acceder a los datos fácilmente.

2.2.3 Extraer, transformar y cargar

El proceso trata de recuperar los datos de las fuentes de información y alimentar el DWH o Data Mart; el proceso de extracción, transformación y cargado (ETL) utiliza desde el 60% hasta el 80% del tiempo de un proyecto de Business Intelligence, por lo que es un proceso clave.

Según Cano Giner (2007) el cual afirma que “el proceso ETL es necesario para acceder a los datos de las fuentes de información al DWH, se divide en 5 subprocesos” (p. 104):

Extracción:

Este proceso recupera los datos físicamente de las distintas fuentes de información. La extracción de los datos se puede realizar de forma manual o utilizando herramientas de ETL, Estas herramientas especializadas han sido diseñadas para llevar a cabo esta función y nos permite visualizar el proceso de detectar los errores durante la carga. Cano Giner (2007)

“El principal objetivo de la extracción es extraer tan solo aquellos datos de los sistemas transaccionales que son indispensables.” Cano Giner (2007)

Limpieza:

“Este proceso recupera los datos y comprueba su calidad, elimina los duplicados y, cuando es posible, corrige los valores erróneos y completa los vacíos, es decir, se transforman los datos para reducir los errores de carga.” Cano Giner (2007)

“Los sistemas transaccionales contienen datos que no han sido depurados y que deben ser limpiados. Las herramientas ETL tienen funcionalidades de limpieza de datos. Si no llevamos a cabo este subproceso, crearíamos dudas al mostrar los resultados.” Cano Giner (2007)

La limpieza de datos se divide en distintas etapas:

Depurar los valores: este proceso localiza e identifica los elementos individuales de información en las fuentes de datos y los aísla en los ficheros destino. Por ejemplo: separa el nombre completo en nombre, primer apellido y segundo apellido.

Corregir: este proceso corrige los valores individuales de los atributos usando algoritmos de corrección y fuentes de datos externas. Por ejemplo: comprueba una dirección y el código postal.

Estandarizar: este proceso aplica rutinas de conversión para transformar valores en formatos definidos aplicando procedimientos de estandarización. Por ejemplo: Sr, Sra.

Relacionar: este proceso busca y relaciona los valores de los registros, corrigiéndolos y estandarizándolos, basándose en reglas de negocio para eliminar duplicados.

Por ejemplo: identificar nombres y direcciones similares.

Consolidar: este proceso analiza e identifica relaciones entre registros relacionados y los junta en una sola presentación.

Transformación:

“Este proceso recupera los datos limpios y de alta calidad y los estructura y suma en los distintos modelos de análisis. El resultado de este proceso es la obtención de datos limpios, consistentes, sumados y útiles.” Cano Giner (2007)

“Transformamos los datos de acuerdo con las reglas de negocio y los estándares que han sido establecidos. La transformación incluye: cambios de formato, sustitución de códigos, valores derivados y agregados.” Cano Giner (2007)

“Los agregados, por ejemplo, la suma de las ventas, normalmente se pre calculan y se almacenan para conseguir mayor rendimiento cuando lanzamos las consultas al DWH.” Cano Giner (2007)

Integración: Este proceso valida que los datos que cargamos en el DWH son consistentes con las definiciones y formatos; además los integra en los distintos modelos de las distintas áreas de negocio que hemos definido en el mismo, estos procesos pueden ser complejos.

La última etapa es la integración en el DWH: es el momento en el que cargamos los datos, comprobamos por ejemplo si las ventas totales que hemos cargado coinciden con la información que residía en nuestro sistema transaccional.

Actualización:

Este proceso es el que nos permite añadir los nuevos datos al DWH y determina cada que tiempo haremos nuevas cargas de datos.

2.2.4 MySQL WorkBench

Es un sistema de administración de bases de datos relacionales, que puede realizar muchas tareas simultáneamente con otras tareas y es accesible a muchos usuarios cuando se necesita información que se encuentra almacenada en este gestor de base de datos.

Una base de datos es una recopilación organizada de datos. La información que puede almacenar una base de datos puede ser tan simple como la de una agenda, un contador, o un libro de visitas, o tan grande como la de una tienda en línea, un sistema de noticias, un portal, o la información generada por una red corporativa. Para manipular los datos almacenados en una base de datos, se necesita un sistema de administración de bases de datos, tal como MySQL.

MySQL se ejecuta en prácticamente todas las plataformas, incluyendo *Linux*, *UNIX* y Windows. A pesar de que se puede utilizar en una amplia gama de aplicaciones, MySQL se asocia más con las aplicaciones basadas en la web y la publicación en línea y es un componente importante de una pila empresarial de código abierto llamado LAMP. LAMP es una plataforma de desarrollo web que utiliza Linux como sistema operativo, Apache como servidor web, MySQL

como sistema de gestión de base de datos relacional y PHP como lenguaje de programación orientado a objetos (a veces, Perl o Python se utiliza en lugar de PHP). (ABC, 2007)

Características de MySQL:

- Aprovecha la potencia de sistemas multiprocesador, gracias a su implementación multihilo.
- Soporta gran cantidad de tipos de datos para las columnas.
- Dispone de API's en gran cantidad de lenguajes (C, C++, Java, PHP, etc.).
- Gran portabilidad entre sistemas.
- Soporta hasta 32 índices por tabla.
- Gestión de usuarios y contraseñas, manteniendo un muy buen nivel de seguridad en los datos.
- Condición de open source de MySQL hace que la utilización sea gratuita y se puede modificar con total libertad.
- Se puede descargar su código fuente. Esto ha favorecido muy positivamente en su desarrollo y continuas actualizaciones.
- Es una de las herramientas más utilizadas por los programadores orientados a Internet.

- Infinidad de librerías y otras herramientas que permiten su uso a través de gran cantidad de lenguajes de programación.
- MYSQL, es el manejador de base de datos considerado como el más rápido de Internet.
- Gran rapidez y facilidad de uso.
- Infinidad de librerías y otras herramientas que permiten su uso a través de gran cantidad de lenguajes de programación.
- Fácil instalación y configuración.

Ventajas

- MySQL software es Open Source.
- Velocidad al realizar las operaciones, lo que le hace uno de los gestores con mejor rendimiento.
- Bajo costo en requerimientos para la elaboración de bases de datos, ya que debido a su bajo consumo puede ser ejecutado en una máquina con escasos recursos sin ningún problema.
- Facilidad de configuración e instalación.
- Soporta gran variedad de Sistemas Operativos.

- Baja probabilidad de corromper datos, incluso si los errores no se producen en el propio gestor, sino en el sistema en el que está.
- Su conectividad, velocidad, y seguridad hacen de MySQL Server altamente apropiado para acceder bases de datos en Internet.

2.2.5 Data Warehouse y Data Mart

Un DWH es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta. La creación de un DWH representa en la mayoría de las ocasiones el primer paso, desde el punto de vista técnico, para implantar una solución completa y fiable de Business Intelligence. (SINEXUS, 2018)

La ventaja principal de este tipo de bases de datos radica en las estructuras en las que se almacena la información (modelos de tablas en estrella, en copo de nieve, cubos relacionales, etc.) Este tipo de persistencia de la información es homogénea y fiable, y permite la consulta y el tratamiento jerarquizado de la misma (siempre en un entorno diferente a los sistemas operacionales. (SINEXUS, 2018)

Características de un Data Warehouse

Según Inmon (2003) afirma que “las características que debe cumplir un Data Warehouse son:”

Orientada a un área: significa que cada fracción del DWH está diseñada para dar solución a un problema de negocio. Por ejemplo: para comprender los hábitos de compras de los clientes de una entidad.

Integrado: La información tiene que ser transformada en medidas, códigos y formatos comunes para que sea útil. Por ejemplo: la moneda en la que están declarados los importes.

Indexado en el tiempo: Es decir que se mantiene la información histórica y se almacena en determinadas unidades de tiempo. Por ejemplo: las ventas en un periodo (horas, días, semanas, meses, trimestres, bimestres, años).

- **No volátil:** Quiere decir que los usuarios no la mantienen, como lo harían en los entornos transaccionales. No se va actualizando incesantemente, sino periódicamente, de forma preestablecida.
- **Data Mart:** reúne información de un número restringido de áreas, por ejemplo, pueden ser marketing y ventas o producción. Comúnmente se definen para responder a usos muy concretos.

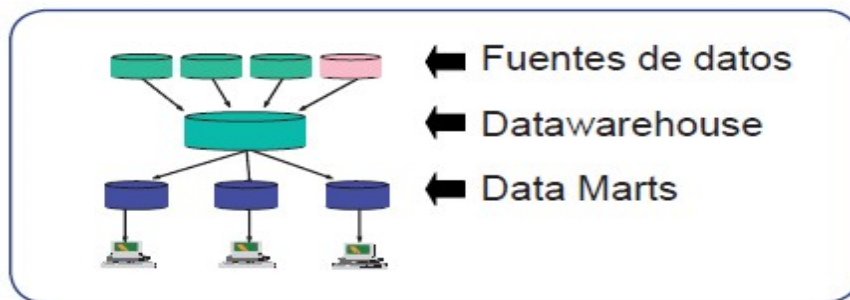


Figura 5 :Tabla Data Warehouse.

Fuente: Business Intelligence Competir con información (2007).

Estructura de un Data Warehouse

“En la estructura de un Data Warehouse encontraremos 4 niveles de esquematización los cuales forman la metadata, estos niveles se diferencian por el nivel de síntesis o depuración de información requerida por la empresa que lo usa” (Pacheco, 2015) Y son:

- **Detalle de datos actuales:** En gran parte, la importancia radica en el detalle de los datos actuales, debido a que los datos evidencian las ocurrencias más recientes, las cuales son de interés. Son voluminosos, ya que se encuentran almacenados al más bajo nivel de granularidad (no están procesados). Por lo general se almacenan en discos, los cuales se tienen fácil acceso, aunque su administración sea costosa y compleja.
- **Detalle de datos antiguos:** Es aquella que se almacena sobre alguna forma de almacenamiento masivo. No es frecuentemente accedida y se almacena a un nivel de detalle, consistente con los datos detallados actuales.

- Datos ligeramente resumidos: Es aquella que proviene desde un bajo nivel de detalle encontrado al nivel de detalle actual. Este nivel del DWH casi siempre se almacena en disco. Los puntos en los que se basa el diseñador para construirlo son:
 - Que la unidad de tiempo se encuentre sobre la esquematización hecha.
 - Qué contenidos (atributos) tendrá la data ligeramente resumida.
- Datos completamente resumidos: El siguiente nivel de datos encontrado en el Data Warehouse es el de los datos completamente resumidos. Estos datos son compactos y fácilmente accesibles por lo general son indicadores que son usados con más frecuencia para el análisis gerencial. (Pacheco, 2015)

Arquitectura de un Data Warehouse

Es una forma de mostrar la estructura global de los datos, la comunicación, los procesos y la presentación al usuario final. La arquitectura está formada por las siguientes partes interconectadas:

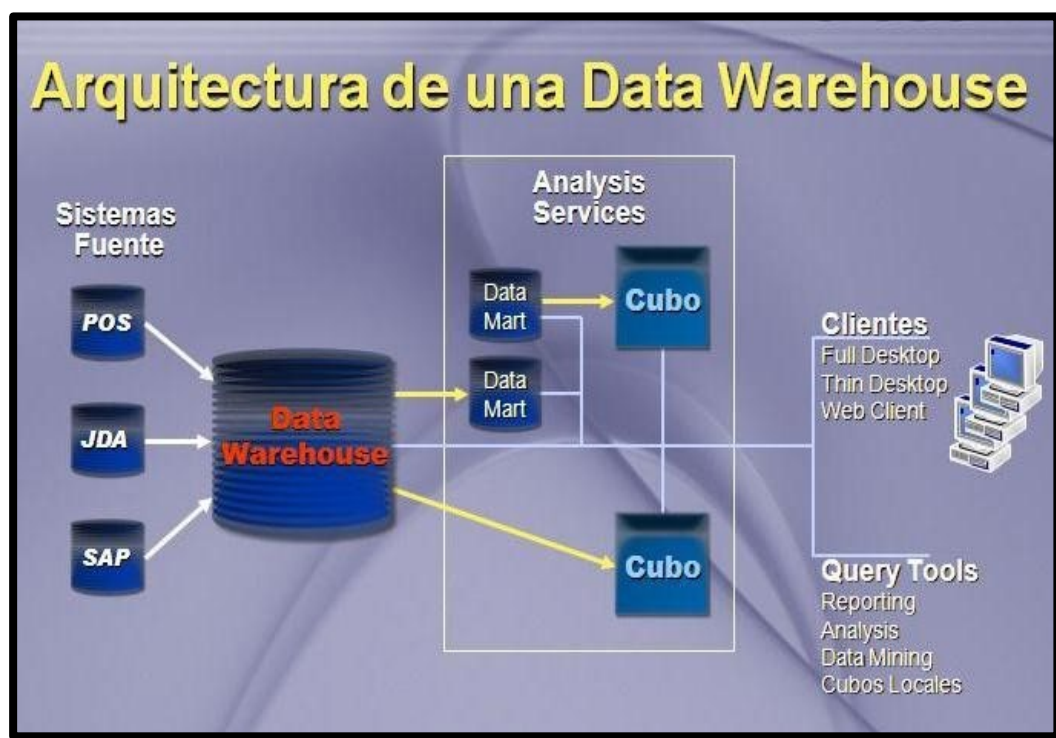


Figura 6 :Arquitectura del Data Warehouse (Srl, 1997)

Fuente: Assert Solutions srl. (1997)

Data Mart

Una data mart es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por disponer la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento (Yalan, 2013).

Un DM es una versión especial de un DWH; una base de datos departamental, especializada en el almacenamiento de los datos de un área del negocio específica. Se caracteriza por disponer de una estructura de datos pensada para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento (Arencibia M., 2016).

Características del Data Mart

- Orientado a un departamento dentro de una empresa
- Implementado para solución a problemas inmediatos
- Cubre necesidades específicas de negocio
- Asegura la consistencia de datos
- Respuestas a corto plazo a pedido de usuarios

Diferencias entre Data Warehouse y Data Mart

Data warehouse:

Se trata del lugar donde toda la data de una compañía es almacenada. Consiste en un sistema computarizado con una gran capacidad de almacenamiento, esencial para reunir y organizar la información proveniente de los distintos departamentos de la organización.

Data Mart:

Esta herramienta se ocupa de almacenar información de un departamento o grupo de trabajo específico. Funciona como una aplicación del DWH o una alternativa para empresas medianas que no pueden afrontar los costos de implementar un sistema tan amplio de almacenamiento de data. Las Data Marts pueden ser dependientes o independientes del DWH.

	Datawarehouse	Datamart
Alcance	Construido para satisfacer las necesidades de información de toda la organización	Construido para satisfacer las necesidades de un área de negocios específica
Objetivo	Diseñado para optimizar la integración y la administración de los datos fuente	Diseñado para optimizar la entrega de información de soporte a 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 esta 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

Figura 7 :Ejemplo de Modelo Estrella.

Fuente: Yalan, 2013

Un DM está pensado para cubrir las necesidades de un grupo de trabajo o de un determinado departamento dentro de la organización, en cambio, el ámbito de un DWH es la organización en su conjunto. El costo del uso de un DM es inferior al de aplicar un DWH y con llevan a un menor tiempo de construcción y puesta en marcha (Arencibia M., 2016).

2.2.6 Modelo Dimensional

Modelo Estrella

El modelo estrella es el más simple en estructura. Cuenta con una tabla principal de “hechos” y distintas “dimensiones”, incluyendo una dimensión de “tiempo”.

“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.” (Fabrizzio, 2010)

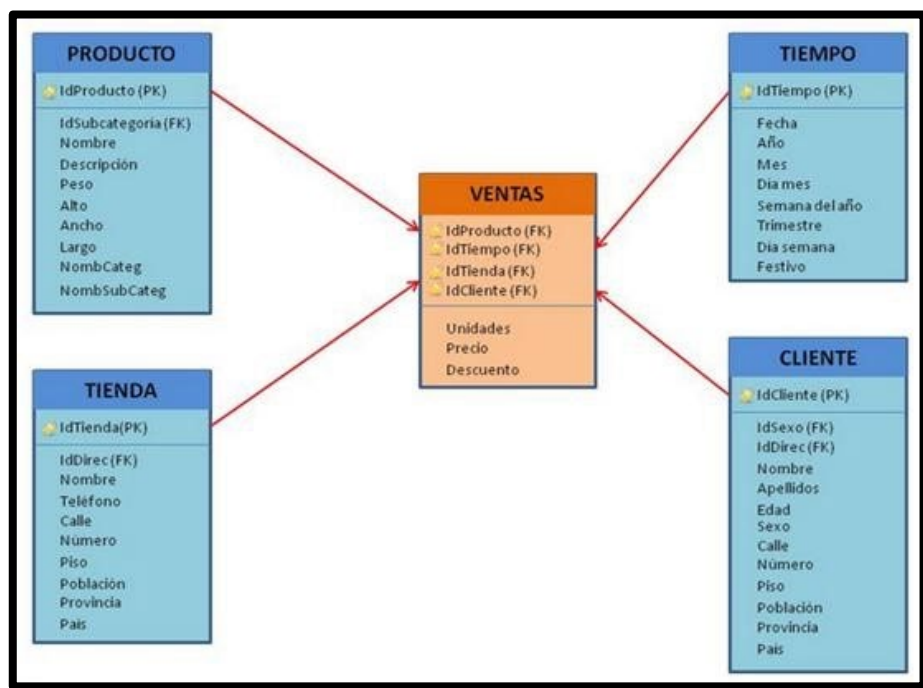


Figura 8 :Ejemplo de Modelo Estrella.

Fuente: Fabrizzio, 2010

Dimensiones

También llamada entidad, característica, dato maestro, indicador, etc. Son los conjuntos de datos que posibilitan identificar quién, cuándo o dónde se genera una operación de negocio. Por ejemplo: compradores, proveedores, tiempo y lugar son dimensiones características .

La asociación de dimensiones según una correspondencia de dependencia lógica, se le dice jerarquías. Tal es el caso del tiempo: Año – Mes – Día.

Una dimensión puede ser creada para usarse en un cubo individual o en múltiples cubos. Una dimensión creada para un cubo individual, es llamada dimensión privada. Por el contrario, si esta puede ser usada por múltiples cubos, se le llama dimensión compartida. Estas podrán ser usadas dentro de todo cubo, en la base de datos, así se optimiza el tiempo y se evita el andar duplicando dimensiones privadas. (Martines, 2015)

“Las dimensiones compartidas, también habilitan la estandarización de las métricas de negocios entre cubos. Por ejemplo, el estandarizar las dimensiones compartidas para el tiempo y localización geográfica, aseguran que los datos analizados, desde diferentes cubos, estén organizados similarmente.” (Martines, 2015)

Tabla de hechos

Una tabla de hechos es la tabla central de un esquema dimensional (en estrella o en copo de nieve) y contiene los valores de las medidas de negocio. Cada medida se toma mediante la intersección de las dimensiones que la definen, dichas dimensiones estarán reflejadas en sus correspondientes tablas de dimensiones que rodearán la tabla de hechos y estarán relacionadas con ella. (Creative Commons Attribution, 2016)

Existen diversos tipos de tablas de hechos las cuales se hace referencia a continuación:

- Tabla de hechos de transacciones: constituyen eventos que acontecen en un determinado espacio-tiempo. Se identifican por permitir analizar los datos con el máximo detalle.
- Tablas de hechos menores/Tablas de coberturas: Son tablas que no poseen medidas y tiene sentido dado que representan el hecho de que el evento suceda.
- Normalmente se añaden contadores a estas tablas para facilitar las consultas SQL.
- Tablas de hechos instantánea periódica: Son tablas de hecho utilizadas para acumular información de forma constante a intervalos de tiempo regulares.

