



UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE COMPUTACIÓN
CENTRO DE INVESTIGACIÓN DE SISTEMAS DE INFORMACIÓN CISI

**SOLUCIÓN DE INTELIGENCIA DE
NEGOCIOS PARA EL SEGUIMIENTO DE
PROYECTOS DE INVESTIGACIÓN
DE LA FACULTAD DE CIENCIAS, UCV**

Trabajo Especial de Grado presentado ante la ilustre
Universidad Central de Venezuela por
Br. Mayra, Perez
Tutores: Prof. Mercy Ospina
Caracas, Mayo 2017.

UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE COMPUTACIÓN

ACTA

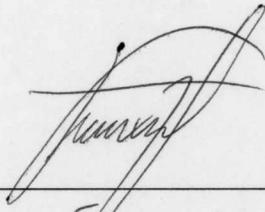
Quienes suscriben, miembros del jurado designado por el Consejo de la Escuela de Computación, para examinar el Trabajo Especial de Grado titulado "SOLUCIÓN DE INTELIGENCIA DE NEGOCIOS PARA EL SEGUIMIENTO DE PROYECTOS DE INVESTIGACIÓN DE LA FACULTAD DE CIENCIAS, UCV" y presentado por el bachiller: Br Mayra Perez, a los fines de optar al título de Licenciado en Computación, dejamos constancia de lo siguiente:

Leído como fue dicho trabajo, por cada uno de los miembros del jurado, se fijó el día 22 de Mayo del 2017 a las 9 a.m horas, para que la autora lo defendiera en forma pública, lo que estos hicieron en la Sala Aula 1 de la Escuela de Computación, mediante una presentación oral de su contenido, luego de lo cual respondieron a las preguntas formuladas. Finalizada la defensa pública del Trabajo Especial de Grado, el jurado decidió aprobar con la nota de 19 puntos.

En fe de lo cual se levanta la presente Acta, en Caracas el día 22 de Mayo del 2017



Prof. Mercy Ospina
(Tutora)



Prof. FRANKY UZCATEGUI
(Jurado)



Prof. Concettina Di Vasta
(Jurado)

AGRADECIMIENTOS

Quiero comenzar agradeciendo a Dios, esa fuerza toda poderosa que siempre me indica el camino correcto a seguir cuando me encuentro desorientada. Quiero agradecerle por cada persona que ha puesto en mi vida durante todos estos años de carrera, ya que todas me han ayudado y enseñado muchas cosas.

Quiero agradecer a mis madres, por su amor, por su apoyo, por la educación y los principios que me han dado, porque me han hecho ser la persona que soy ahora, una persona buena, correcta y que da lo mejor de sí. A mi hermana Maye y Jenni les agradezco por apoyarme a seguir adelante, a pesar de los obstáculos, y mi hermanita Paola que es mi tesoro y mi fuerza a ser cada día mejor.

Quiero agradecer enormemente a mi querido amigo German, por haberme ayudado y ser un ejemplo a seguir en mi carrera. Gracias por haber confiado en mí y darme la oportunidad de aprender tantas cosas de usted.

Quiero agradecer a mis también profesora Mercy y al profesor Franky, quienes me han dado su cariño y confianza. Les agradezco por todo el tiempo que me han dedicado cada vez que he necesito de su ayuda y por todas las palabras de motivación que me han hecho seguir adelante.

Quiero agradecer a mi novio Daniel, tantos años de amistad, amor, de cariño, confianza y tolerancia hacia mi difícil carácter. Gracias por siempre ayudarme, gracias por siempre motivarme, por estar cuando te necesito, gracias por ser quien eres, gracias por estar en mi vida, gracias mi amor.

Quiero agradecer a mi buen amigo Victor, que desde que apareció en mi vida siempre ha confiado en mí, me ha ayudado de manera excepcional y me ha dado su apoyo en todo momento.

Finalmente quiero agradecer a todas aquellas personas que me dijeron "Sí puedes", cuando yo sólo me repetía "No puedo, es imposible". Gracias Sra. Omarira, gracias Leo, gracias Jhonatan, gracias Mafer, gracias Dorjes, gracias a todos.

UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE COMPUTACIÓN
CISI

**SOLUCIÓN DE INTELIGENCIA DE NEGOCIOS PARA EL SEGUIMIENTO DE
PROYECTOS DE INVESTIGACIÓN
DE LA FACULTAD DE CIENCIAS, UCV**

Autor: Br. Mayra, Perez
Tutor: Prof. Mercy Ospina
Fecha: Mayo 2017

RESUMEN:

El presente Trabajo Especial de Grado describe una solución de Inteligencia de Negocio para la Coordinación de Investigación de la Facultad de Ciencias de la UCV, que permite obtener de sus sistemas, la información adecuada para el apoyo en la toma de decisiones en los procesos relacionados con los proyectos de investigación de los profesores y profesionales de la Facultad.

La Coordinación de Investigación cuenta con una base de datos para almacenar y procesar los datos, pero no tiene mecanismos efectivos para la interpretación, análisis y visualización adecuados para la información almacenada en ella.

Conociendo las necesidades existentes, en el presente Trabajo de Grado, se plantea el desarrollo de una solución adecuada de Inteligencia de Negocio, que permite establecer una serie de indicadores de gestión, para apoyar en la toma de decisiones gerenciales a la Coordinación de Investigación, generando información confiable que permita interpretar los datos almacenados.

La solución de Inteligencia de Negocio fue creada en su mayor parte con tecnología de software libre (PostgreSQL y Pentaho Data Integration o Kettle), junto con Tableau que es un software propietario.

El desarrollo se hizo bajo una adaptación de la metodología ascendente (Bottom-Up) de Ralph Kimball.

El presente trabajo es un prototipo debido a que la base de datos no se encuentra con datos reales. Sin embargo, es completamente funcional y está listo para cuando se tengan los datos. Adicionalmente puede ser ampliado cuando sea necesario.

Palabras Claves: Investigación, Inteligencia de Negocios, PostgreSQL, Kettle, Tableau Desktop.

ESTUDIO DE UNA SOLUCIÓN DE
INTELIGENCIA DE NEGOCIOS PARA EL
SEGUIMIENTO DE PROYECTOS DE
INVESTIGACIÓN DE LA FACULTAD DE
CIENCIAS, UCV

Índice general

1. Problema de Investigación	7
1.1. Situación Actual	7
1.2. Planteamiento del Problema	7
1.3. Solución Propuesta	8
1.3.1. Objetivo General	9
1.3.2. Objetivo Específicos	9
1.4. Justificación	10
1.5. Alcance	10
2. Marco Conceptual	11
2.1. Facultad de Ciencias	11
2.1.1. Estructura Organizativa	11
2.1.2. Coordinación de Investigación	12
2.2. Indicadores	14
2.2.1. Criterios de los Indicadores	14
2.2.2. Objetivo de un indicador	15
2.2.3. Características	15
2.2.4. Tipología de Indicadores	16
2.3. Inteligencia de Negocio (Business Intelligence)	17
2.3.1. Inteligencia	17
2.3.2. Negocio	17
2.3.3. Definición de Inteligencia de Negocio	17
2.3.3.1. Características	18
2.3.4. Niveles de Soluciones de Inteligencia de Negocio	19
2.3.5. Arquitectura de una Solución de Inteligencia de Ne- gocio	19

2.3.6.	Fuentes de Datos	21
2.3.7.	Procesos de Extracción, Transformación y Carga (ETL)	21
2.3.8.	Almacén de Datos (Data Warehouse)	22
2.3.8.1.	Definición	23
2.3.8.2.	Características	23
2.3.8.3.	Arquitectura	25
2.3.8.4.	Metadata	27
2.3.8.5.	Requerimientos de un DWH	27
2.3.9.	Bodega de Datos (Data Mart)	28
2.3.9.1.	Estrategias de Construcción	29
2.3.10.	Modelo Dimensional del DWH	29
2.3.10.1.	Definición	30
2.3.10.2.	Tabla de Dimensión	30
2.3.10.3.	Tabla de Hechos	30
2.3.10.4.	Esquema Estrella (Star Scheme)	31
2.3.10.5.	Esquema Copo de Nieve (Snowflake Scheme)	32
2.3.10.6.	Esquema Constelación (Starflake Scheme) .	33
2.3.10.7.	Granularidad	34
2.3.10.8.	Jerarquía	34
2.3.10.9.	Agregación	35
2.4.	Herramientas Tecnológicas	35
2.4.1.	Herramientas de Análisis y visualización	35
2.4.1.1.	Reportes	35
2.4.1.2.	Indicadores	36
2.4.2.	Herramienta de Extracción, Transformación y Carga (ETL)	37
2.4.2.1.	Pentaho	37
2.4.2.2.	Pentaho Data Integration	38
2.4.3.	Herramienta para crear la bodega de datos (Datamart)	39
2.4.3.1.	PostgreSQL	40
2.4.4.	Herramientas de Visualización	41
2.4.4.1.	Tableau	41
3.	Marco Metodológico	46
3.1.	Metodología Descendente (TOP-DOWN)	46

3.1.1.	Metodología de Bill Inmon	46
3.1.2.	Parte 1: Desarrollo de Sistemas Operacionales	48
3.1.3.	Parte 2: Desarrollo del almacén de datos	50
3.1.4.	Parte 3: Uso del Almacén de Datos	51
3.1.5.	Plan de Migración (Migration Path)	51
3.2.	Metodología Ascendente (BOTTOM-UP)	53
3.2.1.	Metodología de Ralph Kimball	54
3.2.2.	Planificación del Proyecto	55
3.2.3.	Definición de los Requerimientos del Negocio	55
3.2.4.	Diseño de la Arquitectura	56
3.2.5.	Selección de productos e instalación	58
3.2.6.	Diseño de Datos o modelado dimensional	58
3.2.7.	Especificación y desarrollo de aplicaciones analíticas	61
3.2.8.	Implementación	61
3.2.9.	Despliegue y Crecimiento	62
3.3.	Metodología a seguir	63
4.	Marco Aplicativo	65
4.1.	Planificación del Proyecto	65
4.2.	Definición de los Requerimientos	66
4.3.	Diseño de la Arquitectura Técnica	67
4.4.	Selección de Productos e Instalación	68
4.5.	Modelo Dimensional	69
4.6.	Diseño físico	76
4.7.	Diseño y Construcción de procesos ETL	79
4.7.1.	Diseño de Procesos de ETL y Desarrollo (almacén de datos)	79
4.7.2.	Verificar la calidad de los datos	81
4.7.3.	Verificación de la calidad de los datos de los compo- nentes del modelo implementado	87
4.7.3.1.	Verificación de la tabla de hecho investiga- ción con sus dimensiones	87
4.7.3.2.	Detallar los requerimientos según el modelo	92
4.8.	Desarrollo de los requerimientos de información	93
4.8.1.	Configuración de la conexión	93

4.8.2.	Generador de los indicadores de gestión	94
4.8.3.	Desarrollo y distribución de consultas	95
4.8.3.1.	Tabla de hecho de investigación	95
4.8.4.	Verificar la calidad de los datos en la herramienta de manipulación y visualización de datos	105
4.9.	Implementación	108
4.9.1.	Ajuste de las consultas	108