- Tablas de hechos instantáneas de acumulación: “representan el ciclo de vida completo de una actividad o proceso, que tiene un principio y final. Se caracterizan por presentar múltiples dimensiones que relacionadas con los eventos presentes en un proceso.” (Datawarehouse, 2016)

Modelo Copo de Nieve

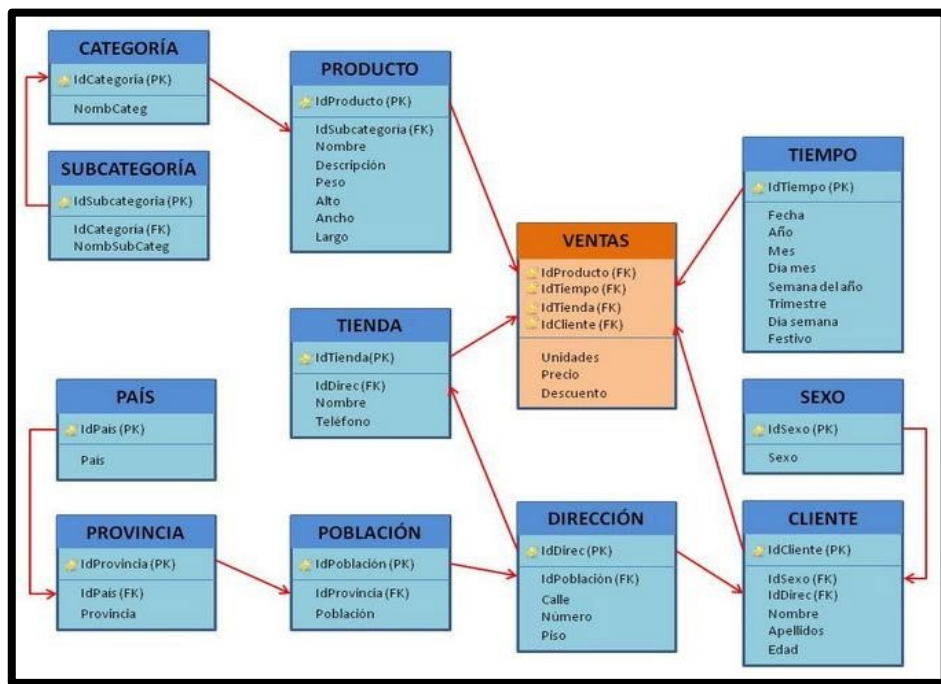


Figura 9 :Ejemplo de modelo copo de nieve

Fuente: Fabrizzio, 2010

El modelo copo de nieve (Figura 8) es una variación o derivación del modelo estrella. En este modelo la tabla de hechos deja de ser la única relacionada con otras tablas ya que existen otras tablas que se relacionan con las

dimensiones y que no tienen relación directa con la tabla de hechos. El modelo fue concebido para facilitar el mantenimiento de las dimensiones, sin embargo, esto hace que se vinculen más tablas a las secuencias SQL, haciendo la extracción de datos más difícil, así como vuelve compleja la tarea de mantener el modelo. (Fabrizzio, 2010)

2.2.7 Modelo multidimensionales

“Los cubos son una estructura de datos organizada mediante jerarquías. Cada indicador se puede evaluar en cualquiera de los niveles de las jerarquías. Así, por ejemplo, pueden obtener las "ventas" a nivel diario, o mensual, para un cliente, o una provincia.” (Urquizu, 2015)

El uso de cubos OLAP tiene dos ventajas fundamentales:

- Facilidad de uso: Una vez construido el cubo, el usuario de negocio puede consultarlo con simpleza, inclusive si se trata de un usuario con escasos o nulos conocimientos técnicos. La estructura jerárquica es sumamente sencilla de entender para la mente humana, y si ésta concuerda con el modelo de negocio, los resultados suelen ser espectaculares, debido a que el cubo se convierte en una gran "tabla dinámica" que el usuario puede explorar en cualquier momento.
- Rapidez de respuesta. Regularmente, el cubo posee pre-calculados las distintas agregaciones, por lo que los tiempos de respuesta son cortos. Si el

cubo correctamente diseñado, será igual de rápido consultar las compras de una ciudad, o las compras de todo el país, o incluso el total de compras de la empresa.

2.2.8 Software libre vs Software propietario

Software Libre

El movimiento de software libre encabezado por Richard Stallman, fundador de la *Free Software Foundation* (FSF) tiene como filosofía mantener público el código fuente de un programa de computador, de tal manera que cualquier persona pueda conocerlo, modificarlo, adaptarlo a sus necesidades específicas, copiarlo, etc.

Según manifiesta Carranza (2004), “el principio básico del Proyecto GNU y de la Free Software Foundation es que el código fuente es fundamental para fomentar la informática y que la libre disponibilidad del código fuente es verdaderamente necesaria para que la innovación continúe”

Mientras tanto, Pino (2014) en su artículo Protección al Software expone algunas características muy importantes acerca del software libre que son las siguientes: “El software libre opera a través de una licencia pública general (GPL) que puede resumirse en 5 puntos: (p.18). (1) El código fuente debe ser revelado. (2) El uso y distribución del código fuente son libres. (3) El código fuente puede ser libremente modificado y copiado. (4) Las modificaciones y mejoras que se

introduzcan en el código fuente son también de uso libre (5) Los derechos sobre el código fuente no pueden ser restringidos por persona alguna”.

Software Propietario

Mientras el software libre es gratuito en algunas ocasiones, sin restricciones y con libertad para su uso, el software propietario es pagado y con algunas limitaciones en cuanto a sus funcionalidades y libertades para usarlo. Así mismo la anterior es una apreciación o definición en términos generales, por ende, vale la pena explicar más detallada y concisamente las características principales del software propietario con el fin de esclarecer su concepto y entender las diferencias puntuales con el software libre.

Juárez, Herrera, y Sánchez (2006) en su artículo “Software libre vs software propietario: ventajas y desventajas” manifiestan que el software no libre o también llamado software privativo “se refiere a cualquier programa informático en el que los usuarios tienen limitadas las posibilidades de usarlo, modificarlo o redistribuirlo (con o sin modificaciones), cuyo código no esté disponible o el acceso a éste se encuentre restringido”

La opinión de Carranza (2004) con respecto al software propietario indica que el software propietario “es de dominio privado, porque una persona tiene la titularidad de los derechos de autor y goza de un derecho exclusivo respecto de su

utilización. Les niega a otras personas el acceso al código fuente del software y el derecho a copiarlo o estudiarlo” (p.103).

Comparación de software

“Cuadro de resumen con los criterios más importantes y sus respectivos pesos. Pentaho Business Analytics vs QlikView.” (Jácome, 2015)

Tabla 1
Comparativa herramientas de Pentaho

Criterios de comparación	Herramientas		
	Puntaje	Pentaho	QlikView
Usabilidad	30	20	24
Curva de aprendizaje	3	1	3
Apariencia estética	3	2	3
Nivel de complejidad en la manipulación de datos	6	5	3
Compatibilidad con distintos sistemas operativos	5	4	2
Flexibilidad	4	3	4
Popularidad y uso en las empresas	6	3	6
Tiempos de respuesta a nivel de tareas	3	2	3
Generación de reportes	30	21	22
Esquemas de selección e interacción en una sola vista	7	3	7
Análisis de contenido con metadatos	5	5	1
Manejo de grandes volúmenes de información	7	7	3
Tiempos de respuesta	5	3	5
Manejo de asociaciones de datos	6	3	6
Funcionalidad	40	32	19
BI Embebido	6	6	1
Estilos tradicionales de análisis	4	3	4
Tableros analíticos y contenidos	4	4	4
Plataforma interna de integración	8	8	1
Gestión de Metadatos	7	7	1
Colaboración e Integración social	4	2	4
Móvil	3	1	3
Despliegue en la nube	4	1	1
Total	100	73	65

Fuente: Jácome, 2015

La herramienta de software libre analizada Pentaho Business Analytics se adapta mejor al contexto de las empresas de comercialización masiva particularmente dentro del enfoque de *Big Data* debido a que posee un propio módulo de integración de datos que permite dar un mejor apoyo y soporte a dicho enfoque en lo que se refiere a procesos, análisis y manipulación de diversas fuentes de datos.(Jácome, 2015)

Dicho estudio nos ayudó a concluir que para este trabajo usaremos software libre, el cual por lo expuesto anteriormente es Pentaho Business Intelligence el cual se detallara en los siguientes ítems.

2.2.9 Pentaho Business Intelligence

Es una suite de BI abierta, que cuenta con la incorporación de las principales herramientas del mercado Open Source. Al día de hoy es la más completa. Cuenta con una gran sociedad de desarrollo, que realiza constantes mejoras y extensiones en la plataforma.

Pentaho se compone de un entorno, al que los usuarios pueden acceder vía web, de forma segura. Cada usuario visualiza todos los elementos habilitados para su perfil, que incluirán informes, análisis OLAP y cuadros de mando con indicadores y tablas. Usted podrá generar nuevas vistas de análisis y nuevos informes y guardarlos asociados a su perfil, para asesorarse más tarde, así como exportarlos a Excel, PDF o realizar su impresión en papel.

2.2.7.1 Pentaho Data Integration

Es el elemento de Pentaho encargado de la extracción, transformación y procesos de carga (ETL). Aunque las herramientas de ETL se utilizan con mayor frecuencia en contexto de DWH, Pentaho Data Integration también puede ser utilizado para otros fines:

- Migración de datos entre aplicaciones o bases de datos
- Exportar datos desde una base de datos hacia archivos planos
- Cargar datos de forma masiva en bases de datos
- Limpieza de datos
- Integración de aplicaciones

Pentaho Data Integration es fácil de usar. Cada proceso se crea con una herramienta gráfica donde se detalla qué hacer sin necesidad de escribir código para indicar el cómo; debido a esto, se podría decir que es una herramienta orientada a metadatos.

Pentaho Data Integration se puede usar como una aplicación independiente, o puede ser utilizado como fracción de la Suite de Pentaho. Como una herramienta ETL.

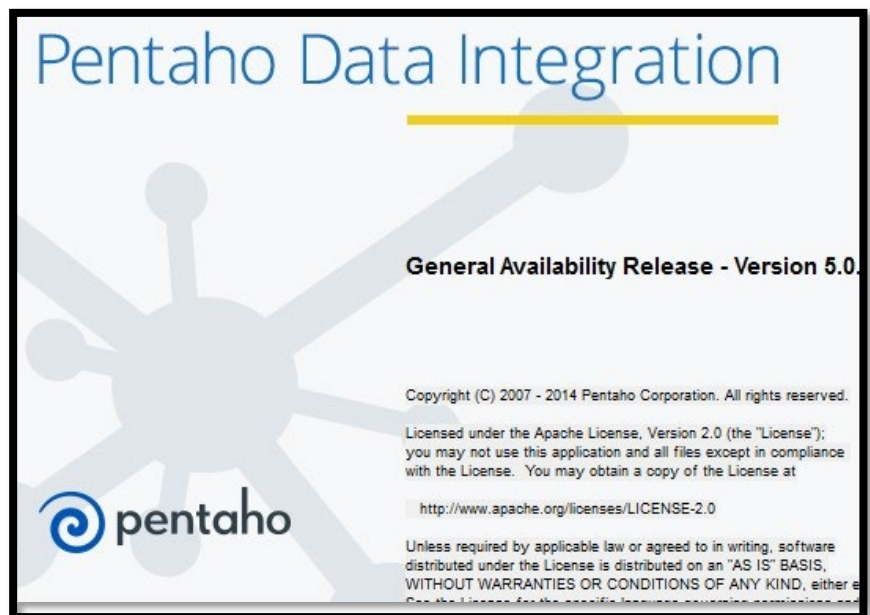


Figura 10 :Pentaho Data Integration

Fuente: Elaboración Propia

2.2.7.2.Pentaho Schema Workbench

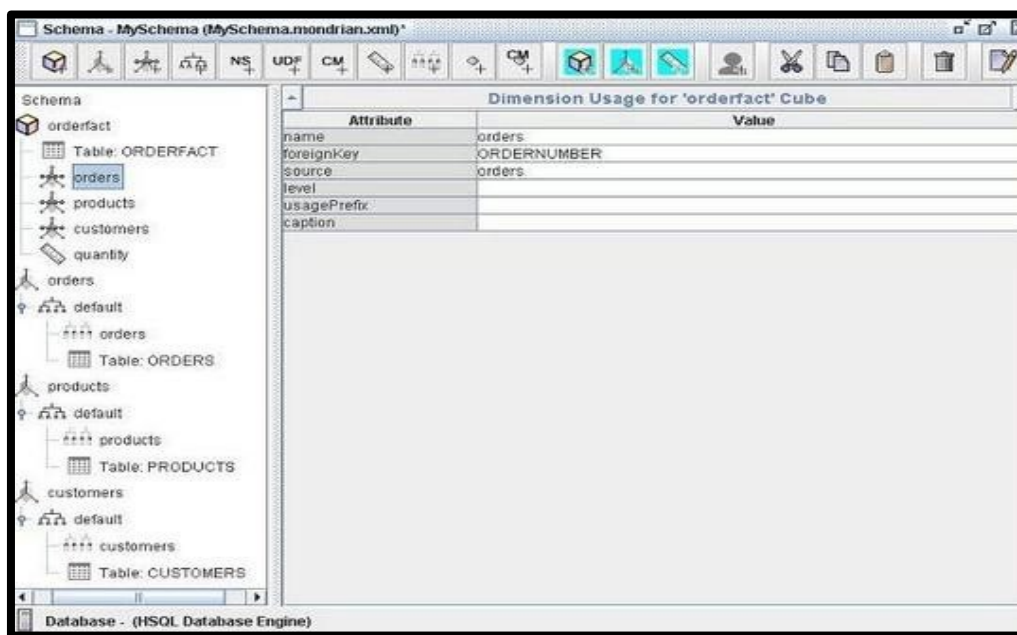


Figura 11 :Pentaho Schema Workbench

Fuente: Elaboración propia

Se define como una interfaz de bosquejo que permite crear y probar esquemas de cubos OLAP. El motor procesa instancias MDX con esquemas relacionales OLAP. Donde estos archivos son esquemas de metadatos XML que se desarrollan en una estructura específica utilizada por el motor de Mondrian. Dichos modelos XML podrían ser considerados estructuras de cubo que utilizan tablas hechos y dimensiones que están en el DWH o en los Data Marts.

Desarrollado originalmente por algunas organizaciones para implementar fácilmente esquemas y utilizarlo como el motor de análisis. Pentaho ha actualizado el entorno de trabajo de workbench adicionando estabilidad, validación y Pentaho BI Server para trabajar con Pentaho BI Suite. Un esquema diseñado en workbench realmente es un documento XML que utiliza el motor de análisis de mondrian.

2.2.7.3 Pentaho Report Designer

Es un aglomerado de herramientas para el desarrollo de informes, con Pentaho *Reporting* los usuarios pueden transformar datos en información significativa de acuerdo de su necesidad. Puede crear reportes en HTML, Excel, PDF, texto o informes impresos.

El desarrollo de esta herramienta es motivado con el objetivo de crear un sistema flexible y sencillo de generar informes. El motor de informes le proporciona flexibilidad para crear informes que se adecuen a sus datos, ya que casi todos los atributos se pueden calcular durante la creación del informe. Los reportes pueden

comprender datos de prácticamente cualquier origen de datos debido a la gran selección de origen de datos, incluyendo bases de datos, fuentes de datos OLAP e incluso la herramienta ETL data integration de Pentaho.

Pentaho Report Designer es la herramienta primordial de diseño para implementar la estructura del informe. Su interfaz de usuario admite crear sin esfuerzo el flujo de datos en el informe y determinar el aspecto visual de éste.

Pentaho Report Designer brinda un acceso completo a todas las opciones de configuración del Pentaho Reporting Engine. Su alto número de posibilidades y flexibilidad extrema puede hacer que sea agobiante para los usuarios novatos. Esta herramienta se encuentra dirigida a usuarios avanzados técnicamente calificados.

Pentaho Report Designer puede tener un uso como herramienta de informes de escritorio mediante la realización de sus informes de forma local. También puede publicar sus informes finalizados en Pentaho BI- servidor para que su informe está disponible para los demás.

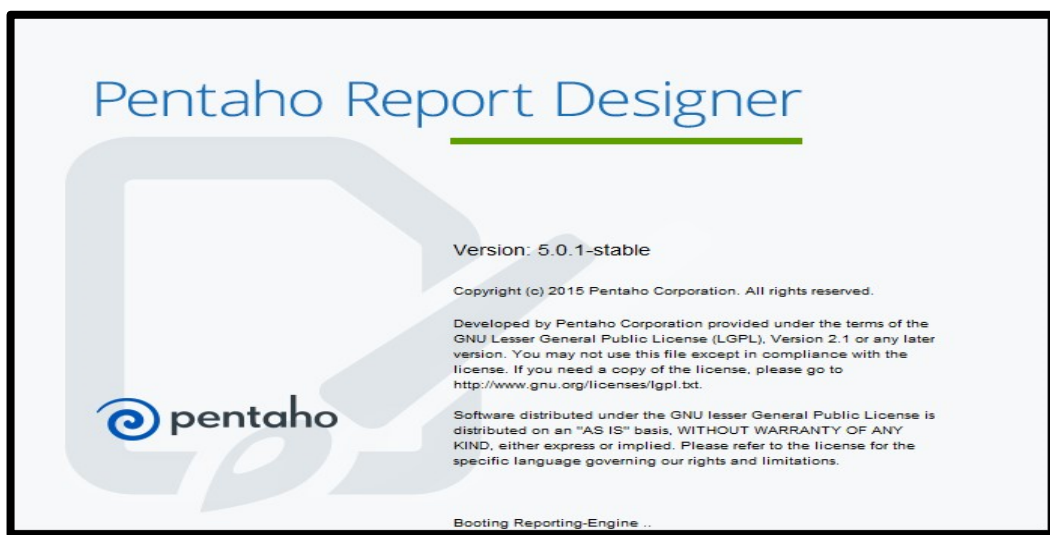


Figura 12 :Pentaho Report Designer

Fuente: Elaboración propia

2.2.10 Rendimiento académico

Según la definición de Jiménez (2000), se entiende por rendimiento académico al “nivel de conocimientos demostrado en un área o materia comparado con la norma de edad y nivel académico.” Por esa razón, se considera que el rendimiento del alumno actualmente es entendido a partir de los procesos de evaluación y exámenes, sin embargo, la simple medición y/o calificación del nivel alcanzado por los alumnos, no provee por sí mismo, todas las pautas necesarias para comprender la relación entre el estudiante, su rendimiento real, y su rendimiento potencial. (Velez y E., 1994) Es decir, el rendimiento de un estudiante, muchas veces se ve impactado negativamente por cuestiones externas a sus propias disposiciones intelectuales.

El rendimiento académico es la expresión de capacidades y de características psicológicas del estudiante desarrolladas y actualizadas a través del proceso de enseñanza-aprendizaje que le posibilita obtener un nivel de funcionamiento y logros académicos a lo largo de un período, año o semestre, que se sintetiza en un calificativo final (cuantitativo en la mayoría de los casos) evaluador del nivel alcanzado.(Chadwick, 1979)

En consonancia con esa caracterización y en directa relación con los propósitos de la investigación, es necesario conceptualizar el rendimiento académico. Para ello se requiere previamente considerar dos aspectos básicos del rendimiento: el proceso de aprendizaje y la evaluación de dicho aprendizaje. El proceso de aprendizaje no será abordado en este estudio. Sobre la evaluación académica hay una variedad de postulados que pueden agruparse en dos categorías: aquellos dirigidos a la consecución de un valor numérico (u otro) y aquellos encaminados a propiciar la comprensión en términos de utilizar también la evaluación como parte del aprendizaje. En el presente trabajo interesa la primera categoría, que se expresa en los calificativos universitarios. Las calificaciones son las notas o expresiones cuantitativas o cualitativas con las que se valora o mide el nivel del rendimiento académico en los estudiantes. Las calificaciones son el resultado de los exámenes o de la evaluación continua a que se ven sometidos los estudiantes. Medir o evaluar los rendimientos es una tarea compleja que exige del docente obrar con la máxima objetividad y precisión.

En el sistema educativo peruano, en especial en las universidades, la mayor parte de las calificaciones se basan en el sistema decimal, es decir de 0 a 20. Sistema en el cual el puntaje obtenido se traduce a la categorización del logro del aprendizaje

2.2.11 Indicadores del rendimiento académico

Es una medida que nos permite ir observando el avance en el cumplimiento del desarrollo de capacidades que proporciona un medio sencillo y fiable para medir logros, reflejar los cambios vinculados con una intervención o ayudar a evaluar los resultados

Indicador	2010-1	2010-2	2011-1	2011-2	2012-1	2012-2	2013-1
Registro de asignaturas inscritas	44039	43352	49566	42167	47931	47673	49918
Aprobados	36814	36713	42701	37643	40605	40308	42588
No aprobados	7225	6639	6865	4524	7326	7365	7330
Índice de Reprobación -Registros	16,4	15,3	13,9	10,7	15,3	15,4	14,7
Créditos Aprobados	122575	122840	143297	126906	137282	136522	143140
Créditos No aprobados	25986	23880	24507	15865	26155	26538	25419
Índice de Reprobación –Créditos	17,5	16,3	14,6	11,1	16,0	16,3	15,1
Eficiencia (%)						78,9	76,8
Repitentes	3496	3386	5455	5009	3458	6187	7421
Índice de repitencia	7,9	7,8	11,0	11,9	7,2	13,0	14,9
Estudiantes matriculados	10003	9914	10206	10246	10398	10151	10176
Registros/Estudiante	4,4	4,4	4,9	4,1	4,6	4,7	4,9
Créditos/Estudiante	14,9	14,8	16,4	13,9	15,7	16,1	16,6
Nota Promedio	3,59	3,61	3,63	3,76	3,64	3,63	3,70
Desviación	1,02	0,98	0,96	0,88	0,98	0,96	0,98
PAPA promedio	3,54	3,55	3,57	3,67	3,58	3,58	3,60
Desviación PAPA	0,60	0,55	0,53	0,51	0,54	0,52	0,53
Estudiantes Reprobados	827	703	654	487	719	631	634
Índice de estudiantes reprobados	8,3	7,1	6,4	4,8	6,9	6,2	6,2

Figura 13 :Indicadores del rendimiento académico.

Fuente: Universidad de Medellín – Oficina de planeación

2.2.12 Marco metodológico

Se realizará con la **metodología en cascada**. “Es un modelo lineal de diseño de software que emplea un proceso de diseño secuencial. El desarrollo fluye secuencialmente desde el punto inicial hasta el punto final, con varias etapas diferentes: planteamiento, iniciación, análisis, diseño, construcción, pruebas, implementación y mantenimiento.”(OBS Business School, 2018)

Ventajas de la metodología en cascada

El énfasis de la metodología en cascada se pone en la planificación de proyecto y, por tanto, antes de comenzar cualquier tipo de desarrollo es necesario que tanto la visión como el plan estén claros. Debido a que el método de cascada requiere un amplio esfuerzo de preparación previa, permite:

- Comenzar con el software con bastante rapidez.
- Estimar calendarios y presupuestos con mayor precisión.
- Lograr un nivel de satisfacción del cliente más elevado que otros enfoques, ya desde el principio.

La metodología en cascada supera algunas de las limitaciones de otros métodos como:

- Scrum: los procesos de desarrollo que siguen la metodología en cascada tienden a ser más seguro, ya que existe una firme orientación al plan. A diferencia de la metodología Scrum, donde el abandono de uno de los miembros del equipo puede suponer un grave problema, con la metodología en cascada no lo sería, ya que se dispone de una completa planificación y documentación que permite suplir este tipo de pérdidas. De esta forma, un nuevo diseñador puede fácilmente tomar la posición vacante siguiendo el plan de desarrollo sin ningún problema.

- **Ágil:** aunque altamente flexible, el enfoque ágil no tiene la estructura que tiene la metodología en cascada, lo que implica algunos inconvenientes. Los más significativos son los que tienen que ver con la dificultad para predecir los presupuestos a partir de líneas de tiempo. Y es que, sin una planificación completa, todos los aspectos quedan vagamente definidos y son susceptibles de generar confusión.

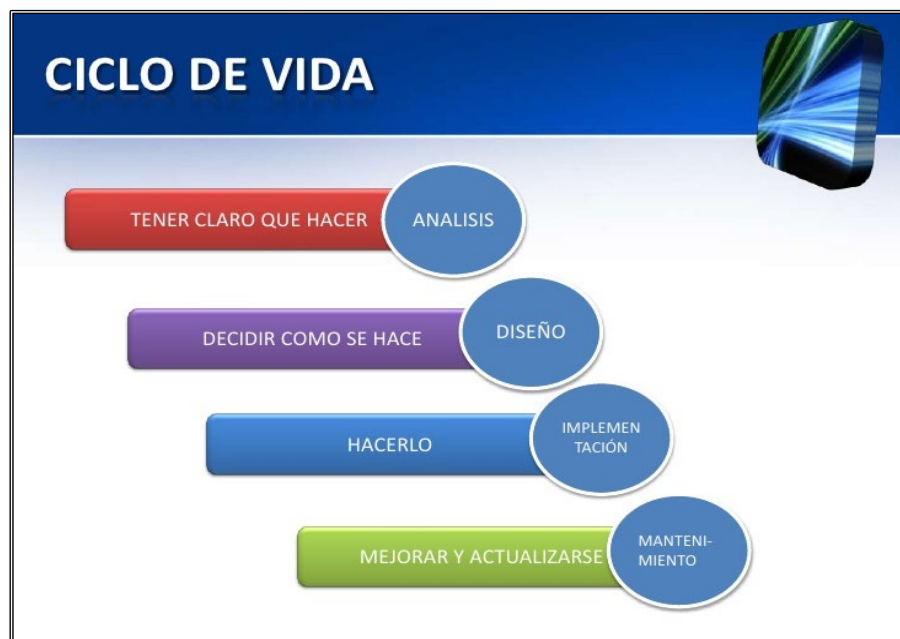


Figura 14 :Ciclo de vida de la metodología
Fuente: Stiven, D. Metodología en Cascada

2.3 Definición de términos

BI: Business Intelligence es un conjunto de procedimientos y métodos que transforman los datos de una organización en información significativa para su desarrollo estratégico, táctico y operativo.

Data Warehouse: Es la base de datos central donde se consolidan todos los datos procedentes de los diferentes sistemas fuente.

Sistemas Fuente: constituyen el origen de los datos que aportan a la DWH.

ETL: Procesos de extracción, transformación y carga son programas encargados de extraer los datos de los sistemas fuente, por medio canales por los que los datos de los fluyen hasta la DWH.

Dimensión: Tablas conectadas a la tabla de hechos en el modelo estrella.

Tabla De Hechos: La tabla principal de un diagrama multidimensional (esquema estrella o copo de nieve) que contiene los atributos de las medidas de negocio.

OLAP: Procesamiento analítico en línea es una solución usada en el campo de Business Intelligence con el objetivo de facilitar la consulta de grandes conjuntos de datos

Metadata: Se refiere a un grupo de datos que describen el contenido informativo de un objeto al que se denomina recurso.

Middleware: Significa lógica de intercambio de información, es un software que actúa entre el sistema operativo y las aplicaciones, permitiendo la comunicación entre ambas.

Data Mart: Es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica

CAPÍTULO III

MARCO METODOLÓGICO

3.1 Diseño experimental o no experimental

Esta investigación se basa en un Diseño No Experimental de corte transversal, podría definirse como “la investigación que se realiza sin manipular deliberadamente variables. Es decir, se trata de estudios en los que no hacemos variar en forma intencional las variables independientes para ver su efecto sobre otras variables. Lo que hacemos en la investigación no experimental es observar fenómenos tal como se dan en su contexto natural, para analizarlos”. (Sciences, 2009)

El diseño no experimental se divide tomando en cuenta el tiempo durante se recolectan los datos, estos son: diseño Transversal, donde se recolectan datos en un solo momento, en un tiempo único, su propósito es describir variables y su incidencia de interrelación en un momento dado, y el diseño Longitudinal, donde se recolectan datos a través del tiempo en puntos o periodos, para hacer inferencias respecto al cambio, sus determinantes y sus consecuencias..(Hernández, Fernández, y Baptista, 2014)

En este estudio se ocupará el diseño no experimental transversal, debido a que se recolectarán datos en un tiempo determinado, sin manipulación de datos.

Para la estadística se puede decir que “El objetivo final, cuando se emplea este método de análisis, es predecir o estimar el valor de una variable, que corresponde al valor dado de otra variable” (Wayne W., 2014). En esta investigación no se busca predecir ni estimar el valor de una variable por consiguiente los resultados no se muestran con estadística descriptiva.

3.2 Población y muestra

3.2.1 Población

Para el presente estudio, se tomará como población al total de registros académicos de la FAIN, del periodo 2014 al 2017.

Tabla 2

Registros académicos FAIN

Registros académicos DASA – FAIN - Periodo 2014 al 2017

Escuelas Profesionales	Alumnos	Periodo	Registros
5	2086	2013	12159
		2014	12249
		2015	13168
		2016	15081
		2017	14077

Fuente: Elaboración propia

3.2.2 Muestra

Muestreo por conveniencia

Creswell (2008) lo define como “un procedimiento de muestreo cuantitativo en el que el investigador selecciona a los participantes, ya que se están dispuestos y disponibles para ser estudiados”. “Muestreo accidental o muestreo por oportunidad” (Cohen, Manion, y Morrison, 2003). Consiste en la elección por métodos no aleatorios de una muestra cuyas características sean similares a las de la población objetivo.

En este tipo de muestreo la “representatividad” la determina el investigador de modo subjetivo, siendo este el mayor inconveniente del método ya que no podemos cuantificar la representatividad de la muestra. Presenta casi siempre sesgos y por tanto debe aplicarse únicamente cuando no existe alternativa. (Casal y Mateu, 2003)

Los individuos de una investigación concreta, son elegidos para el estudio sólo porque son más cómodos de reclutar y el investigador no está considerando las características de inclusión de los individuos que los hace representativos de toda la población.

En las distintas formas de investigación, sería ideal generalizar los resultados a toda la población, pero en casi todos los casos, la población es demasiado grande y resulta difícil incluir cada individuo. Siendo esta la razón por la que la mayoría de investigadores usan técnicas de muestreo, como el muestreo por conveniencia, que es la más común. Varios investigadores prefieren esta técnica, debido a que es rápida, sencilla, económica y, sobre todo, los individuos están disponibles.

Para este proyecto se tomará como muestra el registro académico de los alumnos de la Facultad de Ingeniería, comprendiendo los periodos desde el 2014 hasta el 2017.

Tabla 3

Registros académicos FAIN

Registros académicos DASA – FAIN - Periodo 2014 al 2017			
Escuelas Profesionales	Alumnos	Periodo	Registros
5	2086	2013	12159
		2014	12249
		2015	13168
		2016	15081
		2017	14077
		Total	66734

Fuente: Elaboración propia

3.3 Operacionalización de variables

Una variable es una “propiedad que tiene una variación que puede medirse u observarse” (Hernández et al., 2010, p.93) que es definida estrechamente a la hipótesis de la investigación (p.96). Por consecuencia, no es posible definir una variable para este trabajo de investigación en términos operacionales.

Como definición conceptual se considerará como variable de estudio al “Análisis multidimensional del rendimiento académico”.

Tabla 4
Operacionalización de variables

Variable	Dimensión	Indicador	Descripción	Instrumento	Unidad de medida
Análisis multidimensional del rendimiento académico	Granularidad	Ciclos o semestres	Permite establecer el nivel de detalle con el cual se visualizan los datos	Revisión documental	Cuantitativa
	Dimensiones	Escuela, Sexo, Materia, alumno, estado.	Datos que nos permiten filtrar, agrupar o seccionar la información	Revisión documental	Cuantitativa
	Aprobados	Cantidad de aprobados	Número total de aprobados	Revisión documental	Cuantitativa
	No-aprobados	Cantidad de desaprobados	Número total de desaprobados	Revisión documental	Cuantitativa
	Promedio de notas	Promedio de notas	Promedio de notas de los alumnos	Revisión documental	Cuantitativa

Fuente: Elaboración propia

3.4 Técnicas e instrumentos para recolección de datos

La **revisión documental** es la recopilación de datos mediante el análisis de documentos o fuentes secundarias. “Es trabajar con materiales ya elaborados de tipos secundarios determinan lógicamente las principales ventajas e inconvenientes. Complementará el estudio a través de la utilización de fuentes primarias y secundarias como textos, trabajos de grados, manuales, internet entre otros recursos.” (Castro, 2010)

Para la investigación se solicitó a la Dirección Académica de Actividades y Servicios Académicos (DASA), el registro académico de los alumnos de la Facultad de Ingeniería, desde el periodo 2014 al 2017, con el documento presentado en el Anexo B.

Matriz de datos de identificación de los documentos

En las columnas se colocan los datos que permitan identificar los documentos de acuerdo al estudio que se está realizando

Tabla 5

Matriz de categoría de la revisión documental

Nº	Documento	Referencia	Fuente	Año	Resumen
1	Solicitud registro de notas del sistema académico	Anexo B	Dirección Académica de Actividades y Servicios Académicos	2018	Con el presente documento Anexo B, se solicitó y recibió los datos que corresponden para desarrollar la investigación.

Nota: La matriz de categoría se utiliza luego de recibir la información correspondiendo, se utiliza para registrar el origen y datos relevantes de los documentos recogidos.

3.5 Procesamiento y análisis de datos

En el proceso cuantitativo primero se recolectan todos los datos y luego se analizan, mientras que en la investigación cualitativa no es así, sino que la recolección y el análisis ocurren prácticamente en paralelo; además, el análisis no es uniforme, ya que cada estudio requiere un esquema peculiar. Sin embargo, diversos autores hemos propuesto un análisis genérico o básico común a diversas investigaciones cualitativas, porque en nuestra experiencia, a la mayoría de los alumnos que se inician en la indagación cualitativa, sobre todo si el único enfoque que han utilizado es el cuantitativo, les cuesta trabajo comenzar su tarea analítica y se preguntan por dónde empezar y qué secuencia establecer.

En este apartado se propone un proceso de análisis que incorpora las concepciones de diversos teóricos de la metodología en el campo cualitativo, sin ser una camisa de fuerza. Cada estudiante, tutor o investigador podrá adoptarlo o no de acuerdo con las circunstancias y naturaleza de su estudio en particular.

En el análisis de los datos, la acción esencial consiste en que recibimos datos no estructurados, a los cuales nosotros les proporcionamos una estructura. Los datos son muy variados, pero en esencia consisten en observaciones del investigador y narraciones de los participantes:10 (a) visuales (fotografías, videos, pinturas, entre otras), (b) auditivas (grabaciones), (c) textos escritos (documentos, cartas, etc.) y (d) expresiones verbales y no verbales (como respuestas orales y gestos en una entrevista o grupo de enfoque), además de las narraciones del investigador (anotaciones o grabaciones en la bitácora de campo, ya sea una libreta o un dispositivo electrónico).

Los propósitos centrales del análisis cualitativo son: (1) explorar los datos, (2) imponerles una estructura (organizándolos en unidades y categorías), (3) describir las experiencias de los participantes según su óptica, lenguaje y expresiones; (4) descubrir los conceptos, categorías, temas y patrones presentes en los datos, así como sus vínculos, a fin de otorgarles sentido, interpretarlos y explicarlos en función del planteamiento del problema; (5) comprender en profundidad el contexto que rodea a los datos, (6) reconstruir hechos e historias, (7) vincular los resultados con el conocimiento disponible y (8) generar una teoría fundamentada en los datos.

El logro de tales propósitos es una labor paulatina. Asimismo, algunas características que definen la naturaleza del análisis cualitativo son las siguientes:

- El análisis es un proceso ecléctico (que concilia diversas perspectivas) y sistemático, más no rígido.
- Una fuente de datos importantísima que se agrega al análisis la constituyen las impresiones, percepciones, sentimientos y experiencias del investigador o investigadores.
- La interpretación que se haga de los datos puede diferir de la que podrían realizar otros investigadores, lo cual no significa que una interpretación sea mejor que otra, sino que cada quien posee su propia perspectiva. Esto, pese a que recientemente se han establecido ciertos acuerdos para sistematizar en mayor medida el análisis cualitativo (Lapadat, 2009).
- Como cualquier tipo de análisis, el cualitativo es sumamente contextual y no es un análisis “paso a paso”, sino que consiste en estudiar cada dato en sí mismo y en relación con los demás (“como armar un rompecabezas”).
- Es un camino con rumbo, pero no en “línea recta”, pues continuamente nos movemos; vamos y regresamos entre los primeros datos recolectados y los últimos, los interpretamos y les encontramos un significado, lo cual permite ampliar la base de datos conforme es necesario, hasta que construimos significados para el conjunto de los datos.

- “La interacción entre la recolección y el análisis nos permite mayor flexibilidad en la valoración de los datos y adaptabilidad cuando elaboramos las conclusiones”(Coleman, 2005). Debe insistirse: el análisis de los datos no es predeterminado, sino que es “prefigurado o coreografiado”. Se comienza a efectuar según un plan general, pero su desarrollo sufre modificaciones de acuerdo con los resultados. El análisis es moldeado por los datos (lo que los participantes o casos van revelando y lo que el investigador va descubriendo).
- “El investigador analiza cada dato, que por sí mismo tiene un valor” (Abma, 2004), y deduce similitudes y diferencias con otros datos
- “Los segmentos de datos o unidades son organizados en un sistema de categorías” (Li, 2007; Ritchie, Lewis, y Nicholls, 2013).
- “Los resultados del análisis son síntesis de “orden superior” que emergen en la forma de descripciones, expresiones, categorías, temas, patrones, hipótesis y teoría” (Alvarez-Gayou, 2003)

- “Existen diversos acercamientos al análisis cualitativo de acuerdo con el diseño o el marco referencial seleccionado. Entre estos acercamientos se encuentran varios, como etnografía, teoría fundamentada, fenomenología, feminismo, análisis del discurso, análisis conversacional, análisis semióticos y posestructurales.”(Alvarez-Gayou, 2003) Pero todos efectúan análisis temático.

3.5.1 Fuentes de datos

Son las fuentes donde se encuentran los datos pueden ser archivos transaccionales (digitales), documentos físicos los cuales se encuentran en la organización, en base a estos archivos se el desarrollo el DM:

Se recibió dos archivos transaccionales que alimentarán el DM. Uno que contiene toda la información de los estudiantes, carrera, notas, número de matrículas, cursos, materias, asistencia y notas, y el segundo que contiene un listado de los estudiantes. Este archivo fue facilitado por la oficina de DASA “Anexo B”

- **Archivo transaccional Notas**

Por el número de datos que contiene este archivo ha sido necesario la conversión del formato.

El primer archivo Excel ha sido transformado a un archivo plano separado por “;” con formato CVS, para su mejor tratamiento de la gran cantidad de registros que se tiene almacenado.