Índice de figuras

1.1. Arquitectura de la Solución Propuesta	9
2.1. Estructura organizativa de investigación de la Escuela de Física de la <i>Facultad de Ciencias de la UCV</i> . (Fuente [CoorInv, 2011]).	12
2.2. Indicadores de Gestión (Fuente [Indicador,2014])	15
2.3. Clasificación según los factores claves	16
2.4. Niveles (Fuente [Inteligencia de Negocios, 2017])	19
2.5. Componentes de inteligencia de negocio	20
2.6. Fuentes de Datos	21
2.7. Procesos de Extracción, Transformación y Carga (ETL) (Fuente [Cano, 2007])	22
2.8. Almacén de Datos	22
2.9. Histórico de DWH (Fuente [DWH, 2015])	24
2.10. No Volátil (Fuente [DWH, 2015])	24
2.11. La Arquitectura (Fuente [DWH, 2010])	26
2.12. Bodega de Datos (Fuente [Cano, 2007])	28
2.13. Tablas de Dimensión (Fuente [Kimball, 2002])	30
2.14. Tabla de Hechos (Fuente [Kimball, 2002])	31
2.15. Esquema Estrella de lo tickets (Fuente [Cano, 2007])	32
2.16. Esquema Copo de Nieve o Snowflake (fuente [Cano, 2007])	32
2.17. Esquema Constelación [Bernabeu, 2009]	33
2.18. Jerarquía de lugar (Fuente [Cano, 2007])	34
2.19. Herramienta de Análisis y visualización (Fuente [Cano, 2007])	35
2.20. Ejemplos Reportes (Fuente [Planeaux, 2007])	36
2.21. Ejemplos Indicadores (Fuente [Planeaux, 2007])	36
2.22. Pentaho Arquitectura (Fuente:[Pentaho, 2016])	38
2.23. Interfaz gráfica de Spoon	39
2.24. Tableau Server (Fuente[TableauPDF, 2016])	42
2.25. Arquitectura de Tableau Server (Fuente [TableauPDF, 2016])	43
3.1. Enfoque Inmon (Fuente [Kimball, 2010])	48
3.2. Enfoque R. Kimball (Fuentes [Kimball, 2010])	53
3.3. Ciclo de Vida de la Metodología de Ralph Kimball (Fuente [Kimball, 2002])	55
3.4. Ciclo de vida dimensional adaptado a inteligencia de negocio	64

4.1. Arquitectura de Diseño	67
4.2. Arquitectura de Diseño con sus herramientas	68
4.3. Modelo Dimensiona tipo estrella	70
4.4. Jerarquía de las dimensiones	70
4.5. Dimensión de Área de aplicación	73
4.6. Modelo definitivo tipo estrella	76
4.7. Dimensión Tiempo	77
4.8. Diseño lógico de las dimensiones	77
4.9. ETL de la Dimension Investigador	78
4.10. Modelo dimensional	79
4.11. ETL de Dimensión Investigador	79
4.12. ETL de los Hechos del nivel académico a la Tabla de Hecho de investigación	80
4.13. El llenado del almacén de datos	81
4.14. Verificación de la dimensión investigador	81
4.15. Verificación de la cantidad de los datos de la dimensión investigador . .	82
4.16. Verificación de la dimensión área de conocimiento	82
4.17. Verificación de la cantidad de datos de la dimensión área de conocimiento	82
4.18. Verificación de la dimensión área de aplicación	83
4.19. Verificación de la cantidad de datos de área de aplicación	83
4.20. Verificación de la dimensión institución	83
4.21. Verificación de la cantidad de datos de la dimensión institución	84
4.22. Verificación de la dimensión proyecto	84
4.23. Verificación de la cantidad de datos de la dimensión proyecto	84
4.24. Verificación de la dimensión producto	85
4.25. Verificación de la cantidad de datos de la dimensión producto	85
4.26. Verificación de la dimensión evento	85
4.27. Verificación de la cantidad de datos de la dimensión evento	86
4.28. Verificación de la cantidad de datos de la dimensión nivel académico . .	86
4.29. Verificación de la cantidad de datos de la dimensión nivel académico . .	86
4.30. Verificación de datos de la dimensión tiempo	87
4.31. Comparación de los dato del modelo con la dimensión tiempo	88
4.32. Proyecto en la base de datos	88
4.33. Dimensión proyecto con la tabla de hecho investigación	89
4.34. Dimensión área de conocimiento con la tabla de hecho investigación . .	89
4.35. Área de conocimiento en el almacén de datos	89
4.36. Subárea de aplicación en la base de datos	90
4.37. Dimensión área de aplicación con la tabla de hecho investigación	90
4.38. Dependencia en la base de datos	91
4.39. Dimensión institución con la tabla de hecho investigación	91
4.40. Usuario en la base de datos	91
4.41. Dimensión investigador con la tabla de hecho investigador	92
4.42. Creación de la conexión con la fuentes de dato desde Tableau	93
4.43. La tabla de hecho investigación con sus dimensiones	94

4.44. Elemento de almacén de datos	94
4.45. Conteo de proyectos del investigador	95
4.46. Conteo de proyectos por área de conocimiento	96
4.47. Gráfica de conteo de proyectos del investigador	96
4.48. Conteo de proyectos por área de aplicación	97
4.49. Gráfica de conteo de proyectos por área de aplicación	97
4.50. Conteo de proyectos por institución	98
4.51. Gráfica de conteo de proyectos por institución	98
4.53. Cantidad de productos por institución	99
4.52. Conteo de producto del investigador	99
4.54. Conteo de productos por área de conocimiento	100
4.55. Gráfica de conteo de productos por área de conocimiento	100
4.56. Conteo de productos por área de aplicación	101
4.57. Gráfica de conteo de productos por área de aplicación	101
4.58. Conteo de eventos por investigador	102
4.59. Los proyectos del investigador	102
4.60. Los proyectos del área de conocimiento	103
4.61. Los proyectos de las instituciones	103
4.62. Los productos del investigador	104
4.63. Los productos del área de aplicación	104
4.64. Evento de los investigador	105
4.65. El nivel académico de los investigador	105
4.66. Verificación de la tabla de hecho investigación con base de datos	106
4.67. Verificación de la tabla de hecho investigación con base de datos	106
4.68. Verificación de la tabla de hecho investigación con base de datos	107
4.69. Verificación de la tabla de hecho de nivel académico con base de datos	107

Índice de cuadros

3.1. Marco de trabajo de la Arquitectura	57
3.2. Comparación de las metodologías	63
4.1. Planificación del proyecto	66
4.2. Dimensión Tiempo	71
4.3. Detalle de la Dimensión Investigador	72
4.4. Dimensión de Institución	72
4.5. Dimensión de Área de Conocimiento	73
4.6. Dimensión de Proyecto	73
4.7. Dimensión de Producto	74
4.8. Dimensión de Evento	74
4.9. Dimensión Nivel Académico	75
4.10. Tabla de Hecho Gestión de los proyectos	75

Introducción

*“Las organizaciones pueden ser ricas en datos
y pobres en información...”
Stuart Madnick, 1993*

Trabajo de Grado se presenta un conjunto de conceptos, metodologías y herramientas que permiten llevar a cabo el desarrollo de una solución de Inteligencia de Negocio para la *Coordinación de Investigación* de la *Facultad de Ciencias de la Universidad Central de Venezuela*.

La *Facultad de Ciencias de la Universidad Central de Venezuela* está conformada por Escuelas e Institutos, y éstas a su vez por Centros y Laboratorios de Investigación, donde profesores y otros profesionales llevan a cabo investigaciones.

Para administrar las unidades de investigación de la Facultad, existe una *Coordinación de Investigación*. Esta coordinación dispone de una base de datos relacional, llamada Base de Datos del Sistema de Gestión de Información Científica (BD SIGIC), donde se centraliza la información de los investigadores y sus proyectos. BD SIGIC es usada para almacenar y procesar los datos, pero no tiene mecanismos efectivos para la interpretación de la información.

Para poder interpretar toda esa información, fue necesario extraerla del medio donde se encontraba, para llevarla a un almacén de datos (DWH) y luego procesarla mediante una herramienta de inteligencia de negocio.

La herramienta de inteligencia de negocio permite una gran flexibilidad y manipulación de los datos, dando la capacidad de hacer todo tipo de análisis y visualizaciones para la generación de conocimiento y ayudar así en la toma de decisiones gerenciales.

El presente trabajo es un prototipo debido a que la base de datos de la *Coordinación de Investigación* se encuentra con datos incompletos. Sin embargo, este prototipo es completamente funcional y está listo para cuando se introduzcan los datos. Adicionalmente el sistema es flexible y puede ser ampliado cuando se carguen los datos, no es una actividad de una sola vez sino un proceso continuo y de mejoramiento.

Este documento se encuentra organizado en cuatro (4) capítulos, y una sección de conclusiones y recomendaciones. Dichos capítulos se encuentran estructurados de la siguiente manera:

- **Capítulo 1 - Problema de Investigación.** Se expone la situación actual, el problema, los objetivos generales y específicos contemplados, la justificación del proyecto, la solución propuesta y el alcance definido para la investigación.

- **Capítulo 2 - Marco Conceptual.** Se describen los tópicos más relevantes que se encuentran relacionados con el tema de investigación, los aspectos contemplados se describen a continuación:
 - **Facultad de Ciencias.** Su estructura organizativa, la *Coordinación de Investigación*.
 - **Inteligencia de Negocio.** Definición, arquitectura de una solución de inteligencia de negocio.
 - **Almacén de Datos.** Definición, características y su arquitectura. Adicionalmente se define los conceptos concernientes al modelo dimensional.
 - **Indicadores de Gestión.** Definición, tipos de indicadores de gestión y los criterios de los indicadores de gestión.
 - **Herramientas Tecnológicas.** Componente para el desarrollo de la solución, herramienta ETL, herramienta para construir el almacén de datos y herramientas analíticas y de explotación.
- **Capítulo 3 - Marco Metodológico.** Se describen las distintas etapas que componen la metodología propuesta por *Ralph Kimball* para el desarrollo de una solución de inteligencia de negocio.
- **Capítulo 4 - Marco Aplicativo.** Se especifican las actividades realizadas para el desarrollo de una solución de inteligencia de negocio para la *Coordinación de Investigación* siguiendo las etapas de metodología definida por *Ralph Kimball* de manera no estricta.

Finalmente se presentan las conclusiones y recomendaciones del presente trabajo investigativo, así como también la bibliografía.

Capítulo 1

Problema de Investigación

En este Capítulo se presenta el problema que motiva este Trabajo Especial de Grado (TEG), una descripción general de las soluciones propuesta, alcance, objetivo general y objetivos específicos del TEG.

1.1. Situación Actual

En la Actualidad, la *Coordinación de Investigación*, tiene un *Sistema de Gestión de Información Científica* (SIGIC), la cual permite registrar y gestionar la información, actividades, proyectos y productos producidos por los investigadores de la *Facultad de Ciencias* de manera rápida y sencilla, en una base de datos centralizada, el cual será de gran utilidad para permitir el pleno funcionamiento de cualquier aplicación que requiera de esta información, pero no se cuenta con reportes y indicadores para el análisis de datos y apoyar la toma de decisiones por parte de la *Coordinación de Investigación* [Pérez, 2016].