CORONEL GREGORIO ALBARRACIN LANCHIPA																
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Codigo	ApellidoPat	ApellidoMat	Nombre	FechaNacimi	IdSexo	Sexo	IdEstadoCivi	EstadoCivil	LugarNacimi	LugarResidei	LenguaMate	SituacionLab	CentroEstud	IdMateria	CodigoMateria
2	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			359	19.01104
3	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			356	19.01101
4	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			357	19.01102
5	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			358	19.01103
6	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			360	19.01105
7	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			361	19.01106
8	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			362	19.01107
9	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			4877	19.02108
10	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			4878	19.02109
11	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			4879	19.0211
12	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			4880	19.02111
13	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			4881	19.02112
14	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			4882	19.02113
15	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			4883	19.02114
16	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3147	A.5111
17	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3148	A.5121
18	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3149	A.5131
19	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3150	A.5141
20	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3151	A.5151
21	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3152	A.5161
22	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3153	A.5171
23	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3154	A.5182
24	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3155	A.5192
25	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3156	A.5202

Figura 15 :Archivo transaccional notas

Fuente: Elaboración propia

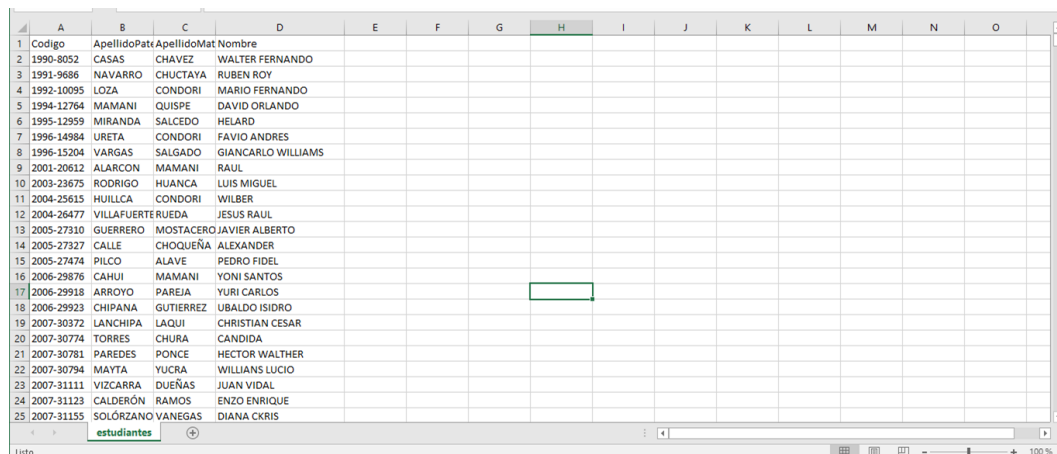
SIN EMPLEO																
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Codigo	ApellidoPat	ApellidoMat	Nombre	FechaNacimi	IdSexo	Sexo	IdEstadoCivi	EstadoCivil	LugarNacimi	LugarResidei	LenguaMate	SituacionLab	CentroEstud	IdMateria	CodigoMateri
2	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			359	19.01104
3	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			356	19.01101
4	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			357	19.01102
5	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			358	19.01103
6	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			360	19.01105
7	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			361	19.01106
8	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			362	19.01107
9	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			4877	19.02108
10	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			4878	19.02109
11	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			4879	19.0211
12	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			4880	19.02111
13	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			4881	19.02112
14	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			4882	19.02113
15	2012-36159	MACHACA	QUINTO	JOSEPH JOEL	13/03/1995	9	MASCULINO	2	SOLTERO	NINGUNO	CIUDAD NUE ESPAÑOL O (SIN EMPLEO	NO ESPECIFI			4883	19.02114
16	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3147	A.5111
17	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3148	A.5121
18	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3149	A.5131
19	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3150	A.5141
20	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3151	A.5151
21	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3152	A.5161
22	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3153	A.5171
23	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3154	A.5182
24	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3155	A.5192
25	2009-33316	LIMACHE	CONDORI	DANIEL ISMA	1/03/1992	9	MASCULINO	2	SOLTERO	TACNA	CORONEL GR ESPAÑOL O (SIN EMPLEO	CORONEL GR			3156	A.5202

Figura 16 :Archivo plano notas

Fuente: Elaboración propia

Notas.csv es un archivo plano separado por “;” con formato CVS, transformado para su mejor tratamiento, debido a la gran cantidad de registros que tiene almacenado.

- Archivo transaccional Estudiantes



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Codigo	ApellidoPat	ApellidoMat	Nombre											
2	1990-8052	CASAS	CHAVEZ	WALTER FERNANDO											
3	1991-9686	NAVARRO	CHUCTAYA	RUBEN ROY											
4	1992-10095	LOZA	CONDORI	MARIO FERNANDO											
5	1994-12764	MAMANI	QUISPE	DAVID ORLANDO											
6	1995-12959	MIRANDA	SALCEDO	HELARD											
7	1996-14984	URETA	CONDORI	FAVIO ANDRES											
8	1996-15204	VARGAS	SALGADO	GIANCARLO WILLIAMS											
9	2001-20612	ALARCON	MAMANI	RAUL											
10	2003-23675	RODRIGO	HUANCA	LUIS MIGUEL											
11	2004-25615	HUILLCA	CONDORI	WILBER											
12	2004-26477	VILLAFUERTE	RUEDA	JESUS RAUL											
13	2005-27310	GUERRERO	MOSTACERO	JAVIER ALBERTO											
14	2005-27327	CALLE	CHOQUEÑA	ALEXANDER											
15	2005-27474	PILCO	ALAVE	PEDRO FIDEL											
16	2006-29876	CAHUJ	MAMANI	YONI SANTOS											
17	2006-29918	ARROYO	PEREJA	YURI CARLOS											
18	2006-29923	CHIPANA	GUTIERREZ	UBALDO ISIDRO											
19	2007-30372	LANCHIPA	LAQUI	CHRISTIAN CESAR											
20	2007-30774	TORRES	CHURA	CANDIDA											
21	2007-30781	PEREDES	PONCE	HECTOR WALTHER											
22	2007-30794	MAYTA	YUCRA	WILLIAMS LUCIO											
23	2007-31111	VIZCARRA	DIÑAS	JUAN VIDAL											
24	2007-31123	CALDERÓN	RAMOS	ENZO ENRIQUE											
25	2007-31155	SOLÓRZANO	VANEGAS	DIANA CKRIS											

Figura 17 : Archivo plano estudiantes.cvs

Fuente: Elaboración propia

Con estos archivos se empezó a implementar y diseñar un modelo de DM que involucre los datos que se encuentran dentro de estos archivos, antes de llevar a cabo el poblado de del DWH se realizó una limpieza y homologación de datos.

3.5.2 Modelo lógico data stage area

El siguiente modelo se origina para que los usuarios puedan extraer información directamente de las fuentes de dato, teniendo como finalidad no interferir con el funcionamiento normal de las fuentes de datos y tener los datos recolectados en una base de datos propia del DM.

Basándonos en las fuentes recibidas se implementó el siguiente modelo lógico, el cual contiene representaciones de entidades y atributos, identificadores exclusivos, subtipos y supertipos y restricciones entre relaciones.

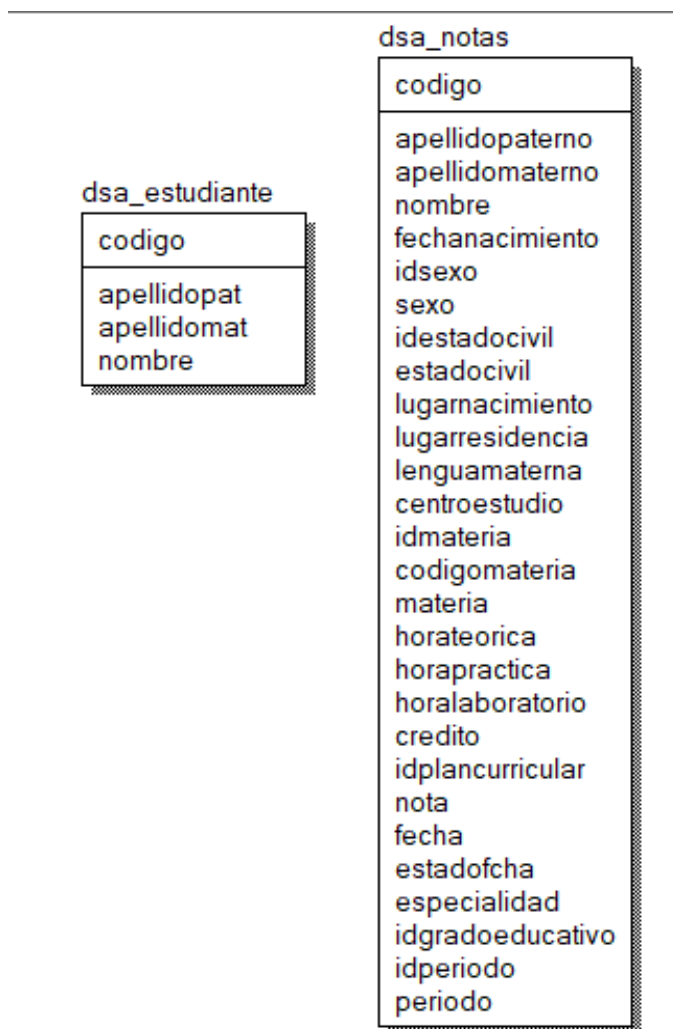


Figura 18 :Data stage area notas

Fuente: Elaboración propia

El modelo contiene 2 tablas. La correspondiente descripción a continuación:

Data stage area notas

Tabla implementada para acoger toda la información recibida en el archivo transaccional Notas, esta tabla es capaz de guardar los datos de códigos de los estudiantes, materias, cursos, carrera, número, procedencia, créditos, notas y horarios.

Data stage area estudiantes

Entidad implementada con la finalidad de guardar todos los datos de los estudiantes recibidos del archivo estudiantes, esta tabla contendrá los datos como código de estudiante, nombres, apellido paterno y materno.

3.5.3 Modelo físico stage

Forma parte de la implementación de una DM, es la estructura de cómo se encuentra desarrollado el físico de la base de datos.

Este modelo ha sido creado partiendo del modelo lógico, es un modelo específico de bases de datos que representa objetos de datos relacionales (por ejemplo, tablas, columnas, claves principales y claves externas) y sus relaciones.

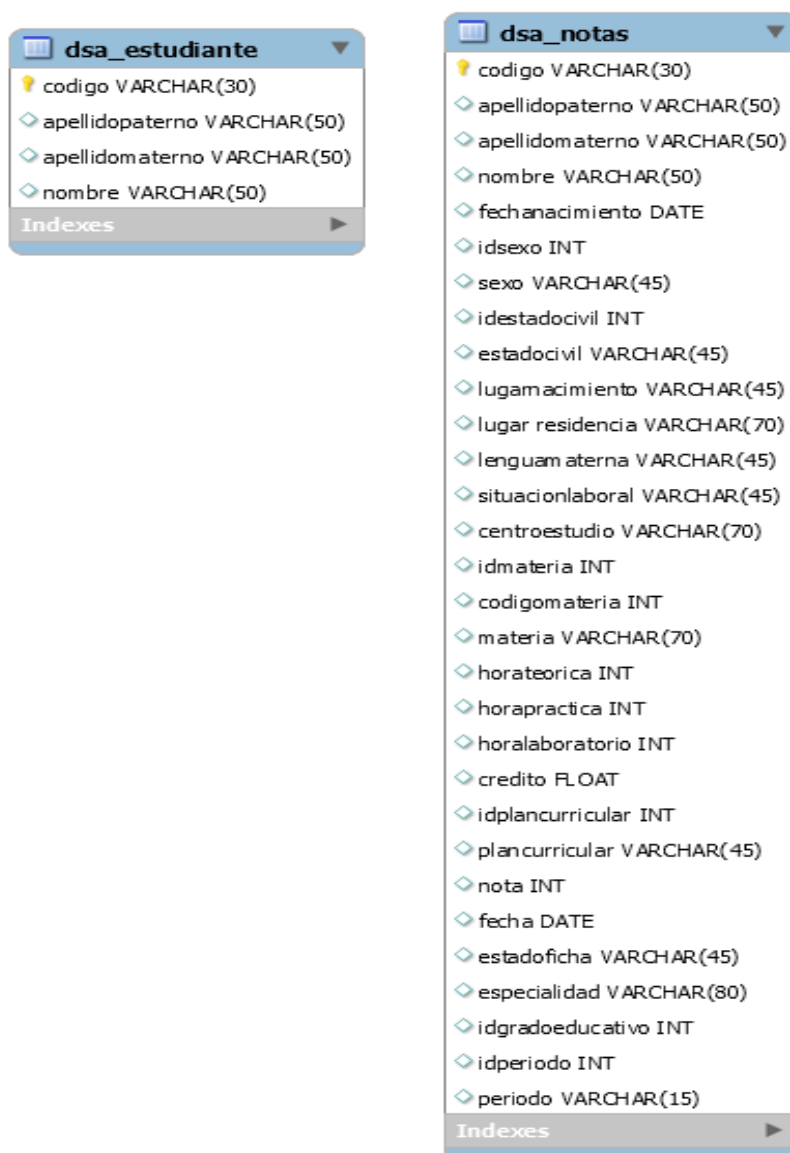


Figura 19 :Modelo lógico stage
Fuente: Elaboración propia

A continuación, se describe la estructura de cada una de las tablas que conforman el modelo Data Stage Area (DSA) de esta solución:

Tabla 6

Tablas del modelo data stage area

Owner	Name	Comment
	dsa_estudiantes	
	dsa_notas	

Fuente: Elaboración propia

Nota: Se menciona las tablas que contiene el modelo DSA

Estructura Data Stage Area Estudiantes

Tabla 7

Tabla del modelo DSA estudiantes

Name	Datatype	Is PK	Is FK
codigo	VARCHA	YES	No
ApellidoPa	VARCHA	No	No
ApellidoM	VARCHA	No	No
Nombre	VARCHA	No	No

Fuente: Elaboración propia

Nota: Descripción de los campos contenidos en la tabla DSA Estudiantes

Estructura Data Stage Area notas

Tabla 8

Tabla del modelo DSA notas

Name	Datatype	Is PK	Is FK
Codigo	VARCHAR(30)	YES	No
ApellidoPaterno	VARCHAR(50)	No	No
ApellidoMaterno	VARCHAR(50)	No	No
Nombre	VARCHAR(50)	No	No
FechaNacimiento	DATE	No	No
IdSexo	INT	No	No
Sexo	VARCHAR(45)	No	No
IdEstadoCivil	INT	No	No
EstadoCivil	VARCHAR(45)	No	No
LugarNacimiento	VARCHAR(45)	No	No
LugarResidencia	VARCHAR(70)	No	No
LenguaMaterna	VARCHAR(45)	No	No
SituacionLaboral	VARCHAR(45)	No	No
CentroEstudio	VARCHAR(7')	No	No
IdMateria	INT	No	No
Materia	VARCHAR(70)	No	No
HoraTeorica	INT	No	No
HoraPractica	INT	No	No
HoraLaboratorio	INT	No	No
Credito	FLOAT	No	No
IdPlanCurricular	INT	No	No
PlanCurricular	VARCHAR(45)	No	No
Nota	INT	No	No
Fecha	DATE	No	No
EstadoFicha	VARCHAR(45)	No	No
Especialidad	VARCHAR(80)	No	No
IdGradoEducativo	INT	No	No
IdPeriodo	INT	No	No
Periodo	VARCHAR(15)	No	No

Fuente: Elaboración propia

Nota: Descripción de los campos contenidos en la Tabla DSA Notas

3.5.4 Calidad de datos

Luego de haber recibido el archivo fuente en el que se basa todos los procesos diseñados, se procedió a realizar un paso previo que es calidad de datos a cada campo enviado en el archivo transaccional notas y estudiantes.

En este paso facilito observar algunas inconsistencias que se estaban dando en los datos de las notas, así mismo limpiar registros, caracteres especiales y datos basura que no aportaban valor al DM.

De esta manera se garantizó que al DM solo se cargan los datos limpios y validos sin duplicados y datos basura que podrían dañar la información que va a almacenar en la estructura del DM.

Todos los detalles a los que se aplicó calidad de datos se describen a continuación para que el usuario pueda ver los problemas que tenía el Excel donde almacena los datos de las notas, esto con la finalidad que en procesos posteriores la limpieza de datos los haga antes de ejecutar el proceso ETL construidos.

La limpieza de los datos abarco los siguientes puntos:

- Se eliminaron los registros cuyas calificaciones tenían 0.
- Justificación:
- Se eliminan los 0 en base a que el 0 no tiene un origen determinado

- Los que tienen 0 no se consideran ya que no han tenido una evaluación regular que permite determinar su rendimiento académico.
- Estudios regulares, evaluación regular, significa permanencia en clase para poder ser evaluado.
- Solamente los registros que han tenido evaluaciones regulares se consideraran para la población del DM
- Se tienen en total 66734 registros con notas entre 0 y 20.
- Se tiene un total de 11732 registros que no han tenido una evaluación regular.
- Se tiene un total de 55002 registros luego de eliminar los registros que no han tenido una evaluación regular.

	Q	R	S	T	U	V	W	X	
a	Materia	HoraTeorica	HoraPractica	HoraLaborati	Credito	IdPlanCurrici	PlanCurricul	Nota	Fe
	MATEMATIC	3	2	0	4	12	ESIS - 2	0	
	FISICA GENEI	3	2	0	4	12	ESIS - 2	0	
	FUNDAMENT	2	2	2	4	12	ESIS - 2	0	
	MATEMATIC	3	2	0	4	12	ESIS - 2	0	
	METODOLOG	1	2	0	2	12	ESIS - 2	0	
	QUIMICA GEI	2	2	0	3	12	ESIS - 2	0	
	TALLER DE LE	0	4	0	2	12	ESIS - 2	0	
	ANTROPOLO	1	2	0	2	12	ESIS - 2	0	
	BIOLOGIA CC	2	2	0	3	12	ESIS - 2	0	
	CIRCUITOS EI	3	2	0	4	12	ESIS - 2	0	
	ESTRUCTURA	2	2	2	4	12	ESIS - 2	0	
	MATEMATIC	3	2	0	4	12	ESIS - 2	0	
	MATEMATIC	3	2	0	4	12	ESIS - 2	0	
	SEMINARIO I	1	2	0	2	12	ESIS - 2	0	
	DISEÑO DE P	4	2	0	5	91	ESME - 3	14	
	ELECTROMET	4	3	0	5,5	91	ESME - 3	13	
	FUNDICION I	4	3	0	5,5	91	ESME - 3	13	
	GESTION DE	4	2	0	5	91	ESME - 3	11	
	METALURGIA	4	0	0	4	91	ESME - 3	12	
	NORMAS LEC	2	0	0	2	91	ESME - 3	13	
	SIMULACION	4	2	0	5	91	ESME - 3	14	
	CORROSION	2	2	0	3	91	ESME - 3	14	
	FUNDICION I	4	3	0	5,5	91	ESME - 3	14	
	GESTION DE	4	2	0	5	91	ESME - 3	11	

Figura 20 : Archivo Notas

Fuente: Elaboración Propia

3.5.5 Procesos ETL diseñados

Para construir la solución de BI descrita en este documento fue necesario construir los siguientes procesos para poder extraer toda la información entregada por el usuario y almacenarla en el DM.

Pentaho Data Integration: Es una herramienta de o componente de Pentaho Suite que permite utilizar técnicas ETL, ofreciendo datos analíticos muy precisos, eliminando complejidades involucradas en la codificación.