Sin embargo se tiene la necesidad de visualizar ciertos reportes y indicadores de gestión para tener la información de una manera más resumida, es decir transformar los datos crudos en información para la toma de decisiones. Porque la información es uno de los recursos más importantes que tiene una organización siempre y cuando sea confiable y acertada, además es la base para plantear estrategias de negocio. Además la mayoría de las organizaciones tienden a tener un gran número de datos, pero muy poca información ya que no saben como resumir y categorizar los datos.

1.2. Planteamiento del Problema

La *Coordinación de Investigación de la Facultad de Ciencias de la Universidad Central de Venezuela* cuenta con una base de datos centralizada donde se almacena información de los proyectos de investigación.

A pesar de la cantidad de datos que se recolecta continuamente y del control, se tiene muy poca información de los proyectos, productos de investigación..

Dado a que usualmente se manejan los datos, se dificulta la tarea de obtener información útil para analizar y visualizar los reportes generados, los cuales requieren de mucho tiempo para su elaboración. Esta demora dificulta la toma de decisiones acertadas y oportunas.

La *Coordinación de Investigación* cuenta con una base de datos donde se almacena la información relacionada con las actividades de investigación de la Facultad, junto con una aplicación que permite de manera sencilla, registrar y administrar dicha información. Este tipo de almacenamiento tiene problemas en cuanto a mecanismos efectivos para la interpretación, análisis y visualización de los datos almacenados en ella. Es por ello que se crea un módulo de análisis de datos para toma de decisiones partiendo de las fuentes de datos y llevarla a un almacén de datos (Data Warehouse) y que esta sea visualizada mediante una herramienta de visualización y de exploración. Se desarrolla una capa de análisis de procesos, que podrá comunicarse directamente con los esquemas propios de la bases de datos de la *Coordinación de Investigación*. La idea es tener toda la información que servirá para generar una serie de responder y indicadores de gestión de los proyectos, productos, eventos, y nivel académico.

1.3. Solución Propuesta

Dados los inconvenientes mencionados, se necesita desarrollar una serie de reportes y indicadores de gestión relacionados con de los proyectos, productos, eventos, y nivel académico de los investigadores para la toma de decisiones, de manera consistente y oportuna.

Se propone una solución de Inteligencias de negocio (BI), cuya arquitectura se describe en figura 1.1, el cual parte de las fuentes de datos, se incorpora los datos mediante el proceso de extracción, transformación y carga de los datos (ETL) al almacén de datos, bodega de datos (Data Mart), y mediante las herramientas de análisis y de exploración se visualizara la información. Dicho prototipo se desarrolló para la *Coordinación de Investigación de la Facultad de Ciencias* y permite la obtención de reportes, indicadores para apoyar la toma de decisiones, ya que por cierta problemática no se logró emplear dicha solución.

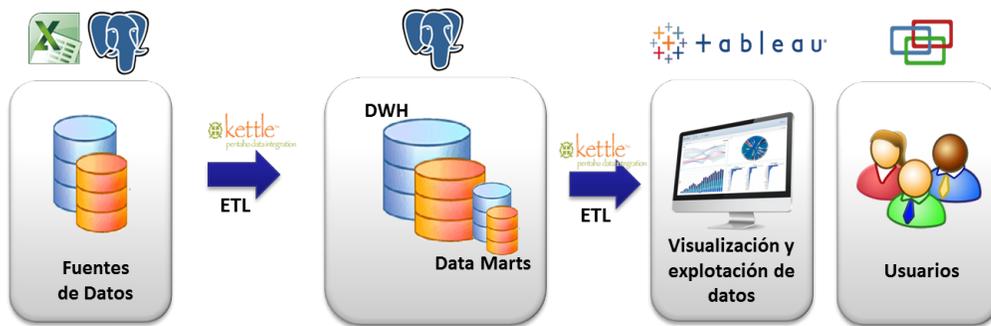


Figura 1.1: Arquitectura de la Solución Propuesta

- **Fuentes de datos.** Corresponde a la fuente de datos de la organización que sean relevantes para el BI.
- **ETL.** Encargada de la integración desde múltiples fuentes existentes para la construcción del almacén de datos.
- **Almacén de datos.** La colección de datos orientados a temáticas, integrados, no volátiles que cambian en el tiempo que apoya la toma de decisiones. Bodegas de Datos que se refiere a una vista del DWH orientado solo a un aspecto de la organización.
- **Visualización y explotación de datos.** Es la encargada de desplegar los análisis de la información.

La metodología que se va a emplear para desarrollar esta solución de Inteligencia de Negocio, es una adaptación la metodología Ascendente (Bottom-up) de Ralph Kimball, pero no se empleara de manera estricta.

1.3.1. Objetivo General

Desarrollar una solución de Inteligencia de Negocio para generar reportes y indicadores de investigación de la *Facultad de Ciencias*.

1.3.2. Objetivo Específicos

1. Analisis de los requerimientos, indicadores de investigación de la *Facultad de Ciencias*.
2. Analizar y definir las fuentes de datos que permitan alimentar al almacén de datos (DWH).
3. Diseñar el modelo dimensional.

4. Construcción del almacén de datos.
5. Desarrollar los reportes y indicadores de gestión para el monitoreo del DWH.
6. Realizar pruebas de integración, funcionales, calidad de datos, de aceptación.

1.4. Justificación

Se pretende desarrollar un prototipo de inteligencia de negocio que se adapte a los requerimientos de la coordinación, con respecto a los proyectos de investigación, que permita observar y evaluar la gestión de los procesos, para realizar el análisis basándose en los hechos del sistema.

Adicionalmente, se pueden hacer comparaciones con los datos históricos y usar éstos para hacer proyecciones en el tiempo y análisis, con la finalidad de solventar, permite corregir o mejorar el proceso.

El desarrollo de una solución de inteligencia de negocio, permitirá a la coordinación disponer de un sistema orientado a la toma de decisiones, utilizando tecnología de software libre como *PostgreSQL*, *Pentaho* y *Tableau* que un software con versión gratuita. Se tiene las herramientas necesarias para generar la solución de inteligencia de negocio orientado al análisis del negocio, y en consecuencia, apoyar la toma de decisiones de la *Coordinación de Investigación de la Facultad de Ciencias de la UCV*.

1.5. Alcance

Un prototipo que plantea ser de gran utilidad para el desarrollo de una solución Inteligencia de Negocio (BI) de la *Coordinación de Investigación*, para la toma de decisiones de una manera precisa y oportuna, ya que existe la necesidad de poder interpretar de una manera sencilla, versátil y rápida todos los datos que se generan a raíz de los procesos de negocios y así tomar decisiones efectivas que incrementen la productividad.

Se realizará el análisis de los proyectos de investigación, diseño e implementación de dicha solución que abarcará, desde diseño y extracción de los datos, además del almacén de datos (DWH), y la creación de los diferentes reportes y indicadores de gestión para la toma de decisiones. No es una actividad de una sola vez, es un proceso de mejoramiento continuo.

Capítulo 2

Marco Conceptual

En este capítulo se abordan todos los elementos a emplear como base teórica para el desarrollo de la solución de inteligencia de negocio. Es decir, se detallan los distintos conceptos, definiciones, arquitectura, esquemas, entre otros, que están relacionado: inteligencia de negocio, almacén de datos, Modelo Dimensional, Bodegas de Datos, Herramientas Tecnológicas.

2.1. Facultad de Ciencias

La *Facultad de Ciencias* es una de las facultades de la *Universidad Central de Venezuela (UCV)* que ofrece continuamente aportes a la Ciencia y a la Tecnología del país, en ella se realizan numerosas investigaciones en las distintas Escuelas e Institutos que la conforman. Dentro de la *Facultad de Ciencias* se encuentra la *Coordinación de Investigación*, encargada de apoyar y administrar los proyectos de investigación.

2.1.1. Estructura Organizativa

La *Facultad de Ciencias* está conformada por cinco escuelas y cuatro institutos. Las escuelas de la *Facultad de Ciencias* son biología, computación, física, matemáticas y química, y los cuatro institutos son ciencias de la tierra, zoología ecología tropical, ciencias de tecnologías de alimentos y biología experimental.

Es estas escuelas e institutos, grupos de investigadores, profesores y profesionales se dedican a la búsqueda y experimentación de diferentes líneas de investigación.

En los Centros de Investigación se realizan constantemente actividades de investigación científica y/o tecnológica, que favorezcan el desarrollo del país y el bienestar de la sociedad venezolana, además de: captación y entrenamiento de capital humano, transferencia de tecnología, difusión, divulgación científica y gestión, seguimiento y evaluación de procesos de ciencia y tecnología [Borges, 2006].

En los Laboratorios de Investigación se elaboran y ejecutan proyectos de investigación, se contribuye con la formación de investigadores, se organizan y/o dictan cursos, se brinda asesoramiento y se gestionan recursos económicos y materiales.

Normalmente, los Laboratorios de Investigación están adscritos a un Centro de Investigación, pero también pueden existir Laboratorios que no pertenezcan a ningún Centro. Esto ocurre porque no hay un Centro relacionado con la línea de investigación del Laboratorio y porque la cantidad de investigadores pertenecientes a dicho Laboratorio no es suficiente para crear un nuevo Centro y así poder cumplir con todas las formalidades necesarias.

En la figura 2.1 se puede ver un ejemplo de una escuela de la *Facultad de Ciencias* que tiene Laboratorios de Investigación que no están adscritos a un Centro de Investigación.

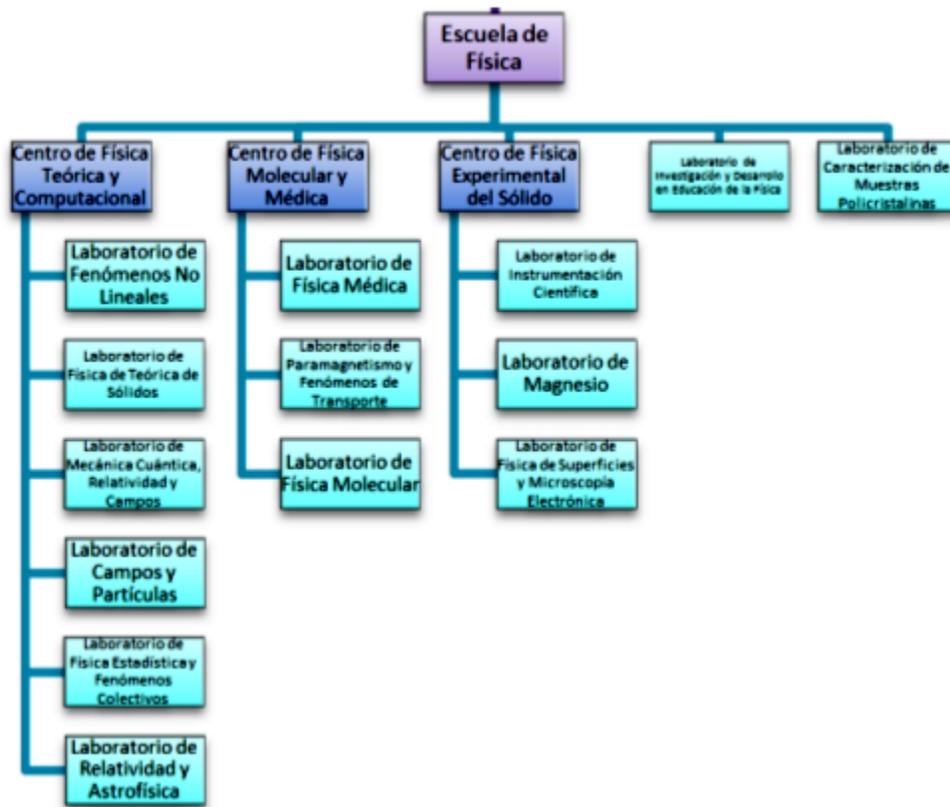


Figura 2.1: Estructura organizativa de investigación de la Escuela de Física de la *Facultad de Ciencias de la UCV*. (Fuente [CoorInv, 2011]).