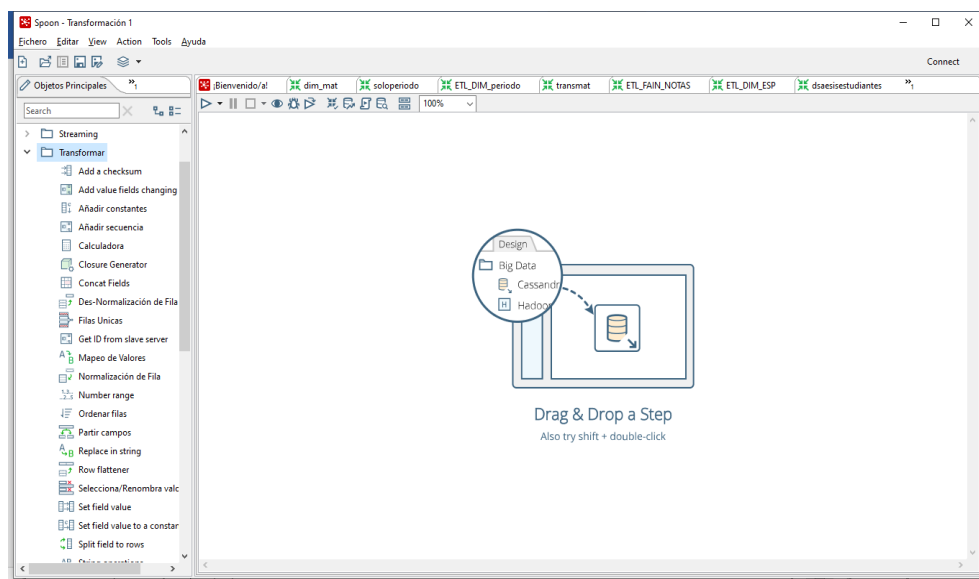


Figura 21 : Pentaho Data Integration
Fuente: Elaboración propia

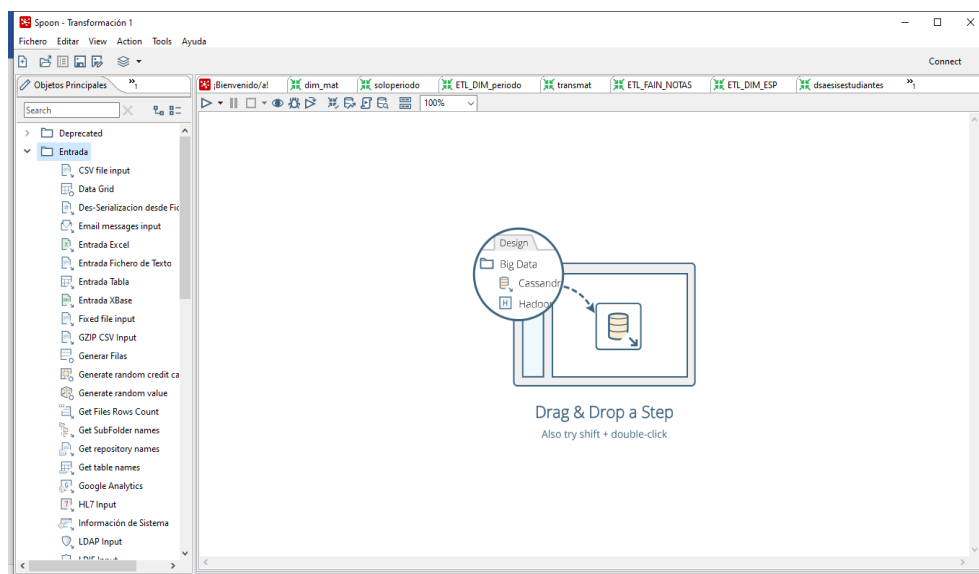


Figura 22 : Pentaho Data Integration
Fuente: Elaboración propia

A continuación, se describe cada uno de los procesos creados.

3.5.5.1 Data stage area estudiantes



Figura 23 :Proceso DSA estudiantes

Fuente: Elaboración propia

Este proceso es el encargado de extraer toda la información que se encuentra almacenada en el archivo estudiantes.csv y la almacena en la tabla dsa_estudiantes.

3.5.5.2 Data stage area notas

Este proceso es el encargado de extraer toda la información que se encuentra almacenada en el archivo Notas.csv y la almacena en la tabla dsa_notas.

La tabla dsa_notas tiene la misma estructura que el archivo csv, esto para garantizar que no se pierdan datos.



Figura 24 :Proceso DSA notas

Fuente: Elaboración propia

3.5.5.3 Dimensión especialidad

Proceso diseñado para extraer las especialidades del dsa_notas, luego mediante una transformación añadimos un surrogate key para poder identificar las distintas especialidades que almacena el dsa_notas.



Figura 25 :Proceso de la dimensión especialidad
Fuente: Elaboración propia

3.5.5.4 Dimensión materias

Proceso construido para extraer todas las materias de la tabla dsa_notas y cargarlas en la tabla dim_materia

Este proceso solo va a cargar valores distintos, es decir que dentro de proceso se quitan los duplicados antes de la carga de la tabla.



Figura 26 :Proceso de la dimensión materia
Fuente: Elaboración propia

3.5.5.5 Dimensión personas

Proceso construido en base a los datos entregados de los estudiantes, este proceso almacena los nombres, apellidos, código del estudiante.

Este proceso solo va a guardar valores distintos, es decir dentro del proceso se quitan los valores duplicados de la carga de la tabla.



Figura 27: Proceso de la dimensión persona

Fuente: Elaboración propia

3.5.5.6 Dimensión estado

Proceso construido para extraer todos los datos estados de la tabla `dsc_notas` y guardar esta información en la tabla `dim_estado`.

Antes de almacenar la información en la tabla definitiva (`dim_estados`) se eliminan los duplicados y se crea la clave subrogada para la correcta identificación de los estados.



Figura 28 :Proceso de la dimensión estado

Fuente: Elaboración propia

3.5.5.7 Dimensión periodo

Proceso construido para extraer todos los datos de la tabla dsa_notas y guardar esta información en la tabla dim_periodo.

Antes de almacenar la información en la tabla definitiva se depuran los datos duplicados, es decir solo se cargar los valores distintos.



Figura 29 : Proceso de la dimensión periodo

Fuente: Elaboración propia

3.5.5.8 Dimensión sexo

Proceso construido para extraer todos los datos estados de la tabla dsa_notas y guardar esta información en la tabla dim_sexo.

Antes de almacenar la información en la tabla definitiva se depuran los datos duplicados, es decir solo se cargar los valores distintos.



Figura 30 : Proceso de la dimensión sexo

Fuente: Elaboración propia

3.5.5.9 Tabla de hechos seguimientos notas

Proceso que realiza la carga de datos desde la tabla dsa_notas, para este proceso se utilizan solo unas columnas de la tabla dsa_notas, posteriormente se utiliza una transformación para reemplazar algunas columnas por sus respectivas llaves subrogadas, quedando la columna nota, acompañadas de las llaves subrogadas de las diferentes dimensiones.

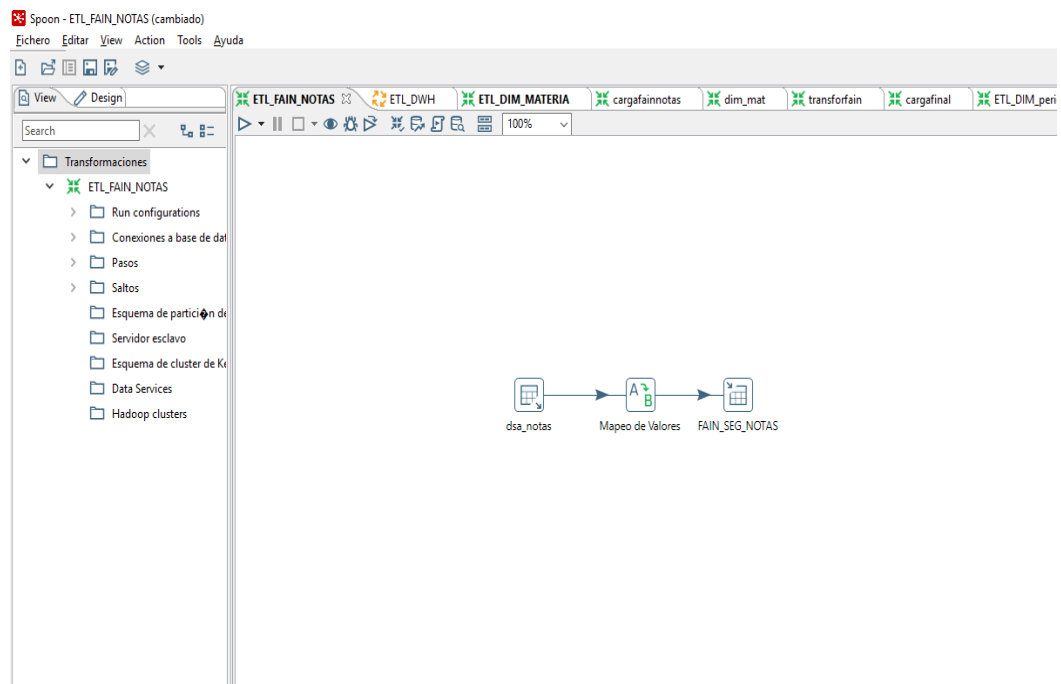


Figura 31 :Proceso tabla hechos FAIN notas

Fuente: Elaboración propia

3.5.5.10 Job Notas

Este es el Job principal diseñado para invocar a todas las transformaciones descritas anteriormente.

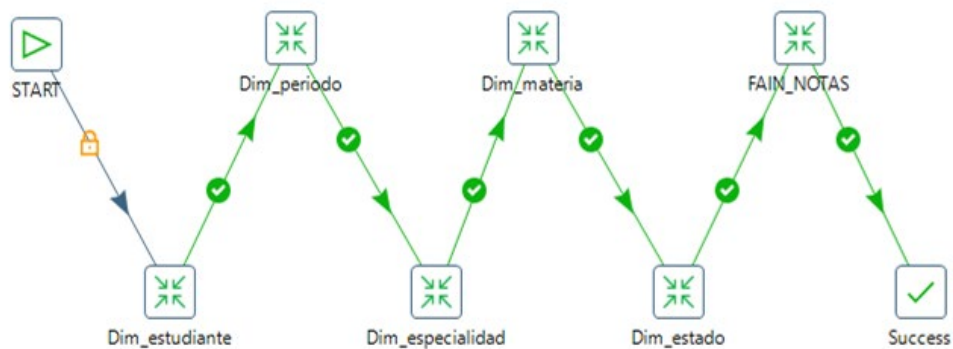


Figura 32 :Proceso job notas

Fuente: Elaboración propia

Cada uno de los procesos creados y desarrollados en el presente proyecto fue probado más de una vez para garantizar que el proceso funcione bien bajo condiciones normales.

De cada ejecución que se realizó de los procesos diseñados se guardó los tiempos de duración que se tomó en terminar su proceso normal de ejecución, esto con la finalidad de tener un tiempo estimado para próximas ejecuciones de los ETL's,

Los tiempos que arrojaron las ejecuciones de los procesos carga del DSA y la DM son los siguientes:

Tabla 9

Tiempo de ejecución de los procesos de integración de datos

	Nombre del Proceso	Tiempo de Ejecución
1	ETL_DSA_NOTAS	0:02:45
2	ETL_DSA_ESTUDIANTE	0:01:49
3	ETL_DIM_MATERIA	0:01:36
4	ETL_DIM_ESTADO	0:00:38
8	ETL_DIM_PERIODO	0:00:20
6	ETL_DIM_PERSONA	0:02:46
7	ETL_DIM_ESPECIALIDAD	0:00:37
9	ETL_DIM_SEXO	0:00:10
8	ETL_FAİN_NOTAS	0:05:28
	Total	16:08:58

Fuente: Elaboración propia

Nota: Se muestra el tiempo de cada uno de los procesos ejecutas para la Integración de los datos

3.5.6 Modelo Físico DM

El modelo físico del DM es una representación del modelo estrella que se va a desarrollar en la base de datos donde se va a almacenar el DM, este modelo detalla las estructuras de las entidades donde se recopilara los datos que se tiene como origen de esta solución.

Para implementar el modelo físico se debe tener en cuenta las siguientes atenciones:

3.5.6.1 Tipos de Datos

Inspeccionar los tipos de datos que se encuentran en la base de datos destino, dando importancia al número de dígitos en los números enteros.

Tabla 10

Tipo de datos del DM

Dato	Tipo de dato
Identificadores	Integer
Notas	Integer
Código	Varchar
Periodo	Varchar
Nombres	Varchar
Apellidos	Varchar
Código estado	Varchar
Nombre de estado	Varchar
Especialidad	Varchar

Fuente: Elaboración propia

Nota: Cuadro donde se muestra los datos del DM con sus respectivos tipos de datos, cabe señalar que para los datos del campo periodo se consideró que sea Varchar debido a que se encontraron dos clases de registros: 2018 – I y 2018, siendo la justificación necesaria para declararlos como Varchar

3.5.6.2 La precisión de los flotantes

Es importante tener un estándar de longitud de las variables o caracteres que se encuentren en el Data Mart. Se estableció la siguiente longitud para algunas variables:

- Char: Longitud 45, se declara como char(45)
- Varchar: Longitud 45, se declara como varchar(45)
- Integer: Longitud 11, se declara como int o int(11)

3.5.6.3 Llaves primarias

En algunas circunstancias se pueden presentar casos en donde la llave primaria no puede representarse en algún tipo de datos ofrecidos por la base de datos elegida, en ese caso se opta por la creación de una llave primaria, llevando esta transformación dentro del proceso de extracción, transformación y cargado (ETL).

Pocas bases de datos poseen la capacidad de *autoincrement* o *identity property* con la cual pueden automáticamente manejar algún atributo para crear llaves incrementales. Pero es importante verificar: como se manejan internamente.

En el modelo físico del DM se cuenta con las siguientes llaves primarias.

Tabla 11
Llaves primarias del Data Mart

Tabla	Llave Primaria
Dimensión materia	sidmateria
Dimensión especialidad	sidespecialidad
Dimensión estado	sidestado
Dimensión periodo	sidperiodo
Dimensión estudiante	idcodigo
Tabla hechos FAIN	

Fuente: Elaboración propia

Nota: En la siguiente tabla se describe las dimensiones del DM con sus respectivas llaves primarias, la tabla hechos FAIN, no cuenta con una llave primaria, pero sí contiene llaves foráneas

3.5.6.4 Orden de los atributos

Para el orden de los atributos dependiendo de la base de datos que se utilice, por lo habitual la secuencia es:

- Columnas de longitud constantes que no se actualizan continuamente.
- Aquellas que al pasar el tiempo no se actualizan que por lo general tendrán longitud variable.
- Las que están siendo actualizadas frecuentemente.

3.5.6.5 Integridad referencial

Se puede indicar información de cuales columnas otorgan o sirven de vínculo entre tablas, teniendo en cuenta que no todos los usuarios cuentan con una capacidad de entendimiento del modelo del DM.

El usuario podría hacerse cargo de esto, pero es mejor que el administrador de la base de datos se haga cargo debido a que se infiere que el conocimiento del administrador suele estar más especializado que el de un usuario común.

3.5.6.6 Índices

Es un atajo desde un campo llave hacia la localización real de los datos. Es el punto crucial de la optimización de velocidad de toda base de datos.

Si se busca alguna “Conjunto de elementos que se guardan de forma consecutiva en memoria” (Elticus.com, 2018) (Tupla) en base a una particularidad que no tiene un índice entonces se debe realizar un escaneo de la tabla completa lo cual implica un costo en tiempo, por eso es recomendable usar índices en:

- Llaves primarias
- Llaves foráneas
- Índices de acceso

3.5.6.7 Granularidad

Para este proyecto se consideran las siguientes granularidades:

- Reporte por promedio de notas por materia, periodo y estado.
- Reporte por porcentaje de aprobados y desaprobados por materia y periodo.
- Reporte porcentaje de aprobados, desaprobados por materia y especialidad a través del tiempo.
- Reporte promedio de notas de los alumnos por materia, especialidad a través del tiempo
- Reporte porcentaje de aprobados y desaprobados por género, especialidad y periodo.

3.5.6.8 Elección de dimensiones

Se consideran las dimensiones: periodo, especialidad, materia, alumno, especialidad, siendo elegidas en base a los indicadores del rendimiento académico y a los campos más relevantes del Data Mart.



Figura 33 : Modelado de las dimensiones

Fuente: Elaboración propia

Estas dimensiones son la base para crear seguir con el procedimiento analítico en línea, cuyo objetivo es agilizar la consulta de los datos.

La figura 33 nos muestra la distribución de las dimensiones que son utilizadas para la creación de la base de datos.

En base a todas las consideraciones descritas se construye el siguiente modelo:

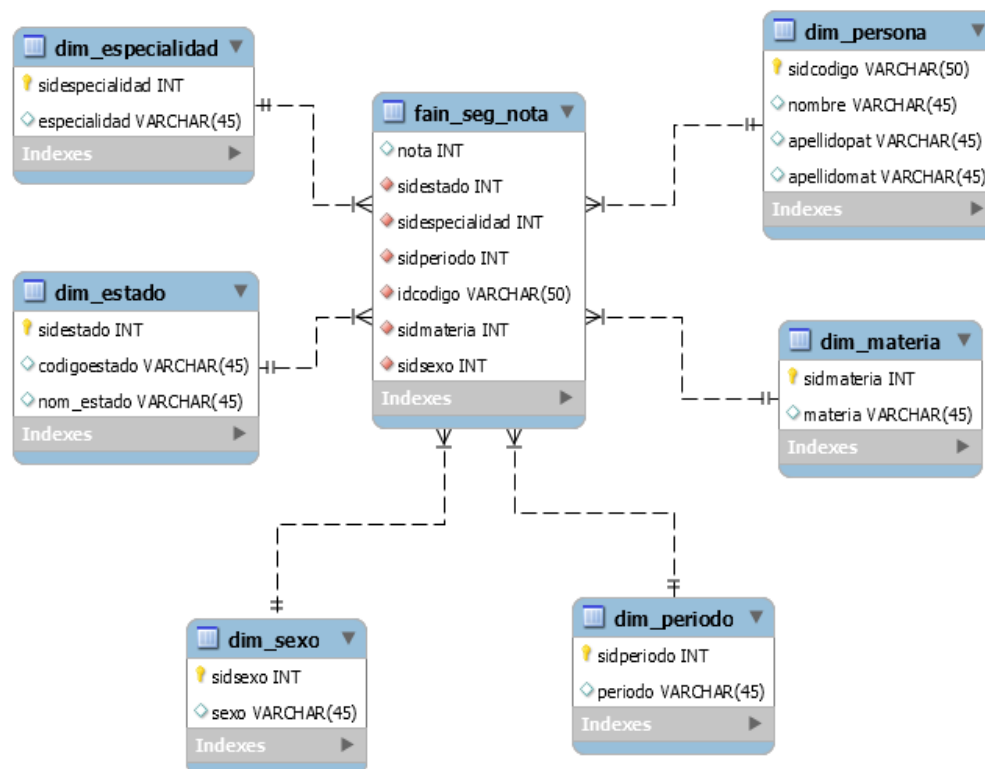


Figura 34 : Modelo físico DM
Fuente: Elaboración propia

Descripción:

El DM tiene las siguientes tablas:

Tabla 12
Estructura del DM

Owner	Name	Comment
	dim_especialidad	
	dim_estado	
	dim_periodo	
	dim_materias	
	dim_sexo	
	dim_persona	
	fain_seguimiento_notas	

Fuente: Elaboración propia

Nota: Descripción del DM

3.5.6.9 Dimensión periodo

Esta es una tabla donde se va a almacenar todo lo que se refiere al tiempo (años, semestres). Esta dimensión va a ayudar a los usuarios a obtener los indicadores, limitados a periodos de tiempo, según como se defina en esta tabla.

Tabla 13
Estructura de la dimensión periodo

Estructura de la Dimensión Periodo				
Name	Datatype	Is PK	Is FK	Comment
sidperiodo	INTEGER	Yes	No	Identificador único de la tabla periodo
Periodo	VARCHAR(45)	No	No	Periodo académico

Fuente: Elaboración propia

Nota: Descripción del contenido de la tabla dimensión periodo

3.5.6.10 Dimensión materia

Esta tabla va a guardar la información de todas las materias de la cual se componen las especialidades de la facultad de FAIN.

Tabla 14

Estructura de la dimensión materia

Estructura de la Dimensión Materia				
Name	Datatype	Is PK	Is FK	Comment
sidmateria	INTEGER	Yes	No	Identificador único de la tabla Materia
materia	VARCHAR(45)	No	No	Nombres de la materia

Fuente: Elaboración propia

Nota: Descripción de la tabla dimensión materia

3.5.6.11 Dimensión especialidad

Tabla con todas las carreras que existen actualmente en la FAIN-UNJBG.

Tabla 15

Estructura de la dimensión especialidad

Estructura de la Dimensión Especialidad				
Name	Datatype	Is PK	Is FK	Comment
sidespecialidad	INTEGER	Yes	No	Identificador único de la tabla especialidad
especialidad	VARCHAR(45)	No	No	Nombre de la especialidad

Fuente: Elaboración propia

Nota: Descripción de la tabla dimensión especialidad

3.5.6.12. Dimensión estado

Tabla que contiene los estados que un estudiante va a tener, es decir aprobado o desaprobado. A estos estados se les va asignar un código:

Tabla 16*Estados del rendimiento académico*

Código Estado	Estado
Aprob	Aprobado
Desap	Desaprobado

Fuente: Elaboración propia

Nota: Estados del rendimiento académico

Tabla 17*Estructura de la dimensión estado*

Estructura de la Dimensión Estado				
Name	Datatype	Is PK	Is FK	Comment
sidestado	INTEGER	Yes	No	Identificador único de la tabla estado
codigoestado	VARCHAR(45)	No	No	Código establecido a los estados
nom_estado	VARCHAR(64)	No	No	Nombre de los estados

Fuente: Elaboración propia

Nota: Descripción de la tabla dimensión estados

3.5.6.13 Dimensión sexo

Esta tabla almacena los diferentes géneros de los estudiantes de la facultad, es decir se encontrarán datos como: Masculino y femenino.

Tabla 18*Estructura de la dimensión sexo*

Estructura de la Dimensión Estado				
Name	Datatype	Is PK	Is FK	Comment
sidseo	INTEGER	Yes	No	Identificador único de la tabla sexo
sexo	VARCHAR(45)	No	No	Nombre de tipo de sexo

Fuente: Elaboración propia

Nota: Descripción de la tabla dimensión sexo

3.5.6.14 Dimensión persona

Esta tabla almacena toda la información de los estudiantes de la facultad, es decir que va a guardar código, nombres, apellidos.

Tabla 19*Estructura de la dimensión persona*

Estructura de la dimensión persona				
Name	Datatype	Is PK	Is FK	Comment
sidcodigo	VARCHAR(45)	Yes	No	Identificador único de la tabla persona
nombre	VARCHAR(45)	No	No	Nombre del estudiante
apellidopat	VARCHAR(45)	No	No	Apellido paterno del estudiante
apellidomat	VARCHAR(45)	No	No	Apellido materno del estudiante

Fuente: Elaboración propia

Nota: Descripción de la tabla llamada dimensión persona

3.5.6.14 Tabla hechos FAIN

Esta es la tabla hechos, la tabla principal del modelo estrella, aquí se van a guardar todos los indicadores, en este caso se va a guardar las notas, se van a crear muchos indicadores dependiendo de que reportes requiera el usuario final.

Tabla 20*Estructura de la tabla Hechos*

Estructura de la Tabla Hechos FAIN_SEG_NOTAS				
Name	Datatype	Is PK	Is FK	Comment
nota	INTEGER	NO	No	
sidestado	INTEGER	YES	YES	Identificador único de la tabla estado
sidespecialidad	INTEGER	YES	YES	Identificador único de la tabla especialidad
sidperiodo	INTEGER	YES	YES	Identificador único de la tabla periodo
sdcodigo	INTEGER	YES	YES	Identificador único de la tabla persona
sidmateria	INTEGER	YES	YES	Identificador único de la tabla materia

Fuente: Elaboración propia

Nota: Descripción de la tabla hechos llamada FAIN_SEG_NOTAS

3.5.7 Modelo Lógico DM

Modelo donde se describe la estructura lógica del Data Mart

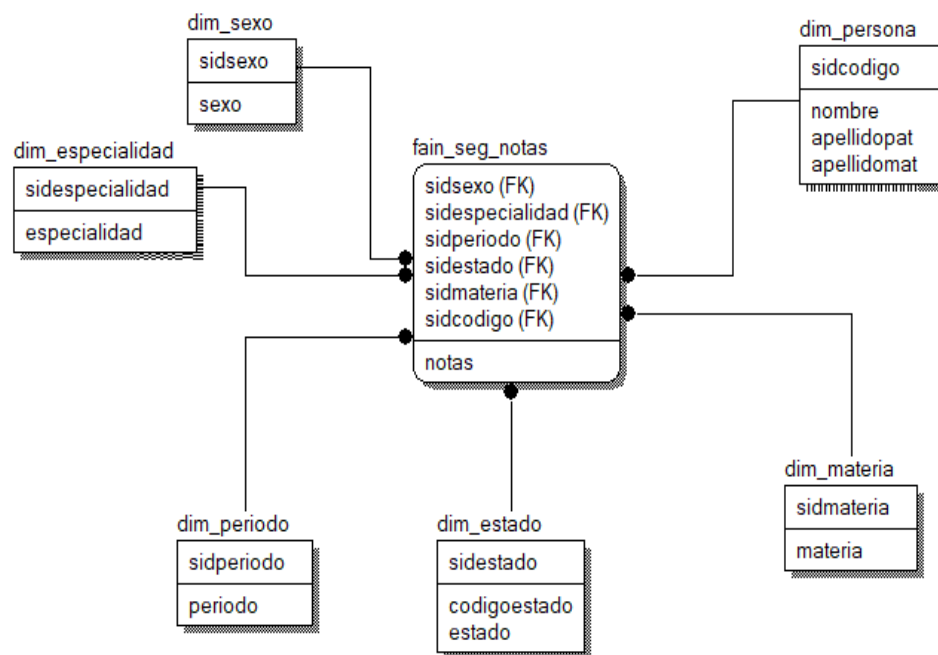


Figura 35 : Modelo lógico DM

Fuente: Elaboración propia

3.5.8 Diseño OLAP

Para la explotación de los datos del DM se realizó mediante un análisis OLAP, la cual resume el Data Warehouse en un cubo OLAP llamado FAIN notas, el cual contiene las dimensiones y medidas necesarias para que el usuario pueda realizar un análisis interactivo e intuitivo del rendimiento académico, la cual serán visualizadas a través de reportes estadísticos desplegadas por la herramienta de Business Intelligence.

Para la construcción y diseño del cubo OLAP se utilizó la herramienta Pentaho Schema Workbench la cual ayudara a los usuarios funcionales y administradores a dar mantenimiento a los cubos construidos y permitirá desarrollar más cubos si fuese el caso.

La plataforma Pentaho permite que todos los cubos OLAP que sean diseñados en Schema Workbench se los pueda publicar en el servidor de dicha plataforma, de esta manera el usuario pueda acceder vía web a visualizar los cubos OLAP y realizar análisis interactivo con todos los datos almacenados en el DM.

El cubo diseñado, contiene la siguiente estructura:

- Estudiantes: Muestra la cedula y los nombres de cada uno de los estudiantes de la facultad.
- Materia. Muestra la materia que se imparten o se impartieron en la facultad.

- Especialidades: Tiene la información de las carreras existentes en la Facultad de Ingeniería.
- Periodo: Muestra los diferentes periodos también llamados años académicos del cual se tiene registros.
- Estado: Muestra los estados del cual se califica el estado de la nota del alumno sea aprobado o desaprobado.
- Nota máxima: Muestra la nota máxima obtenida de un alumno o conjunto de ellos
- Nota mínima: Muestra la nota mínima obtenida de un alumno o conjunto de ellos.
- Total, de alumnos: Contiene el número de total de estudiantes que se tiene registrado.
- Promedio de notas: Contiene el promedio del total de notas dividido por el número de estudiantes que la obtuvo.

Cubo OLAP: Es una estructura de datos que supera las limitaciones de las bases de datos relacionales y proporciona un análisis rápido de datos. Los cubos pueden mostrar y sumar grandes cantidades de datos, a la vez que proporcionan a los usuarios acceso mediante búsqueda a los puntos de datos.

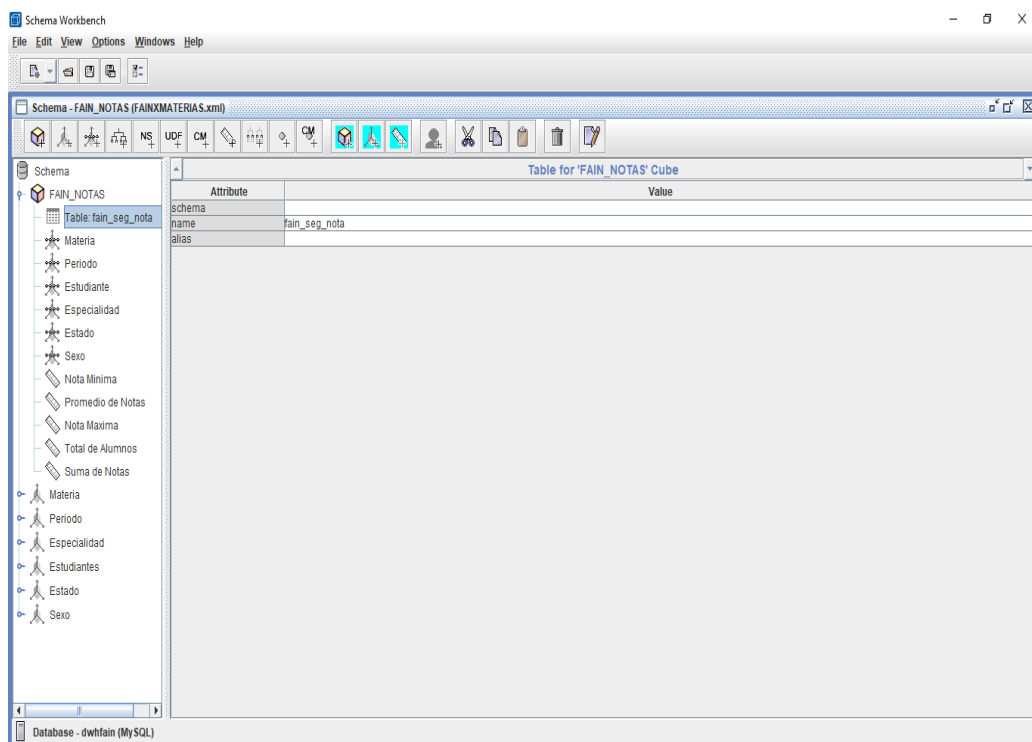


Figura 36 : Cubo OLAP FAIN
Fuente: Elaboración propia

3.5.9 Diseño de Reportes

Para la visualización de la información se utilizó *Saiku Analytics* que es un visor OLAP, trabaja como un *plugin* sobre sobre nuestro servidor ofrecido por Pentaho.

3.5.9.1 Interfaz de Saiku Analytics

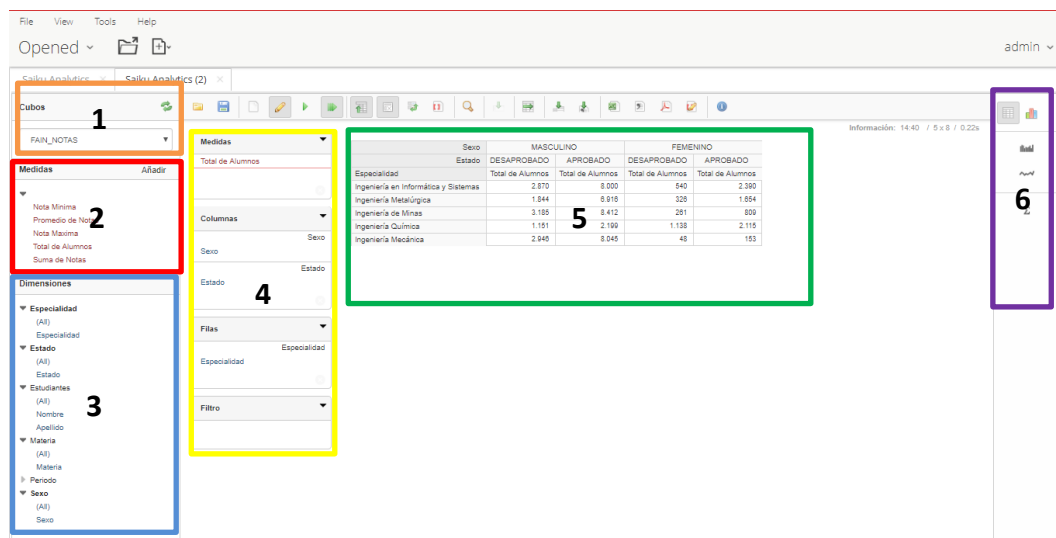


Figura 37 : Interfaz Saiku Analytics
Fuente: Elaboración Propia

A continuación, se describirá los recuadros remarcados dentro de la Figura 34.

Descripción de la interfaz

1. En el cuadro 1, se encontrarán los diferentes cubos diseñados para el análisis con Saiku Analytics
2. En el cuadro 2, se encuentran las medidas y operaciones que sean necesarias para el desarrollo de los reportes. En este caso se tienen en cuenta medidas como nota mínima, nota máxima, promedio de notas, total de alumnos y suma de notas.

3. En el cuadro 3, se encuentran las dimensiones del cubo, las cuales el usuario podrá interactuar y realizar reportes de acuerdo a sus necesidades en forma dinámica. Siendo necesario para ello realizar un doble clic en la dimensión o arrastre, hacia el cuadro de columnas o filas según sea la necesidad.
4. El cuadro 4, se encuentran las medidas y dimensiones disponibles, que se necesitan para realizar el reporte especializado.
5. El cuadro 5, se visualiza el resultado de la elección de dimensiones y medidas necesarias del reporte, resumidas en un cuadro o grafico estadístico. Según sea la necesidad del usuario.
6. El cuadro 6, se encuentran las funciones para realizar diferentes visualizaciones del reporte contando con diferentes gráficos estadísticos y algunas operaciones de estadísticas básicas.

3.5.9.2 Reporte A

Promedio de notas de los estudiantes de FAIN, del periodo 2017: El siguiente reporte mostrara el promedio del rendimiento académico de las 5 especialidades de la facultad de Ingeniería, durante el periodo 2017, considerando Matemática I, como indicador del rendimiento académico de los estudiantes, siendo este curso uno de los pilares que tienen en común dichas especialidades.

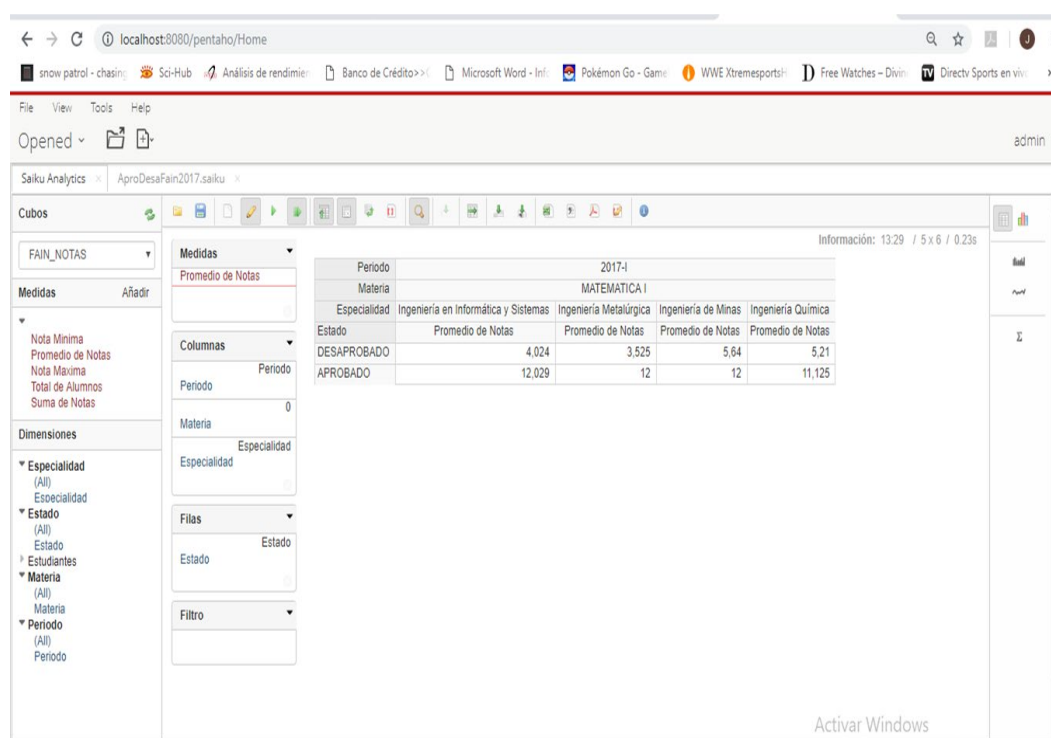


Figura 38 : Cuadro estadístico del reporte A
Fuente: Elaboración Propia

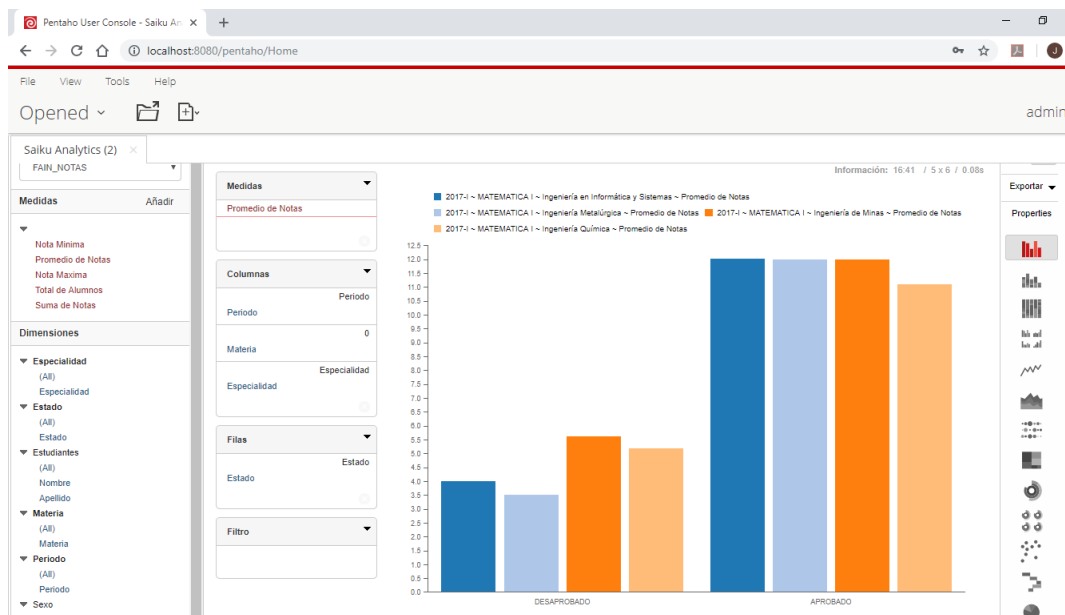


Figura 39 : Grafico de barras del reporte A
Fuente: Elaboración Propia

Como resultado se obtuvo un cuadro estadístico y su respectivo gráfico de barras, del cual se infiere que la especialidad con el promedio más alto es la E.P. de Ingeniería en Informática y Sistemas con un promedio de 12.029.

Siendo Ingeniería Química la última con un promedio 11,125.

Resaltando que, las interpretaciones, análisis y toma de decisiones, que se puedan disponer son a criterio de un personal especializado y relacionado con el rendimiento académico de los estudiantes.

3.5.9.2 Reporte B

Porcentaje de aprobados y desaprobados de los estudiantes de FAIN, del periodo 2017 – I: El siguiente reporte muestra el porcentaje de alumnos aprobados y desaprobados de la Facultad de Ingeniería, durante el periodo 2017, del curso de Matemática I, siendo este un indicador del rendimiento académico, debido a que es una materia fundamental y base para el correcto desempeño dentro de las de ingenieras en general.