2.1.2. Coordinación de Investigación

La *Coordinación de Investigación* de la *Facultad de Ciencias* es un ente administrativo que se ocupa de la promoción y el apoyo de actividades de investigación, así como también del fomento de relaciones con entes universitarios, gubernamentales y privados, dedicados al financiamiento de la ciencia. Fue creada por aprobación de la *Facultad de Ciencias*, en diciembre de 1991. Inicialmente existía una Comisión de In-

vestigación adscrita a la Coordinación de Postgrado formada por los Coordinadores de Investigación de las escuelas, Centros e Institutos de la *Facultad de Ciencias*, la cual había sido creada por el Consejo de Facultad en enero de 1986 [CoorInv, 2011].

La *Coordinación de Investigación* está constituida por un Coordinador y por un Consejo de Investigación. El Consejo está formado por el Coordinador de Investigación, quien lo preside, los representantes de cada uno de los Consejos Técnicos de los institutos, los representantes de las Comisiones de Investigación de cada una de las escuelas y representantes de la Facultad ante el Consejo de Desarrollo Científico y Humanístico (CDCH) [CoorInv, 2011].

Entre sus funciones esenciales está realizar el balance de la investigación que se realiza anualmente en la Facultad, analizar y promover el potencial productivo de la misma, dar a conocer a los diferentes entes del entorno nacional el resultado de las investigaciones y facilitar la conexión entre los investigadores y las instituciones externas que puedan promover financiamiento a la actividad de investigación [CoorInv, 2011].

El Consejo de Investigación tiene las siguientes atribuciones [CoorInv, 2011]:

- Proponer al Consejo de Facultad lineamientos generales de política para las actividades de investigación de la Facultad, así como las normas que obliguen a su cumplimiento.
- Velar por el uso adecuado de los recursos ordinarios y extraordinarios destinados a la investigación en la Facultad.
- Mantener una base de datos de los recursos de investigación en la *Facultad de Ciencias*.
- Proponer al Consejo de Facultad los Representantes de la *Coordinación de Investigación* ante organismos o entes del Gobierno.
- Proponer programas de investigación que se consideren necesarios a objeto de solucionar problemas prioritarios en nuestra sociedad.
- Velar porque las actividades de investigación de la Facultad reciban el apoyo de los recursos y hacer la distribución racional del mismo.
- Estudiar y evaluar las proposiciones de creación de Institutos y Centros de Investigación, Laboratorios de Investigación y Laboratorios de Apoyo a la Investigación dentro de la Facultad, y hacer las recomendaciones correspondientes.
- Estimular y orientar las solicitudes de fondos por parte de las diferentes estructuras organizativas de investigación de la Facultad ante entes intra y extra universitarios que la fomentan.
- Divulgar el potencial de investigación de la Facultad de Ciencias, así como los resultados de investigaciones ya realizadas, distinciones y/o premios.

- Fomentar la creación de Grupos Interdisciplinarios de Investigación, dentro y fuera de la Facultad.
- Propiciar la realización de eventos que beneficien las actividades de investigación de la Facultad tales como: Seminarios, Jornadas, Conferencias, Cursos y Reuniones.

2.2. Indicadores

Los Indicadores de Gestión son expresiones cualitativas o cuantitativas que describen características, comportamiento o fenómenos de una organización, permitiendo evaluar el desempeño y evolución en el tiempo, es decir es un instrumento para la evaluación, medición y análisis de las variables asociadas a las metas y objetivos de la organización. El valor de los indicadores expresan cualitativa o cuantitativa el desempeño de toda una organización o sus partes que permiten proponer una solución y toma acciones que conlleven a los objetivos planteados. Por medio de los indicadores podemos observar la evolución de una variable, establecimiento de una relación entre variables, comparación con períodos anteriores, productos similares o una meta.

2.2.1. Criterios de los Indicadores

Los indicadores para que sean útiles y efectivos tienen que cumplir los siguientes criterios:

- **Relevante.** Debe tener relación y estar alineado con los objetivos y metas de la organización.
- **Comparable.** Se debe poder comparar los valores a lo largo del tiempo en la organización e incluso entre otras organizaciones.
- **Verificable.** Se debe poder confirmar que los valores obtenidos sean correctos. (con respecto a lo que se desea medir)
- **Fácil de comprender.** Se debe mostrar de una forma clara y sencilla lo que se desea expresar como vemos 2.2.1.



Figura 2.2: Indicadores de Gestión (Fuente [Indicador,2014])

2.2.2. Objetivo de un indicador

A través de los indicadores puedes observar la evolución de una variable, establecer una relación entre variables es decir:

- Genera información útil para mejorar el proceso de toma de decisiones.
- Monitorear y evaluar el desempeño.
- Cuantificar los cambios.
- Identificar problemas y oportunidades.
- Medir comportamiento y facilitar la delegación.

2.2.3. Características

Las características de los indicadores “SMART” son:

- **Específicos (Specific)**. Permite describir la situación o fenómeno determinado.
- **Medible (Measurable)**. Debe ser cuantificable que permitan obtener unos resultados objetivos. Siguiendo el principio “lo que no se puede medir, no se puede controlar”.
- **Accionable (Actionable)**. Que los resultados y análisis permitan tomar decisiones.
- **Relevante (Relevance)**. Un objetivo este realmente orientado a la obtención de los resultados en un momento determinado.

- **Oportuno (Timely).** Estableciendo el período de tiempo en el que se debe completar cada uno de ellos.

2.2.4. Tipología de Indicadores

Existen muchos tipos de clasificaciones comunes en la teoría sobre indicadores:

- **Medición (Cuantitativos/ Cualitativos).** Es una representación numérica.
 - **Cuantitativos.** Su característica más importante es que, al encontrarse valores diferentes, estos pueden ordenarse de forma ascendente o descendente.
 - **Cualitativos.** Su características principal es que su resultado se refiere a una escala de cualidades.
 - **Calidad (Eficacia, Eficiencia, Efectividad).** Dan cuenta de la dinámica de actividades específicas.
 - **Eficacia.** Miden el grado de cumplimiento de los objetivos propuestos o metas programadas. Se concentra en **Qué** se debe hacer.

$$EficaciaDeObjetivos = \frac{ObjetivosAlcanzados}{ObjetivosPlanteados}$$
 - **Eficiencia.** Miden el nivel de ejecución del proceso con un mínimo de recursos, se concentran en el **Cómo** se hicieron las cosas.

$$EficienciaDeObjetivos = \frac{ObjetivosLogrados}{RecursosUtilizados}$$
 - **Efectividad.** Este concepto relaciona la eficiencia y la eficacia, es decir el logro de los resultados programados en el tiempo y con los costos mas razonables posibles.

$$EfectividadDeObjetivos = \frac{(PuntajeEficiencia+PuntajeEficacia)/2}{MáximoPuntaje}$$
- Los indicadores se clasifican según los factores claves de éxito:



Figura 2.3: Clasificación según los factores claves

- **Jerarquía de objetivos (Gestión/ Estratégicos).**
 - **Gestión.** Refleja cuáles fueron las consecuencias de acciones tomadas en el pasado en el marco de una organización. La idea es que estos indicadores

sienten bases para acciones o tomar en el presente y en el futuro. Es importante que los indicadores de gestión reflejen datos veraces y fiables, ya que el análisis de la situación, de otra manera, no será correcto.

2.3. Inteligencia de Negocio (Business Intelligence)

En la actualidad la información es uno de los recursos importantes que poseen las organizaciones, sin embargo, su obtención, análisis y manejo no siempre es tan sencillo. La información confiable y acertada es la base para la creación de un estrategia de negocio rentable y productiva.

Es por ello, que las distintas organizaciones han enfocado sus esfuerzos no solo en la recopilación de los datos para la generación de información, sino que también, en soluciones que permitan la visualización de reportes e indicadores que ayuden a realizar análisis para la toma de decisiones.

2.3.1. Inteligencia

“La inteligencia es información procesada y explotada a un muy alto nivel” [Prior, 2007], es decir, la inteligencia es la capacidad de abstracción para solucionar un problema y de organizar grandes cantidades de información para luego razonar y pensar sobre un aspecto en específico.

2.3.2. Negocio

Un negocio o una organización es “La acción y el efecto de articular, disponer y hacer operativos un conjunto de medios, factores o elementos para la consecución de un fin concreto” [Andrade, 2005], es decir, es un conjunto de personas que trabajan en conjunto desempeñando diversas actividades para lograr un objetivo en común (satisfacer las necesidades de la sociedad).

2.3.3. Definición de Inteligencia de Negocio

inteligencia de negocio por las siglas en ingles BI (Business Intelligence) por Howard Dresner 1989, cuando era consultor de Gartner, mediante el uso de tecnología y metodologías se quiere convertir datos en información y a partir de la información ser capaces de descubrir conocimientos basados en hechos. También puede intervenir en todos los procesos de una compañía, actuando en las tareas y actividades de los empleados, creando nuevas actividades y nuevas habilidades, mejorando la comunicación entre departamentos e incrementando la capacidad de reacción de la compañía. Para definir BI partiremos de la definición de Gartner ¹. [Cano, 2007]

¹Glosario de Gartner, www.gartner.com, enero 2006. Gartner es una consultora internacional especializada en Tecnologías de Información y Comunicación.

“BI es un proceso interactivo para explorar y analizar información estructurada sobre un área (normalmente almacenada en un datawarehouse), para descubrir tendencias o patrones, a partir de los cuales derivar ideas y extraer conclusiones. El proceso de Business Intelligence incluye la comunicación de los descubrimientos y efectuar los cambios.”

El origen de la Business Intelligence va ligado a proveer acceso directo a la información a los usuarios de negocio para ayudarles en la toma de decisiones. [Cano, 2007] Una solución de BI nos proporcionara los siguientes beneficios a la empresa.

2.3.3.1. Características

La Inteligencia de Negocio tiene características como [Cano, 2007]:

- **Proceso interactivo.** Hablar de BI estamos suponiendo que se trata de un análisis de información continuado en el tiempo, no sólo en un momento puntual. Aunque evidentemente este último tipo de análisis nos puede aportar valor, es incomparable con lo que nos puede aportar un proceso continuado de análisis de información, en el que por ejemplo podemos ver tendencias, cambios, variabilidades, etc.
- **Explorar.** En todo proyecto de BI hay un momento inicial en el que por primera vez accedemos a información que nos facilita su interpretación. En esta primera fase, lo que hacemos es “explorar” para comprender qué sucede en nuestro negocio; es posible incluso que descubramos nuevas relaciones que hasta el momento desconocíamos.
- **Analizar.** Pretendemos descubrir relaciones entre variables, tendencias, es decir, cuál puede ser la evolución de la variable, o patrones. Si un cliente tiene una serie de características, cuál es la probabilidad que otro con similares características actué igual que el anterior.
- **Área de análisis.** Todo proyecto de BI debe tener un objeto de análisis concreto. Nos podemos centrar en los clientes, los productos, los resultados de una localización, etc. que pretendemos analizar con detalle y con un objetivo concreto: por ejemplo, la reducción de costes, el incremento de ventas, el aumento de la participación de mercado, el ajuste de previsiones de venta, el cumplimiento los objetivos de venta presupuestados, etc.
- **Comunicar los resultados y efectuar los cambios.** Un objetivo fundamental del BI es que, una vez descubierto algo, sea comunicado a aquellas personas que tengan que realizar los cambios pertinentes en la organización para mejorar nuestra competitividad.