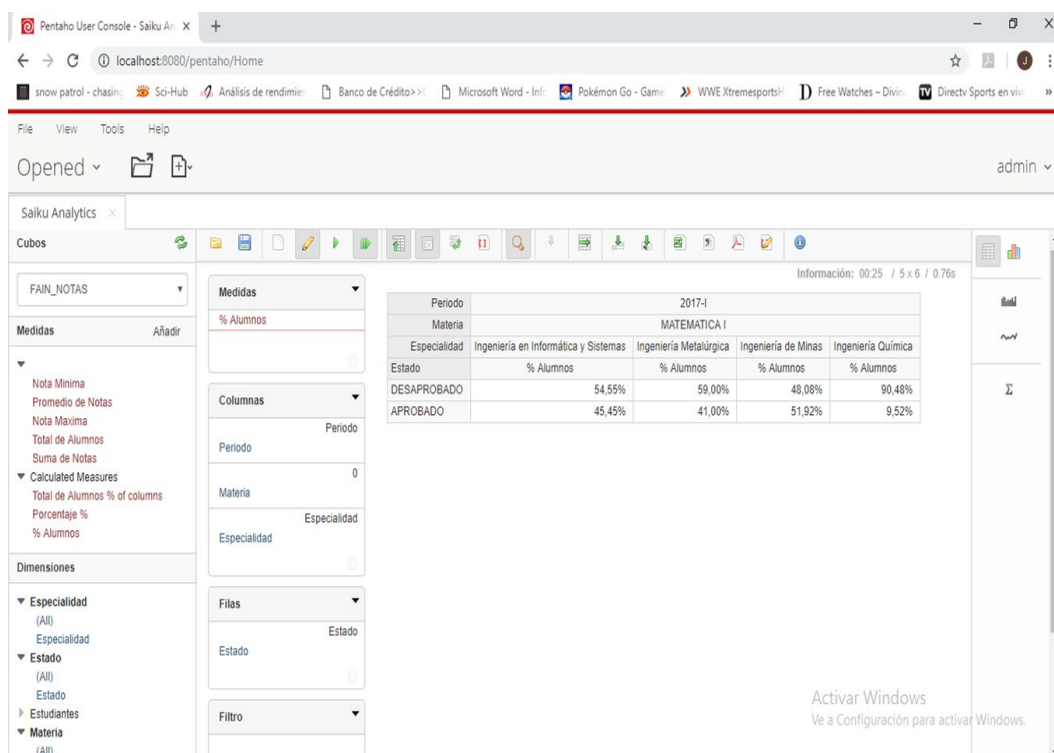


Figura 40 :Cuadro estadístico del reporte B

Fuente: Elaboración Propia

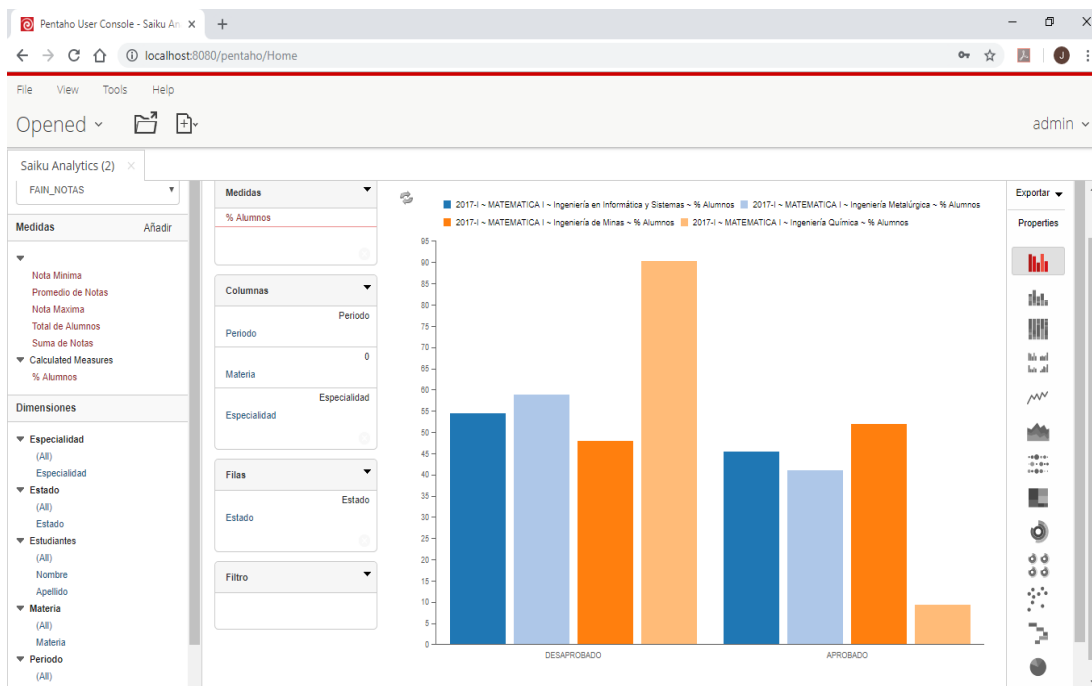


Figura 41 :Gráfico de barras del reporte B

Fuente: Elaboración Propia

Como resultado se obtuvo un gráfico de barras el cual nos muestra, el resumen del reporte, del cual se puede inferir que en la carrera de Ingeniería de Minas cuenta con el mayor porcentaje de aprobados en la materia de matemática I, por otro lado, podemos inferir que en la E.P. de Ingeniería Química cuenta con el menor porcentaje de aprobados en dicha materia en el periodo 2017.

Esta información nos permitirá analizar y proveer decisiones, para mejorar y entender la situación del rendimiento académico en la Facultad de Ingeniería

3.5.9.3 Reporte C

Porcentaje de aprobados y desaprobados de la E.P. de Ingeniería en Informática y Sistemas, durante el periodo 2015 y 2017: Para el siguiente reporte se considera como indicadores al curso de estadística y probabilidades, y estructura de datos, midiendo el rendimiento académico basándose en el porcentaje de aprobados y desaprobados de estos cursos, para así poder tener una idea del rendimiento académico de ambos cursos, que tienen un impacto diferente en la especialidad.

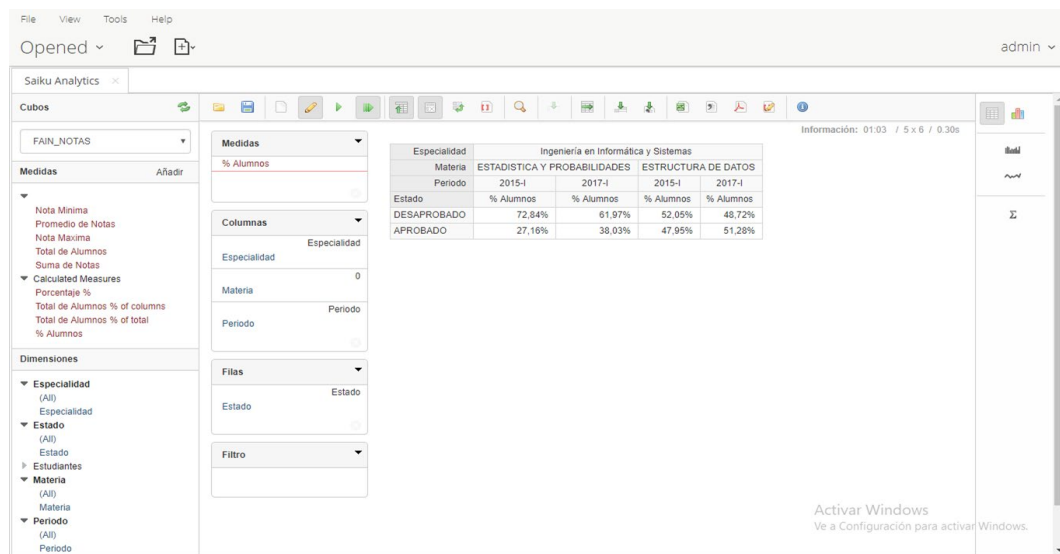


Figura 42 : Cuadro estadístico del reporte C

Fuente: Elaboración propia

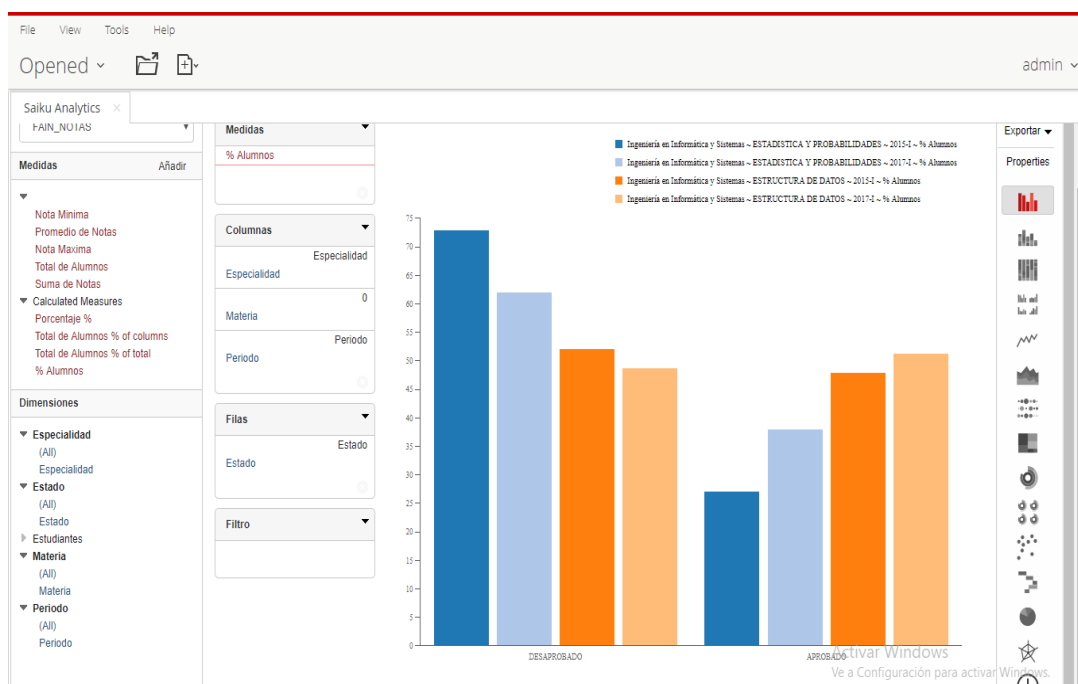


Figura 43 : Grafico de barras del reporte C

Fuente: Elaboración propia

Como resultado del reporte se puede inferir que en el periodo 2015 la materia de estadística y probabilidades contaba con 27% de aprobados, posteriormente en el periodo 2017 aumento a un 38%, pudiendo ser este un resultado positivo, pero quizás no el esperado, siendo un porcentaje bajo de aprobados, motivo por el cual se necesitan implementar estrategias para mejorar estos resultados.

3.5.9.4 Reporte D

Promedio de notas de los estudiantes de la E.P. de Ingeniería en Informática y Sistemas, durante el periodo 2015 y 2017, por materias:

El siguiente reporte muestra el promedio de las notas obtenidas de los estudiantes, frente a los cursos de estadística y probabilidades, y estructura de datos, de la E.P. de Ingeniería en Informática y Sistemas. Considerando que el resultado proporciona una visión del rendimiento académico a través de diferentes periodos académicos, pero considerando los mismos cursos.

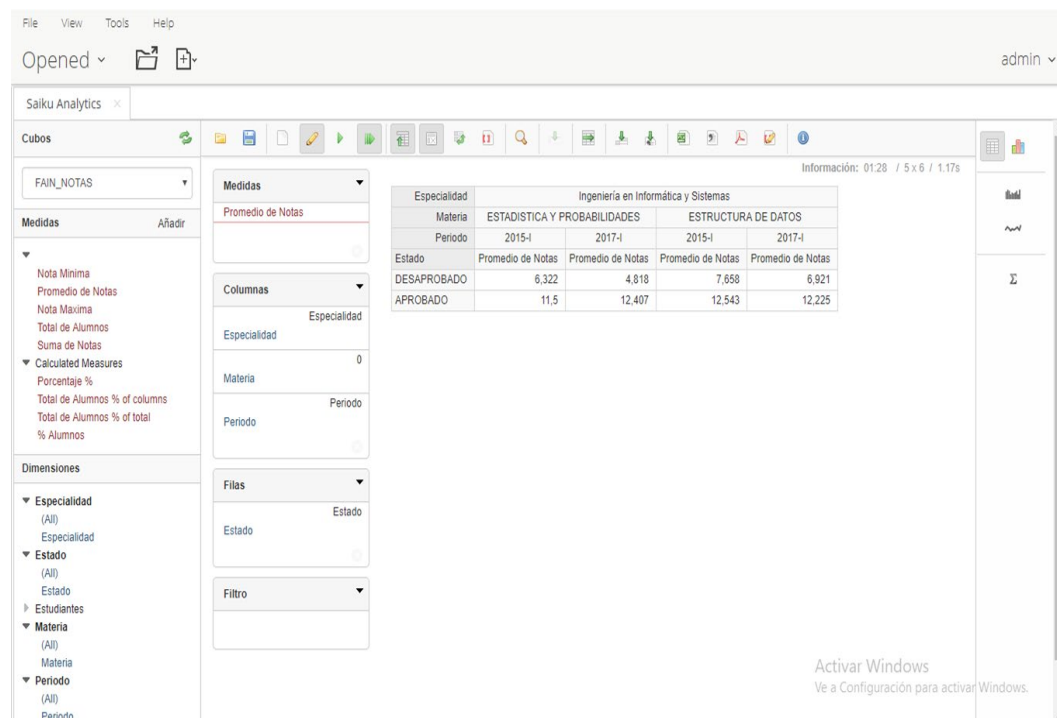


Figura 44 :Cuadro estadístico del reporte D

Fuente: Elaboración propia

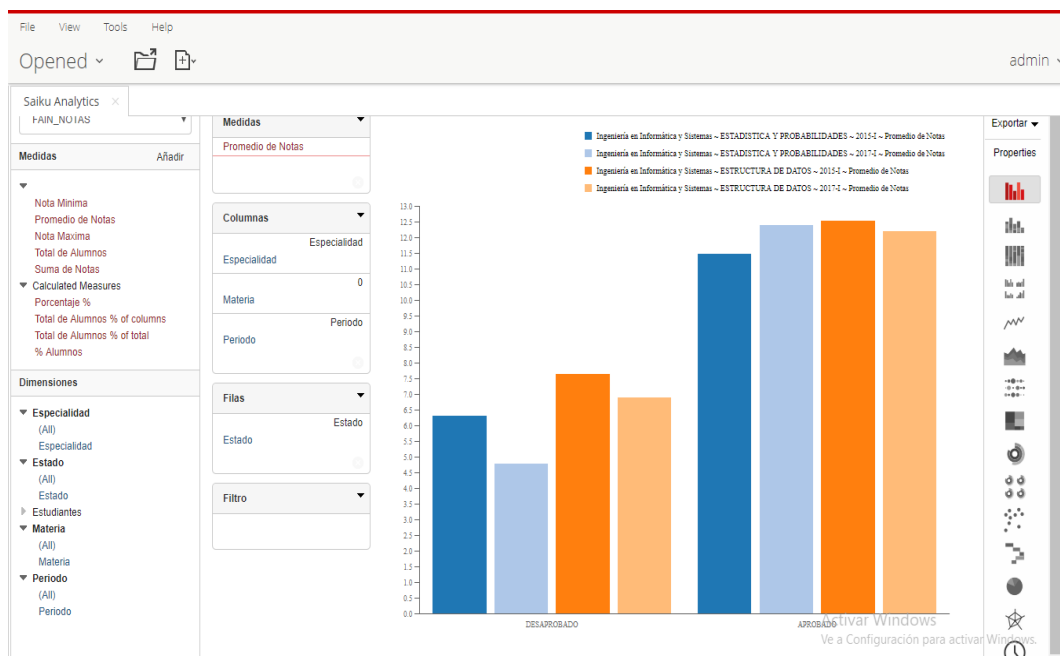


Figura 45 : Grafico de barras del reporte D

Fuente: Elaboración propia

Como análisis del reporte se puede inferir que para la materia de estadística y probabilidades ha habido una mejora en el rendimiento académico, durante los periodos 2015 y 2017, considerando que en el 2015 el promedio alcanzando de aprobados fue 11.5, y en el 2017 alcanzando un promedio de 12.4. Para el curso de Estructura de datos, se puede inferir que ha habido una disminución del promedio de las notas de los estudiantes, alcanzando en el 2015 un promedio de 12.54 y en el periodo 2017 teniendo un promedio de 12.22.

3.5.9.5 Reporte E

Porcentaje de aprobados y desaprobados por género de la E.P. de Ingeniería en Informática y Sistemas, e Ingeniería de Minas, del periodo 2017.

En el siguiente reporte se muestra el porcentaje alumnos de dos carreras profesionales elegidas por el usuario, también se toma en cuenta al curso de Matemática I, como materia fundamental de ambas carreras, y se considera reunir la información por género.

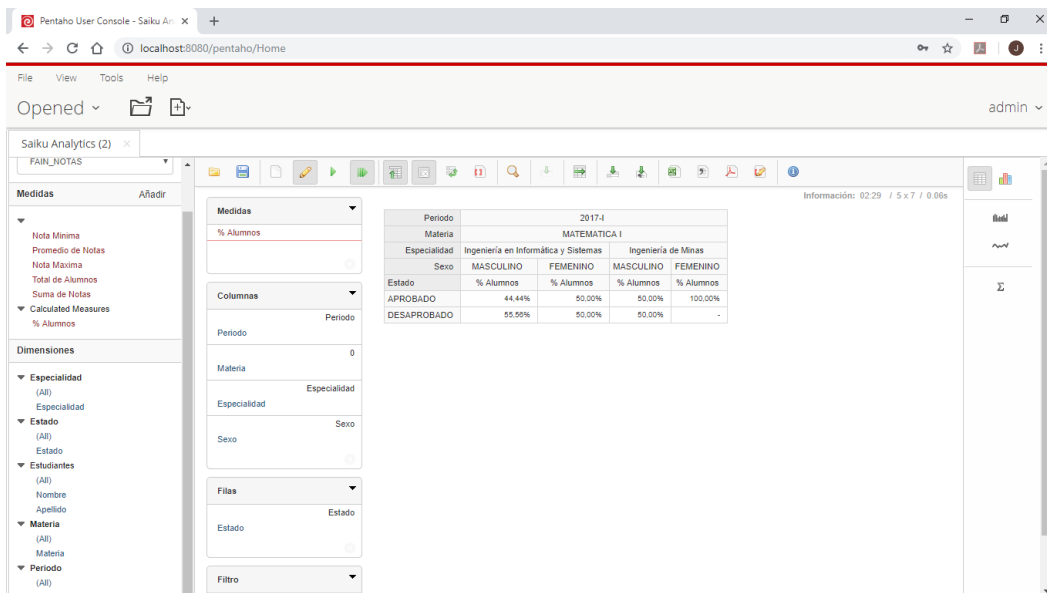


Figura 46 Cuadro estadístico del reporte E

Fuente: Elaboración propia

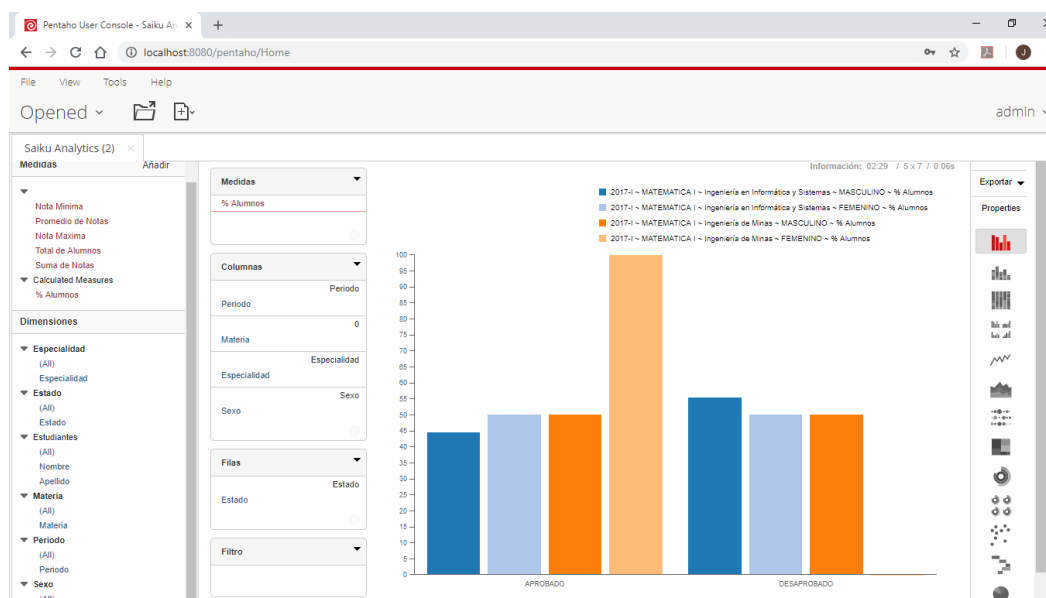


Figura 47 Grafico de barras del reporte E

Fuente: Elaboración Propia

Como resultado del reporte se obtiene un gráfico de barras y un cuadro estadístico, siendo la información necesaria para inferir que, en la E.P. de Ingeniería Química el total de mujeres logro aprobar satisfactoriamente la materia de Matemática I, en tanto en la E.P. de Ingeniería en Informática y Sistemas el porcentaje se reparte de manera equitativa, también ocurre algo similar con los hombres de ambas carreras.