2.3.4. Niveles de Soluciones de Inteligencia de Negocio

La inteligencia de negocio muestra resultados, de acuerdo a las necesidades de los distintos niveles jerárquicos de la organización 2.4 se presenta una pirámide con los distintos tipos de BI, mostrando de izquierda a derecha el personal y a la derecha las herramientas o instrumentos que utilizan para realizar las tareas de la organización.



Figura 2.4: Niveles (Fuente [Inteligencia de Negocios, 2017])

Apoyan a los distintos niveles de la organización [Inteligencia de Negocios, 2017]:

- **Nivel Operativo.** Se permite a los empleados de la empresa recibir de forma oportuna, exacta y adecuada la información operativa, basándose en herramientas de trabajos como reportes, hojas de cálculos, manteniendo siempre un formato fijo cuya información se actualiza cada cierto tiempo.
- **Nivel Táctico.** Permite a los analistas de datos y a la gerencia media de una empresa, utilizar herramientas de análisis y consultas con el fin de obtener acceso a la información sin intervención de terceros.
- **Nivel Estratégico.** Permite que los directivos de la empresa pueda analizar y seguir día a día las tendencias, patrones, metas y objetivos estratégicos de la empresa. Un ejemplo de ello, lo constituye el cuadro de mando integral, entre otros.

2.3.5. Arquitectura de una Solución de Inteligencia de Negocio

Una solución de inteligencia de negocio tiende a ser una de las piezas más complejas de software que se puedan llegar a implementar en una organización, debido a que se requiere de la integración de diversos sistemas que usualmente no tienen una conexión directa entre sí. Por esta razón, vale destacar que una solución BI es mucho más que un software que despliega los datos de la empresa al usuario.

Una organización que maneja grandes cantidades de datos, necesita para maximizar sus ingresos e incrementar su eficiencia, monitorear una serie de indicadores claves que informen sobre el funcionamiento de la empresa en tiempo real.

Para poder generar conocimientos en una organización a partir de los datos que esta maneje, es necesario involucrar un conjunto de elementos mínimos indispensables, los cuales, permitirán aplicar inteligencia de negocio en las organizaciones que lo requieran.

Los diversos componentes que componen a una solución de inteligencia de negocio se ven expresados en la 2.5:

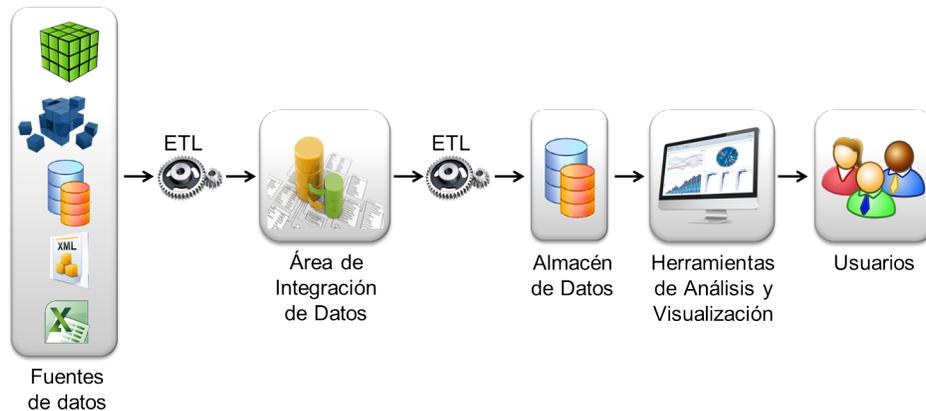


Figura 2.5: Componentes de inteligencia de negocio

- **Fuentes de Datos.** Es el proceso de donde se parte DWH para alimentar de información el DWH.
- **ETL Proceso de extracción, transformación y carga de los datos en el DWH.** Antes de almacenar los datos en un DWH, éstos deben ser transformados, limpiados, filtrados y redefinidos. Normalmente, la información que se tiene en los sistemas transaccionales no está preparada para la toma de decisiones.
- **Almacén de Datos (DWH). Metadata o Diccionario de datos.** Se busca almacenar los datos de una forma que maximice su flexibilidad, facilidad de acceso y administración.
- **Motor OLAP ².** Provee capacidad de cálculo, consultas, funciones de planeamiento, pronóstico y análisis de escenarios en grandes volúmenes de datos, sin embargo, en la actualidad existen otras alternativas tecnológicas al OLAP.
- **Las herramientas de acceso.** Permiten el análisis y la navegación de la información a través de los mismos.

²OLAP: On-Line Analytical Processing / Procesamiento Analítico en Línea

2.3.6. Fuentes de Datos

A la hora de desarrollar una un solución de BI, se debe determinar que datos se necesitan para realizar los análisis requeridos, por esta razón, un sistema de este tipo sin una fuente de datos correcta válida, tendría poco valor para la organización.

El conjunto de datos que se puede usar 2.6 para una solución de BI es muy amplia, se puede usar datos internos de la empresa, así como también datos externos que sirvan para efectos de los análisis a realizar.



Figura 2.6: Fuentes de Datos

En esta etapa deberíamos plantearnos una serie de preguntas, como: ¿Qué información necesita la organización?, ¿Cuándo se requiere dicha información?, ¿En que formato se necesita?, ¿Que información posee la empresa? y ¿Para quién está dirigida?, esto servirá para complementar los análisis para implementar la solución.

La mayoría de las implementaciones de BI usan snapshots (imágenes instantáneas de uno o más sistemas de archivos) o copias de seguridad como datos fuentes, debido a que esto garantiza que los datos en el sistema sean precisos y correctos.

Si se tomaran los datos de los sistemas transaccionales de la empresa a medida que se van generando, una actualización a un registro en el sistema transaccional generaría una pérdida de exactitud en los datos que se emplearán para los análisis, lo cual conlleva a que los resultados que se obtengan sean incorrectos o pocos precisos y adicionalmente, el rendimiento del sistema operacional bajaría considerablemente.

2.3.7. Procesos de Extracción, Transformación y Carga (ETL)

Una vez que ya hayamos seleccionado la data necesaria para el sistema procedemos ETL de los datos a la plataforma de BI. Es la encargada de la integración de los datos desde las multiples fuentes existentes para la construcción del almacén de datos (DWH) y los Data Marts, formando el ciclo de vida de una implementación de BI. Cada proceso se define como:



Figura 2.7: Procesos de Extracción, Transformación y Carga (ETL) (Fuente [Cano, 2007])

- **Extracción.** Consiste en la identificación y selección de los datos de los sistemas operacionales para satisfacer las necesidades de la organización. Una vez seleccionados los datos, se extraen los datos. Se considera la consolidación de los datos desde las diversas fuentes que usan diferentes estructuras y formatos, estableciendo el acceso a las fuentes de datos, extrayendo de manera eficiente cada una de los datos requeridos.
- **Transformación.** Los datos de la transacciones se convierten en datos con un formato consistente y orientado a los negocios, es decir, unos datos correctos, no ambiguos, consistentes y completos. Aplicando una serie de reglas o funciones a los datos extraídos desde la fuente para generar los datos a cargar al destino final.
- **Carga.** Se toman los datos y registran en el almacén de datos siguiendo las especificaciones que se definieron con anterioridad

2.3.8. Almacén de Datos (Data Warehouse)

El almacén de datos (DWH) ³ es una base de datos corporativa que proporciona múltiple funciones a un número determinado de usuarios, como el de disponer de Sistemas de Información de apoyo a la toma de decisiones.

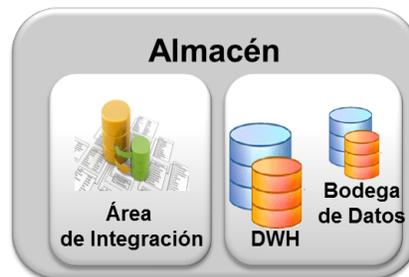


Figura 2.8: Almacén de Datos

La tecnología DWH se ha desarrollado rápidamente, y representa mejor que otros sistemas, la compleja estructura de una organización a la hora de administrar sus datos generacionales [Inmon, 2002].

³Data Warehouse siglas en inglés

2.3.8.1. Definición

Un DWH es una colección de información creada para soportar las aplicaciones de toma de decisiones.⁴ Es decir es una BD corporativa que abarca los procesos, las herramientas y las tecnologías para convertir datos en información, en conocimiento y planes para conducir de forma eficaz las actividades de negocios. [Cano, 2007]

El término DWH tiene sus orígenes en dos fuentes, “An Architecture for a Bussinesdd and Information Systems” de Barry Devlin y Paul Morphy⁵⁶ y ”Building the Data Warehouse” de Inmon, William H.⁷

- La primera fuente, “An Architecture for a Bussinesdd and Information Systems”, publicado en el IBM Systems Journal, es considerado como uno de los primeros artículos que describe la arquitectura de un DWH, desarrollado entre 1985 y 1986 para uso interno de IBM Europa en Dublín. En dicho artículo se hizo uso del término “Information Warehouse” para hacer referencia al proyecto. En la actualidad, suele usarse de manera general, para referirse a un DWH, la siguiente expresión que aparecía en el artículo: “El almacén de datos es un único, completo y consistente almacén de datos obtenido de una variedad de fuentes, y puesto a la disposición de los usuarios finales, de manera que ellos puedan entender y usar en el contexto del negocio” [Cano, 2007].
- La segunda fuente, ”Building the Data Warehouse”, describe al DWH como “un conjunto de datos orientados hacia una materia, integrados, no transitorios y que varían en el tiempo, los cuáles apoyan el tema de toma de decisiones de una administración” Inmon,1992. Inmon es considerado el padre del DWH y ha escrito una serie de libros, junto con otros autores, donde abordan diversos tópicos relacionados con la arquitectura almacén de datos [Cano, 2007].

2.3.8.2. Características

Las características principales del almacén de datos son las siguientes:

- **Integrado.** Los datos almacenados en el DWH deben estar integrados en una estructura que sea capaz de eliminar cualquier tipo de inconsistencias existentes entre los diversos sistemas operacionales que sirven de fuentes de datos. Al mismo tiempo, la información suele estructurarse en distintos niveles de detalle para adecuarse a los requerimientos preestablecidos por todos los usuarios finales.
- **Temático u Orientado a un área.** Sólo los datos necesarios para generar información del negocio se integran desde el entorno operacional. Los datos se organizan por temas, no por aplicación, así se facilita el acceso y entendimiento de los

⁴La primera aproximación es la del Watson, que define DWH

⁵Devlin, 1997

⁶“An Architecture for a Bussinesdd and Information Systems” de Barry Devlin y Paul Morphy

⁷”Building the Data Warehouse” de William H. Inmon

datos contenidos en el DWH. Eventualmente, esta característica permite realizar análisis y minería de datos.

- Histórico o Indexado en el tiempo (figura: 2.9).** El tiempo es parte implícita de la información contenida en un DWH. , en el DWH se carga con los distintos valores que puede tomar una variable en el tiempo, para poder hacer comparaciones y análisis, soportando de esta manera el proceso de toma de decisiones. Es decir, el DWH se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones. Los datos históricos son de poco uso en el procedimiento operacional. La información del depósito por el contrario, debe incluir los datos históricos para usarse en la identificación y evaluación de tendencias.



Figura 2.9: Histórico de DWH (Fuente [DWH, 2015])

- No volátil.** El DWH se construye para ser leído, y no modificado. La información que existe en un DWH es permanente, su actualización consiste en la incorporación de los últimos valores que tomaron las distintas variables sin alteración de los datos que ya existían se actualiza de manera periódica ya establecida.

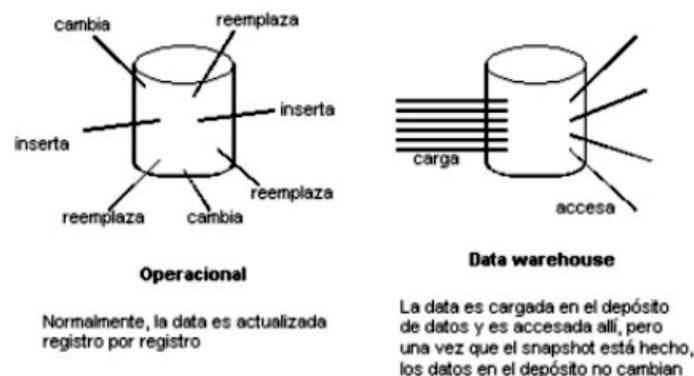


Figura 2.10: No Volátil (Fuente [DWH, 2015])

Como se puede observar en la figura: 2.10 en la BD operacional la actualización (actualizar, borrar y modificar) se hace regularmente, mientras en el DWH sea

una sola actualización esto hace que cuando tengamos que tomar una decisión con esta información tengamos seguridad de esta.

Además de las características previamente mencionadas, otras, quizás menos relevantes, pero que igualmente sirven para remarcar la potencialidad de un DWH:

- **Contiene datos diversos.** Es un repositorio unificado de información. Los datos de toda la organización, aunque pertenezcan a aplicaciones diferentes, son integrados en el DWH.
- **Optimizado para la consulta masiva.** El diseño físico de un DWH tiene un objetivo diferente al de las bases de datos transaccionales, mejorar los tiempos de respuesta de procesos de consultas de datos masivos de información, tomando en cuenta que deben ser orientados a temas, así como también brindar facilidad de entendimiento al usuario final.
- **La interfaz de usuario.** Está dirigida a los ejecutivos y diversos analistas, por lo que deben ser intuitivas.
- **Gran volumen.** Cuando se habla de almacén de datos, el espacio de almacenamiento suele medirse en Gigabytes y Terabytes debido a la cantidad de información sumariada. Cabe destacar que un DWH admite redundancia de datos y un tiempo de vida de información entre 5 y 10 años.

2.3.8.3. Arquitectura

Una arquitectura de almacén de datos es una forma de representar la estructura global de los datos, la comunicación, los procesos y la presentación al usuario final. La arquitectura está constituida por las siguientes partes interconectadas [DWH, 2010]:

Elementos que constituyen la arquitectura de un almacén de datos:

- **Nivel de base de datos externos, base de datos operacional.** Las organizaciones adquieren datos de Bases de Datos externas a la propia organización, que incluyen datos demográficos, económicos, datos sobre la competencia, etc. Mediante el proceso de DWH se extrae la información que está en la bases de datos operacionales y se mezcla con otras fuentes de datos. Enriquecemos la información.
- **Nivel de acceso a la información.** Es la capa con la que trata el usuario final. La información almacenada se convierte en información fácil y transparente para las herramientas que utilizan los usuarios. Se obtienen informes, gráficos, diagramas, etc.
- **Nivel de acceso a los datos.** Comunica el nivel de acceso a la información con el nivel operacional, es el responsable de la interfaz entre las herramientas de acceso a la información y las bases de datos.

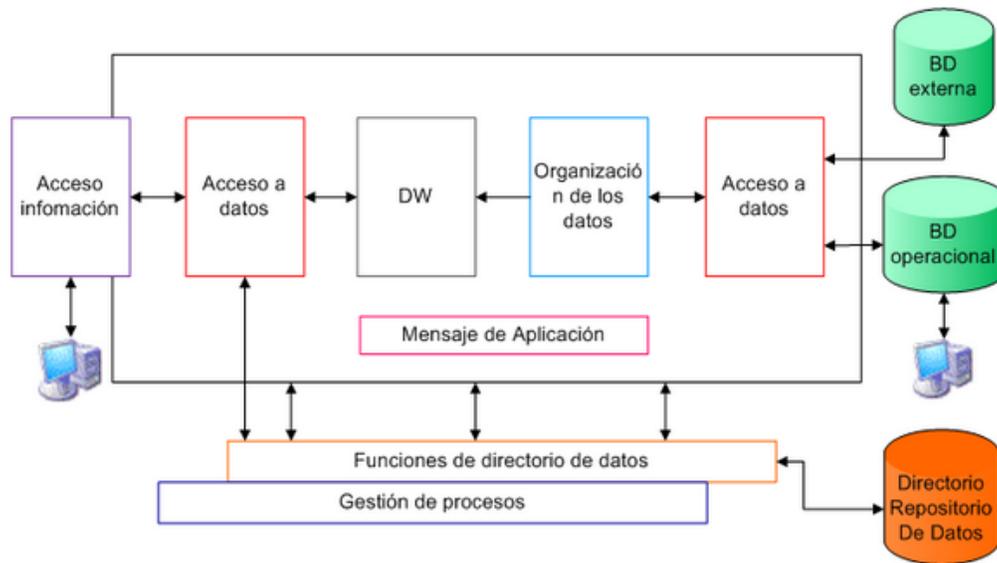


Figura 2.11: La Arquitectura (Fuente [DWH, 2010])

La clave de este nivel está en proveer al usuario de un acceso universal a los datos, es decir, que los usuarios sin tener en cuenta la ubicación de los datos o la herramienta de acceso a la información, deberían ser capaces de acceder a cualquier dato del DWH que les fuera necesario para realizar su trabajo.

- **Nivel de directorio de datos (metadatos).** Para proveer de un acceso universal, es absolutamente necesario mantener alguna clase de directorio de datos o repositorio de información de metadato que ayude a mantener un control sobre los datos. El metadato aporta información sobre los datos de la organización, de dónde proviene, qué formato tenía, cuál era su significado y si se trata de un agregado, cómo se ha calculado éste.

Para mantener un almacén completamente funcional, es necesario disponer de una amplia variedad de metadatos, información sobre las vistas de datos para los usuarios finales y sobre las bases de datos operacionales.
- **Nivel de gestión de proceso.** Este nivel tiene que ver con la planificación de las tareas que se deben realizar, no sólo para construir, sino también para mantener el DWH y la información del directorio de datos. Es o el controlador de alto nivel de los procesos que se han de llevar a cabo para que el almacén de datos permanezca actualizado.
- **Nivel de mensaje de la aplicación.** El nivel de mensaje de la aplicación tiene que ver con el transporte de información alrededor de la red de la empresa. Puede usarse para aislar aplicaciones operacionales o estratégicas a partir de un formato de datos exacto, recolectar transacciones o los mensajes y entregarlos a una ubicación segura en un tiempo específico.

- **Nivel Almacén de Datos (físico).** Es el repositorio central, altamente flexible de información, donde residen las copias de datos operacionales usados. En un DWH físico las copias de datos, operaciones y/o externos se almacenan de forma que sea fácil de acceder. En la actualidad un DWH se almacena en plataformas cliente/servidor, pero también existen configuraciones sobre mainframes, equipos externos para su rápido acceso mediante consulta.
- **Nivel de organización de datos.** El componente final de la arquitectura es la organización de los datos. Incluye todos los procesos necesarios para seleccionar, editar, resumir, combinar y cargar datos en el y para acceder a la información desde bases de datos operacionales y/o externas.

La forma de implementar un está sujeta a la forma en la que se va a estructurar el almacenamiento de los datos dentro del mismo. Independientemente del modelo que se escoja el objetivo principal es escoger uno que satisfaga las necesidades empresariales.

2.3.8.4. Metadata

La Metadata es el repositorio central de la información y es un componente crítico del DWH. Nos da el significado de cada uno de los componentes y sus atributos que se encuentran en el DWH. Puede incluir definiciones de negocio gestión del Almacén de Datos, descripciones detalladas de los tipos de datos, formatos y otras características [Cano, 2007]. Son los datos que describen la información, campos y formatos de BD y del DWH.

Para gestionar metadatos es necesario proporcionar una guía del punto de vista técnico y comercial de éstos. Los metadatos pueden ser categorizados en dos tipos:

- **Los metadatos técnicos.** Son metadatos que se crean durante el principio de un repositorio de datos , los cuales sirven para apoyar la gestión del repositorio. Éstos incluyen normas de adquisición, la transformación de los datos dentro del formato requerido por el repositorio, horarios para realizar copias de seguridad y la actualización de los datos.
- **Los metadatos de negocio.** Permiten al usuario final comprender información que se tiene guardada en el DWH y de qué manera se puede acceder a ella.

2.3.8.5. Requerimientos de un DWH

Se define los requerimientos que debería cumplir un DWH ⁸:

- **Fácil acceso.** Lograr que la información de la organización sea de fácil acceso. El alcance del Almacén de Datos puede ser bien un departamento o un corporativo. Aportando datos entendibles, manejables, rápidos de acceder y navegables, facilitándole al personal obtener dicha información para poder lograr un objetivo en específico.

⁸(Kimball & Margy, 2002), Define los objetivos de un DWH

- **Consistencia.** Lograr que la información de la organización sea consistente. Logrando que la información que se trabaja en la organización tenga el mismo nombre, significado, lógica, coherencia y solidez.
- **Adaptable y elástica.** Proporcionar información adaptable y elástica. La información en el DWH puede ser separada y combinada para analizar cada una de las posibles medidas del negocio. Tanto los datos, como la tecnología existente, no cambian en absoluto.
- **Seguridad.** Proporcionar un mecanismo de seguridad que protege los valores de la información. Permite al usuario final la manipulación de los datos, según la permisología asignada, logrando brindar seguridad en la información, ya que no todos los usuarios pueden modificar la información.
- **Fundamentar la toma de decisiones.** Almacenando los datos correctos, como el volumen de datos necesarios, para apoyar la toma de decisiones en la organización.

2.3.9. Bodega de Datos (Data Mart)

Los DWH representan una gran cantidad de información, es decir una gran base de datos que puede estar distribuida en distintas bases de datos, sería muy costoso y tiempo que ninguna organización podría aceptar. Es por ello que surge la bodega de datos, que está dirigido a un grupo de usuarios dentro de la organización.