No se puede dar un mayor análisis debido a que este debe ser dado por una persona especializada, y relacionada de alguna forma con la toma de decisiones del rendimiento académico de los estudiantes.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1 Resultados

1. Los resultados obtenidos del análisis del modelo multidimensional son reportes detallados del rendimiento académico, que muestran indicadores que ayudarán notablemente a observar un panorama de la situación del rendimiento académico estudiantil y se podrán tomar mejores acciones para la creación de estrategias que ayuden a evitar potenciales fracasos académicos. Analizando el reporte A del apartado 3.4.7 del Capítulo III; Promedio de notas por materia: Muestra el promedio de las notas, de los alumnos matriculados en el periodo 2017 – I, de la materia Matemática I, de las especialidades de la facultad de Ingeniería. Se puede interpretar que la especialidad con mayor promedio de notas aprobatorias es la de Ingeniería en Informática y Sistemas con 12,029, y teniendo en último lugar a la especialidad de Ingeniería Química con 11,125. Por otro lado, analizando el promedio de los desaprobados de FAIN se encuentra a Ingeniería de Minas con un promedio de 5,64, y en último lugar la escuela de Metalurgia con el promedio más bajo de desaprobados 3,525.

Este conocimiento puede servir a las autoridades a interpretar si los promedios de los alumnos son los esperados, determinar qué escuela necesita una mejora en la estrategia de formación académica y así poder elevar el promedio tanto como los desaprobados y aprobados. Resultado un importante aporte para apoyar la toma de decisiones.

2. Se implementó un modelo multidimensional para OLAP (ver ítems 3.5.2

Modelo lógico data stage area y 3.5.3Modelo físico stage) a partir de los archivos transaccionales del rendimiento académico que fueron obtenidos de las oficinas correspondientes. Luego de un proceso de limpieza, que es parte de la metodología en base a la que se implementó el modelo referido y por lo que se excluyen las notas con valor de cero, así como los campos no relacionados directamente con la granularidad de la investigación (Ver Item 3.5.6 Modelo físico DM), se estableció un DM basado en un modelo dimensional estrella (3.5.6 Modelo Físico).

3. La herramienta escogida para la implementación de BI es Pentaho BI. Esta aplicación cumple con los principales requerimientos para el desarrollo de la solución, reconociendo el acceso que el investigador tiene a las herramientas. Se eligió el uso de software libre u open source, para el desarrollo de la solución de BI (apartado 2.2.8 Comparación de Software), se justifica en base a un cuadro donde se examinan criterios tales como usabilidad, funcionabilidad y generación de reportes, que la herramienta llamada Pentaho BI es la adecuada a satisfacer las necesidad de la arquitectura de Inteligencia de Negocios, obteniendo una mejor valoración sobre la herramienta de Business Intelligence con licencia de propietario QlikView. Los reportes realizados con la herramienta Pentaho BI, muestran la información consolidada en cuadros y gráficos estadísticos, cumpliendo con la granularidad establecida y las dimensiones escogidas, permitiendo la generación de diversos reportes especializados, que proporcionan conocimiento del estado del rendimiento académico en la Facultad de Ingeniería.

4.2 Discusión

La diferencia de este trabajo con el que presentaron Zambrano, Rojas, Carvajal, y Acuña (2011), es que ellos se basaron en minería de datos haciendo un análisis con redes neuronales; y en este se analizó a través del procesamiento analítico en línea, el cual organiza y resume los datos de DM en cubos OLAP para mejorar la eficiencia de las consultas analíticas.

Como primer objetivo de la investigación fue determinar un modelo multidimensional a partir de los archivos transaccionales, en la tesis de Alcántara Castro (2016) determina las dimensiones del modelo multidimensional a partir de un análisis de requerimientos que se aplica a autoridades del entorno académico, por otro lado, Antamba (2015) basa la multidimensionalidad del modelo, en base a lógica y conceptos adquiridos a través del tiempo en su formación profesional. Considerando que para la investigación expuesta se desarrolló el modelo multidimensional basándose en conceptos de rendimiento académico y archivos transaccionales los cuales ya describen el rendimiento académico, mas no en un análisis de requerimientos a las autoridades correspondientes, lo cual no desestima el objetivo cumpliendo con una correcta consolidación de datos, los cuales son una fuente confiable para el desarrollo del modelo multidimensional.

Los indicadores obtenidos en la investigación son el resultado del análisis del modelo multidimensional, estos son generados a través de los reportes especializados de la Facultad de Ingeniería, los cuales brindan el conocimiento necesario para apoyar en la toma de decisiones, tanto Alcantara Castro (2016) en su tesis “Implementación de una solución de Inteligencia de Negocio orientada a la mejora del proceso de formación académica estudiantil en la UCH” y Antamba (2015) en su tesis “Desarrollo De Una Solución De Business Intelligence Orientado Al Análisis Del Rendimiento Académico Y La Planificación De Los Cursos De La Carrera De Ingeniería Informática En La Facultad De Ingeniería, Ciencias Físicas Y Matemática”, coinciden en la premisa que mediante los reportes obtenidos y cubos se obtienen los indicadores del rendimiento académico, dando un panorama del rendimiento académico “el cual servirá para tomar mejores acciones para la creaciones estrategias”(Alcantara, 2016)

Para determinar la herramienta de BI que permita obtener los reportes especializados es importante determinar el tipo de software que el investigador pueda adquirir, sin dejar de lado criterios como la funcionabilidad, usabilidad y la generación de reportes, aspectos considerados para esta investigación. En la tesis Alcantara Castro (2016) utiliza Sql server 2008 para el modelamiento y solución de Reporting Services, por otro lado, Cobo et al. (2016) desarrolla la solución con Pentaho debido a que la Universidad desde el 2002 adopto la decisión de trabajar con software libre, conformando un grupo para el estudio de esta herramienta y así

poder continuar el desarrollo sobre la base del software libre. También Díaz et al. (2013) se basa en Pentaho y O3, siendo herramientas provistas por el Ministerio de Educación basas en software libre. Por tanto, la herramienta de Business Intelligence es una elección que puede ser adoptada de forma personal o provista por alguna institución, ambas alternativas cumplen con las expectativas y objetivos de BI, brindando reportes especializados que ayuden a la toma de decisiones.

CONCLUSIONES

1. Se implementó un modelo multidimensional, desarrollado en un DM. Contiene datos consistentes y ordenados en un solo lugar, resultando una fuente confiable y estandarizada, permitiendo un eficiente manejo de la gran cantidad de datos que fueron recibidos a partir de los archivos transaccionales del rendimiento académico.
2. Se estructuró un modelo multidimensional a partir de la base de datos entregada por la oficina correspondiente de la UNJBG.
3. Se determinaron cinco dimensiones, que son especialidad, materia, alumnos, periodo y estado. Mediante un análisis OLAP aplicado a estas dimensiones, se obtuvo artefactos que son los indicadores que ayudarán notablemente a observar un panorama de la situación del rendimiento académico. La autoridad competente podrá usar el conocimiento consolidado para tomar mejores acciones e innovar estrategias que ayuden a evitar potenciales fracasos académicos.
4. Se determinó como herramienta a Pentaho BI, que ofrece una suite completa de software libre especializada en el desarrollo de la Inteligencia de Negocios en las empresas o instituciones. Los artefactos se crearon con Saiku Analytics, aplicación web de código libre que proporciona un visor de consulta y análisis OLAP, ofrecida dentro la suite de Pentaho.

RECOMENDACIONES

1. Para trabajos futuros trabajar con dimensiones sociales, económicas y dimensiones con datos de encuestas de perfiles biopsicosociales de los estudiantes para generar una nueva arquitectura del DM y nuevos indicadores.
2. Para el desarrollo de un proyecto de DWH o DM y/o Inteligencia de Negocios, trabajar con metodologías que favorezcan las fases de desarrollo y que faciliten llevar a cabo de una forma organizada todas las etapas.
3. Se recomienda tener una idea clara de las limitaciones del investigador o de la institución, para la elección correcta del software que se utiliza como herramienta de Business Intelligence, ya que se dispone de software libre o también llamado Open Source la cual no conlleva ningún costo monetario para su uso, y también existe el software propietario la cual para su utilización es necesario adquirir una licencia la cual implica un costo para el usuario.

REFERENCIA BIBLIOGRÁFICAS

ABC. (2007). Obtenido de Definición ABC:
<http://www.definicionabc.com/tecnologia/mysql.php>

Alcantara Castro, Nadia V. (2016). *Implementación de una solución de Inteligencia de Negocio orientada a la mejora del Proceso de Formación Académica Estudiantil en la UCH*. Lima.

Alvarez-Gayou, J. (2003). *Cómo hacer investigación cualitativa*.

Antamba, J. G. (2015). *Desarrollo de una Solución de Business Intelligence orientado al análisis del Rendimiento Académico y la planificación de los cursos de la carrera de Ingeniería Informática en la Facultad de Ingeniería, Ciencias Físicas y Matemática*. Quito, Ecuador.

Arencibia M., A. y. (2016). Herramienta informática para la toma de decisiones de las Reacciones Adversas a Medicamentos en Cuba. *Revista Cubana de Ciencias Informáticas*.

Boeije, H. R. (2009). *Analysis in Qualitative Research*. SAGE Publications Ltd.

Buigues, A. (2015). <http://anabuigues.com/2010/03/05/arquitectura-de-un-data-warehouse/> . Obtenido de <http://anabuigues.com/2010/03/05/arquitectura-de-un-data-warehouse/>

Cano Giner, J. (2007). Business Intelligence: Competir con información. En B. I. información, *Business Intelligence: Competir con información*.

Carranza Torres, M. (2004). *Problemática jurídica del software libre*. Buenos Aires: Lexis Nexis.

Castro, L. (2010). *Metodología de la Investigación*. Universidad Central de Venezuela. Caracas. Venezuela.

Castro, N. V., & Huaynate, R. D. (2016). *Implementación de una solución de Inteligencia de Negocio orientada a la mejora del proceso de Formación Académica Estudiantil en la UCH*. Lima - Perú.

Yalan, J. y. (2013). Implementation of a Datamarts Solution Business Intelligence for T-Impulso Logistics Area. *Revista de Investigación de Sistemas e Informática*, 53-63.

Chadwick, C. (1979). Teorías del aprendizaje y su implicancia en el trabajo en el aula. *Revista de Educación*.

Cobo, H., Gurmendi, L., Menéndez, M., & others. (2016). Business Intelligence en el Sistema Universitario Nacional. *Sexta Conferencia de Directores de Tecnología de Información, TICAL 2016*.

Cohen, L., Manion, L., & Morrison, K. (2003). *Methods in education*.

Coleman, H. y. (2005). *Social work: Research and evaluation. Quantitative and qualitative approaches*.

Creative Commons Attribution. (2016). *datawarehouse.wikispaces.com*. Obtenido de <http://datawarehouse.wikispaces.com/>

Creswell, J. w. (2008). *Diseño de la investigación*. Argentina.

Davenport, T. H. (1998). working knowledge: how organizations manage what they know.

DIARIUM. (2016). <https://diarium.usal.es/id00710310/2016/03/16/business-intelligence/>. Obtenido de <https://diarium.usal.es/id00710310/2016/03/16/business-intelligence/>

elticus.com. (2018). *elticus.com/*. Obtenido de <http://elticus.com/diccionario.php?palabra=Tupla>

Díaz, F. J., Osorio, M. A., Amadeo, A. P., & Romero, D. (2013). Aplicando estrategias y tecnologías de Inteligencia de Negocio en sistemas de gestión académica. *XV Workshop de Investigadores En Ciencias de La Computación*, 225–229. Retrieved from <http://hdl.handle.net/10915/27157>

Fabrizio. (2010). *http://biverano2011.blogspot.com*. Obtenido de <http://biverano2011.blogspot.com/2011/09/modelo-estrella-y-modelo-copo-denieve.html>

Grbich, C. (2007). *Qualitative Data Analysis*. SAGE Publications.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. del P. (2014). *Metodología de la investigación* (6a ed., Vol. 136). Mc Graw Hill Education.

Inmon, W. (2003). Características de un Data Warehouse.

Jordi Casal & Enric Mateu (2003). *Tipo de muestreo*. Rev. Epidem Med.

Jácome, J. S. (2015). *Análisis comparativo de herramientas de Software Libre y Propietario para la gestión de Big Data en empresas de Comercialización Masiva*. Quito, Ecuador.

Jiménez, M. (2000). *Competencia social: intervención preventiva en la escuela*. Infancia y Sociedad.

Juárez, C., Herrera, G., & Sánchez, T. (2006). Software libre vs software propietario.

Li, S. y. (2007). *Learning to do qualitative data analysis: An observational*.

- López Beltrán, C. P. (2007). Análisis, diseño e implementación de un data MART para la Dirección Financiera y Recursos Humanos de la Escuela Politécnica del Ejército para una toma de decisión efectiva, 1–226. Retrieved from <http://repositorio.espe.edu.ec/handle/21000/400>
- Málaga Tejada, Gianfranco A. (2016). *Modelo de Gestión de Incidentes Basado en ITIL v. 3*. Tacna - Perú.
- Martines, J. A. (2015). *www.gestiopolis.com*. Obtenido de <http://www.gestiopolis.com/olap-y-el-diseno-de-cubos/>
- Moss, L. T. (2013). *Business Intelligence Road Map*.
- OBS Business School. (2018). *www.obs-edu.com/*. Obtenido de <https://www.obs-edu.com/int/blog-project-management/metodologia-agile/pros-y-contras-de-la-metodologia-en-cascada>
- Pacheco, O. (2015). *Data warehouse. Obtenido de Data Warehouse*. Obtenido de <http://dwhucv.blogspot.com/p/estructuar-de-un-data-warehouse.html>
- Pascal, G., Servetto, D., Lobo Mirassón, U., & Luna, Y. (2017). Aplicación de Business Intelligence para la toma de decisiones en Instituciones Universitarias. Implementación de Boletines Estadísticos en la Universidad Nacional de Lomas de Zamora (UNLZ), 4. <https://doi.org/10.1093/rfs/hhi018>
- Pino, S. A. (2014). Derecho Penal Informático.
- Pollo Cattaneo, Maria F. (2008). *Análisis de precisión de técnicas de agregación en contextos experimentales pocos maduros*. Buenos Aires.
- Ritchie, J., Lewis, J., & Nicholls, C. y. (2013). *Qualitative research practice:.* Londres: SAGE.
- SINEXUS. (2018). Obtenido de <http://www.sinnexus.com/>: http://www.sinnexus.com/business_intelligence/datawarehouse.aspx

The SAGE Glossary of the Social and Behavioral Sciences. (2009).

Universidad Nacional Jorge Basadre Grohmann. (2018). <http://www.unjbg.edu.pe/>.

Obtenido de <http://www.unjbg.edu.pe/institucion/vision.php>

Urquizu, P. (2015). Obtenido de www.businessintelligence.info:
<http://www.businessintelligence.info/definiciones/cubos-olap.html>

Velez, E., & Schiefelbein, E. (1994). *Factores que afectan el rendimiento académico en la educación primaria: Revisión de la literatura de America latina y el Caribe*. Revista latinoamericana de innovaciones educativas argentinas.

Wayne W. Daniel (2014) Bioestadística : Bases para el análisis de las ciencias de la salud. Editorial Willey Milusa, D.F., Mexico

Zambrano Matamala, C., Rojas Díaz, D., Carvajal Cuello, K., & Acuña Leiva, G. (2011). Análisis de rendimiento académico estudiantil usando data warehouse y redes neuronales. *Ingeniare. Revista Chilena de Ingeniería*, 19(3), 369–381.
<https://doi.org/10.4067/S0718-33052011000300007>

ANEXOS

Anexo 1

Matriz de Consistencia: Título: Análisis multidimensional del rendimiento académico de la UNJBG utilizando herramientas Business Intelligence

PROBLEMA	OBJETIVOS	HIPÓTESIS	VARIABLES Y SUBVARIABLES	INDICADORES	TIPO Y Población
1. Problema Principal: ¿Es factible aplicar un análisis multidimensional del rendimiento académico de la UNJBG utilizando herramientas Business Intelligence? 2. Problema específico ¿Es factible implementar un multidimensional a partir de los archivos transaccionales de rendimiento académico? ¿Es factible determinar las dimensiones que permitan generar indicadores para la toma de decisiones? ¿Es factible elegir una herramienta BI que permita obtener reportes especializados?	1. Objetivo General: Aplicar un análisis multidimensional del rendimiento académico en la UNJBG utilizando una herramienta BI 2. Objetivos Específicos: • Implementar un modelo multidimensional a partir de los archivos transaccionales de rendimiento académico • Determinar las dimensiones que permitan generar indicadores para la toma de decisiones. • Elegir una herramienta BI que permita obtener reportes especializados	1. Hipótesis General: No aplica	1. Variables de la Investigación. Análisis multidimensional del rendimiento académico	Para la Variable: • Granularidad • Dimensiones • Estudiantes Matriculados • Aprobados • No-Aprobados	1. Tipo y nivel de la Investigación: 1.1. Tipo de la Investigación: Descriptiva – No Experimental 2. La Población (N) y Muestra (n): 2.1 La Población: Registro Académico FAIN - UNJBG 2.2 La Muestra: Por conveniencia Registro de notas FAIN 2014-2017

Anexo 2

SOLICITO: REGISTRO DE NOTAS DEL
SISTEMA ACADÉMICO

DR. OSCAR MAMANI AGUILAR, DIRECTOR DE LA DIRECCIÓN ACADÉMICA DE
ACTIVIDADES Y SERVICIOS ACADÉMICOS

Yo, JORGE RAFAEL MAQUERA
PARIHUANA identificado(a) con D.N.I. N°
47398345 con. Domiciliado Calle San Camilo N°815-
Tacna ante usted con el debido respeto me
presento y expongo:

Que siendo estudiante de la carrera profesional INGENIERÍA EN INFORMÁTICA Y
SISTEMAS, y que actualmente vengo desarrollando mi tesis de pregrado denominada
"Análisis multidimensional del rendimiento académico de la UNJBG utilizando
herramientas Business Intelligence" con resolución N° 04553-2017-FAIN/UNJBG, y por
razones de mi investigación es que solicito a Ud. se me proporcione el registro de
notas de la facultad de ingeniería del periodo 2014 al 2017.

Por todo lo expuesto ruego a Ud. Se sirva acceder a mi solicitud.

Tacna, 21 de Noviembre del 2018



[Signature]
JORGE RAFAEL MAQUERA PARIHUANA
DNI: 47398345