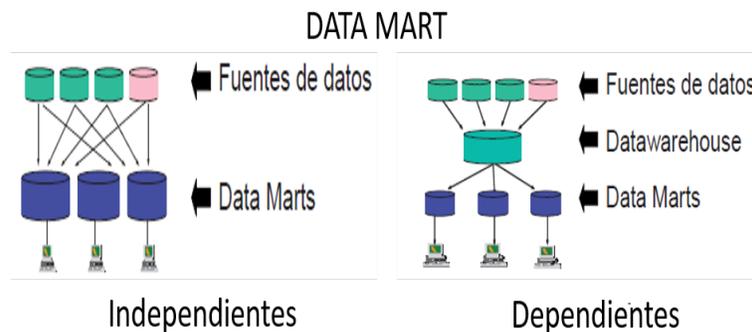


Figura 2.12: Bodega de Datos (Fuente [Cano, 2007])

Una bodega de datos es una bases 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 [sinnexus, 2016].

Los bodega de datos son más pequeños que los DWH y tienen menos información, menos modelos de negocios y son utilizados por un grupo menor de usuarios, pueden

ser dependientes e independientes. Un bodega de datos independiente puede originar inconsistencia con otro bodega de datos ver en la figura: 2.12.

2.3.9.1. Estrategias de Construcción

Dos estrategias básicas para la construcción de un DWH, una propuesta por Kimball y la otra por Inmon:

- W.H Inmon, Propone construir un DWH corporativo y a partir de él ir construyendo los modelos de análisis para los distintos niveles y departamentos de la organización; es decir, una estrategia de arriba abajo, desde la estrategia a lo más operativo se conoce como Descendente [Cano, 2007].
- R. Kimball, Propone construir distintos bodega de datos que cubran las distintas necesidades de la organización, sin la necesidad de construir un DWH se conoce como Ascendente [Cano, 2007].

Las dos estrategias son válidas.

De construir un DWH corporativo es desarrollo en fases es decir se construye un bodega de datos dependiente con una parte de información del DWH y en partes posteriores se van desarrollando bodega de datos usando subconjuntos del DWH. Igual que los proyectos complejos, es caro, y necesita mucho tiempo y es propenso al fracaso. Cuando tenemos éxito conseguimos un DWH integrado y escalable [Cano, 2007].

Si optamos por la más común, de construir distintos bodega de datos, el proyecto comienza con un bodega de datos único al que posteriormente se irán añadiendo otros bodega de datos que cubrirán otras áreas de negocio. Normalmente no requiere de grandes inversiones y es fácil de implementar, aunque con lleva algunos riesgos. Si seguimos esta estrategia debemos tener claro el plan de acción, es decir, qué áreas cubriremos y la integración de los distintos modelos. Esta estrategia se utiliza a veces como un paso previo al desarrollo de un DWH corporativo.

2.3.10. Modelo Dimensional del DWH

El modelo dimensional se describe en el año 1996 por Ralph Kimball, como propuesta para el diseño del Almacén de Datos, partiendo de la visión multidimensional que los usuarios tienen de los datos empresariales cuando se enfrentan a ellos con propósito de análisis (análisis multidimensional – OLAP – en concreto).

Todo DWH comienza con modelo dimensional, ya que se identifican los hechos, que es la tabla central de dicho modelo y las dimensiones son categorías que describen el contexto en el cual se analizan la tabla de hecho. Existen dos tipos de esquema en el modelo relacional: esquema estrella y copo de nieve.

2.3.10.1. Definición

El modelo dimensional es una técnica de diseño para el DWH (a nivel lógico) que pretende representar los hechos del negocio. Este modelo está optimizado para llevar a cabo consultas con un alto rendimiento.

El modelo dimensional es una de las técnicas de modelado de un Almacén de Datos. En el proceso de modelado dimensional, un modelo de tablas y relaciones se constituye con el propósito de optimizar el soporte de toma de decisiones a nivel de desempeño del query en bases de datos Relacionales relativo al proceso de negocio que se esté modelando. En contraste, los modelos convencionales de Entidad-Relación se constituyen para eliminar la redundancia en el modelo de datos, facilitar la obtención de registros en base a indicadores establecidos y por lo tanto optimizar el nivel de desempeño de los sistemas OLTP.

Por tal motivo, es necesario definir y tener claros los conceptos básicos al modelo dimensional, los tipos de tablas involucradas, y los esquemas de trabajo que pueden emplearse en una solución.

La Dimensión es una entidad independiente en el modelo dimensional que sirve como un punto de entrada o como un mecanismo de reordenamiento y fraccionamiento de las medidas sumariadas, localizadas en la tabla de hechos del modelo [Kimball, 2002].

2.3.10.2. Tabla de Dimensión

Las tablas de dimensiones contiene descripciones textuales, tablas que integran la tabla de hechos, cada dimensión se define por su clave primaria única, designada por la notación PK como vemos en la figura: 2.13, que sirve como base para la integridad referencial con cualquier tabla de hechos a la que se une y una columnas de atributos descriptivos que juegan un papel fundamental en el DWH son lo que queremos medir.

Product Dimension Table
Product Key (PK)
Product Description
SKU Number (Natural Key)
Brand Description
Category Description
Department Description
Package Type Description
Package Size
Fat Content Description
Diet Type Description
Weight
Weight Units of Measure
Storage Type
Shelf Life Type
Shelf Width
Shelf Height
Shelf Depth
... and many more

Figura 2.13: Tablas de Dimensión (Fuente [Kimball, 2002])

2.3.10.3. Tabla de Hechos

Es la representación en el DWH de los procesos de negocio de la organización. La tabla principal del DWH con las mediciones de rendimiento numéricas que son caracterizadas por una clave compuesta, donde cada elemento de la misma es una clave

foránea que corresponde a una tabla de dimensiones se encuentra todo lo que queremos medir, y analizar. Una medida se toma en la intersección de todas las dimensiones (fecha, producto y la tienda) como podemos ver en la figura:2.14. Una fila de una tabla de hechos corresponde a la medida. Todas la medidas de una tabla de hechos deben estar al mismo grano o granularidad [Kimball, 2002].

Daily Sales Fact Table
Date Key (FK)
Product Key (FK)
Store Key (FK)
Quantity Sold
Dollar Sales Amount

Figura 2.14: Tabla de Hechos (Fuente [Kimball, 2002])

2.3.10.4. Esquema Estrella (Star Scheme)

El esquema de estrella es la representación genérica de un modelo dimensional en una base de datos relacional, en la cual una tabla de hechos con una clave compuesta que forma la clave de cada dimensión, es unida a un número de tablas de dimensiones (Empleado, tiempo, centro, etc.), cada una con una clave primaria simple [Kimball, 2002]. El esquema estrella es ideal por su simplicidad y simetría porque son más fáciles de entender, navegar y es altamente reconocible por los usuarios de negocios como podemos ver en la figura 2.15.

Las características del esquema estrella [Cano, 2007]:

- Una tabla de hechos, que contiene los datos sin redundancias.
- Una sola tabla por dimensión.
- La tabla de hechos (Fact table) tiene un atributo columna que forma la clave de cada dimensión.
- Cada tabla de dimensión (Dimension table) es una tabla simple desnormalizada.

Cuando unimos distintos esquemas “estrella” que tienen distintas tablas de hechos, pero comparten las de las dimensiones, hablamos de constelaciones de hechos; algunos autores hablan incluso de esquema “galaxia” [Cano, 2007].

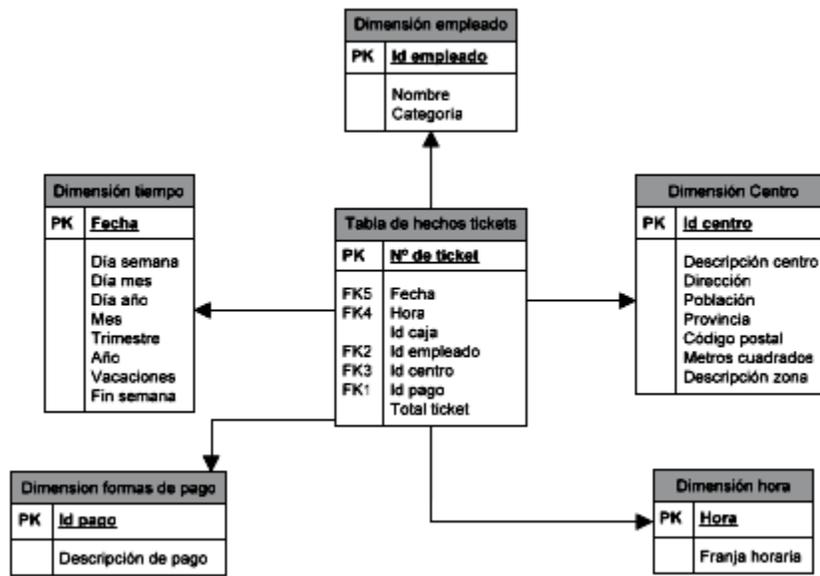


Figura 2.15: Esquema Estrella de lo tickets (Fuente [Cano, 2007])

2.3.10.5. Esquema Copo de Nieve (Snowflake Scheme)

El esquema de copo de nieve es una dimensión normalizada en el cual una tabla de dimensión simple es descompuesta en una estructura de árbol con muchos niveles [Kimball, 2002].

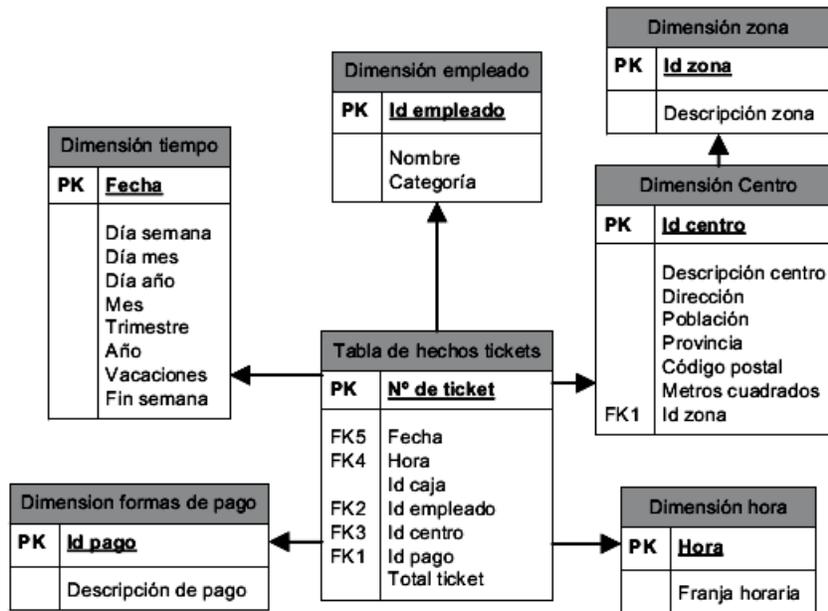


Figura 2.16: Esquema Copo de Nieve o Snowflake (fuente [Cano, 2007])

El esquema “estrella” no es totalmente normalizado, como podemos ver en la figura: 2.15 ya que en la tabla de la “Dimensión Centro” tenemos una redundancia que es “Descripción zona”, se repetirá tantas veces la zona como centros existan en la misma. Es por eso que nace el esquema copo de nieve o Snowflake para solucionar este problema. Como vemos en la figura: 2.16, en el esquema “copo de nieve” aparecen relaciones entre las tablas de dimensiones, mientras que en el esquema “estrella” sólo hay relaciones entre la tabla de hechos y las de dimensiones [Cano, 2007].

2.3.10.6. Esquema Constelación (Starflake Scheme)

El esquema constelación está compuesto por una serie de esquema estrellas y tal como se puede apreciar en la siguiente figura 2.17, está formado por una tabla de hechos principal (“HECHOS_A”) y por una o más tablas de hechos auxiliares (“HECHOS_B”), las cuales pueden ser sumalizaciones de la principal. Dichas tablas yacen en el centro del modelo y están relacionadas con sus respectivas tablas de dimensiones [Bernabeu, 2009].

No es necesario que las diferentes tablas de hechos compartan las mismas tablas de dimensiones, ya que, las tablas de hechos auxiliares pueden vincularse con solo algunas de las tablas de dimensiones asignadas a la tabla de hechos principal, y también pueden hacerlo con nuevas tablas de dimensiones [Bernabeu, 2009].

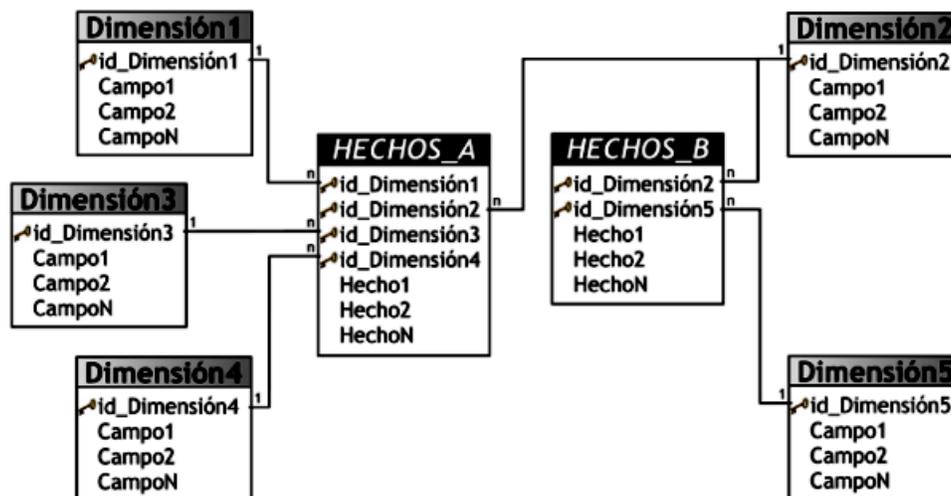


Figura 2.17: Esquema Constelación [Bernabeu, 2009]

Las características del esquema constelación:

Su diseño y cualidades son muy similares a las del esquema en estrella, pero posee una serie de diferencias con el mismo, que son precisamente las que lo destacan y caracterizan. Entre ellas se pueden mencionar [Bernabeu, 2009]:

- Permite tener más de una tabla de hechos, por lo cual se podrán analizar más aspectos claves del negocio con un mínimo esfuerzo adicional de diseño.

- Contribuye a la reutilización de las tablas de dimensiones, ya que una misma tabla de dimensión puede utilizarse para varias tablas de hechos.
- No es soportado por todas las herramientas de consulta y análisis.

2.3.10.7. Granularidad

Presenta el nivel de detalle al que se desea almacenar la información que posee cada registro de una tabla de hechos en un DWH [Kimball, 2002]. Es decir deberemos decidir cuál es el nivel de granularidad necesario para poder construir un modelo que nos permita responder a aquellas preguntas que nos hemos formulado a determinar un nivel de granularidad podemos responder unas preguntas pero no otras. Por ejemplo, los datos referentes a ventas o compras realizadas por una empresa, pueden registrarse día a día, en cambio, los datos pertinentes a pagos de sueldos o cuotas de socios, podrán almacenarse a nivel de mes.

2.3.10.8. Jerarquía

Es una serie de relaciones en cascada de uno a muchos [Kimball, 2002]. Después de que ya se sabe cuál sería la granularidad, se realiza la jerarquía. Esta jerarquía corresponde con las tablas de dimensión que se definen como los niveles de asociación que tienen de los datos. Una dimensión debe contener al menos una jerarquía, la cual puede tener varios niveles.

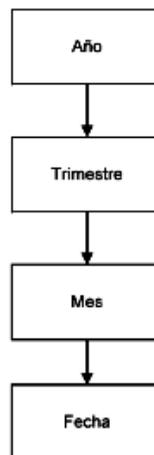


Figura 2.18: Jerarquía de lugar (Fuente [Cano, 2007])

Como por ejemplo la dimensión tiempo, una jerarquía del tiempo como podemos ver en la figura: 2.18. Años se pueden descomponer en trimestres, los trimestres en meses y los meses en días. La existencia de las jerarquías en las dimensiones nos permite pasar del máximo detalle a la agregación en los distintos niveles. En nuestro ejemplo

podemos por tanto analizar las ventas de un artículo por días, por meses, por trimestres o por años.

2.3.10.9. Agregación

Se trata de una medida, es decir un dato contable. Por ejemplo, si se quiere tener el total del salario de una empresa, lo que se hace es hacerle un proceso de agregación, definiendo como operación la suma de los sueldos de los empleados, logrando que cada vez que entre más información en el atributo sueldo, se va a ir mostrando sumariado.

2.4. Herramientas Tecnológicas

El proceso de selección de una herramienta tecnológica puede llegar a ser complejo debido a las múltiples alternativas existentes en el mercado (Oracle BI, Pentaho BI, IBM Cognos, MicroStrategy, entre otros) y a la diversa gama de funcionalidades que cada herramienta brinda.

Independientemente de la herramienta a seleccionar, ésta debe proveer un repositorio centralizado, visualización de cuadros de mando y la posibilidad de construir consultas a la medida.

2.4.1. Herramientas de Análisis y visualización

Con el surgimiento de aplicaciones de inteligencia de negocio, se han creado herramientas especializadas que permiten mostrarle a la alta gerencia los datos más relevantes de una empresa de manera resumida, en forma de indicadores y reportes para la toma decisiones.



Figura 2.19: Herramienta de Análisis y visualización (Fuente [Cano, 2007])

2.4.1.1. Reportes

Es un informe generado por un sistema, que nos muestra de forma estructurada y resumida, datos relevantes generados por las aplicaciones, de manera que se pueda realizar la toma de decisiones.

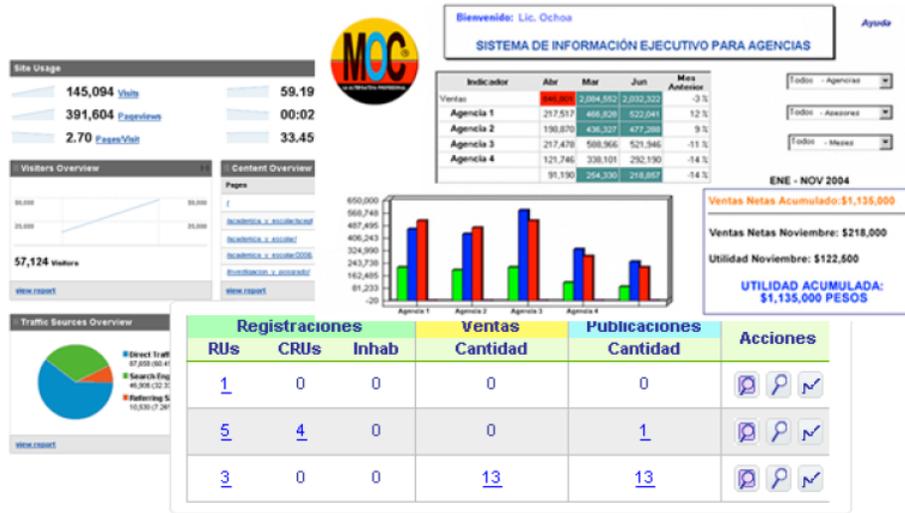


Figura 2.20: Ejemplos Reportes (Fuente [Planeaux, 2007])

2.4.1.2. Indicadores

Es una medida cuantitativa que permite identificar cambios en el tiempo, cuyo objetivo es determinar el buen funcionamiento de un sistema o negocio. Logrando así identificar los problemas y poder tomar medidas para solucionarlos.

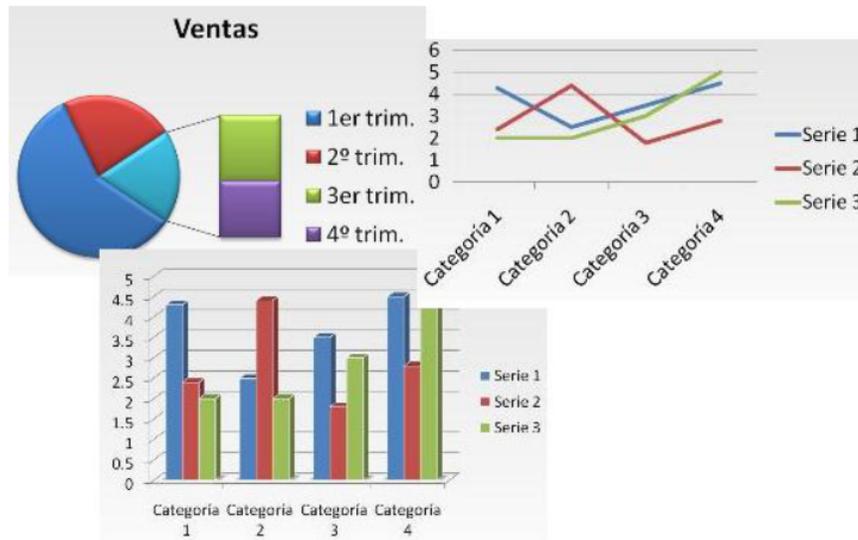


Figura 2.21: Ejemplos Indicadores (Fuente [Planeaux, 2007])

2.4.2. Herramienta de Extracción, Transformación y Carga (ETL)

Las herramientas de Extracción (E), Transformación (T) y Carga (L – de Load en inglés) permiten de manera sencilla recuperar datos de distintas fuentes de información, limpiarlos y realizar las transformaciones necesarias para cargarlos en un repositorio, minimizando fallos comunes como la existencia de campos o valores nulos, tablas de referencia inexistentes, entre otros. Siguiendo la línea de tecnologías de código abierto, se ha seleccionado la herramienta de ETL de Pentaho, llamada Pentaho Data Integration (PDI) o también conocida como Kettle, la cual se explica a continuación.

2.4.2.1. Pentaho

Es una plataforma Open Source de BI orientada a soluciones y centrada en procesos fundada en el 2004. Ofreciendo soluciones para la gestión y análisis de la información, incluyendo el análisis multidimensional OLAP, presentación de informes, minería de datos y creación de cuadros de mando para el usuario.

La plataforma ha sido desarrollada bajo el lenguaje de programación Java y tiene un ambiente de implementación también basado en Java, haciendo así que Pentaho sea una solución muy flexible al cubrir una alta gama de necesidades empresariales.

Arquitectura Los módulos de la arquitectura de Pentaho BI son:

- **Pentaho Data Integration.** Permite tomar información de diferentes fuentes y cargarlas en un repositorio analítico. Provee una consistencia, una sola versión de todos los recursos de información mediante ETL ⁹.
- **Pentaho reporting.** Es un potente generador de informe permitiendo la distribución de los resultados del análisis en múltiples formatos como PDF, XLS, HTML y textos. Los reportes Pentaho permiten también programación de tareas y ejecución automática de informes con una determinada periodicidad.
- **Pentaho Analysis.** Suministra a los usuarios un sistema avanzado de análisis de información es un servidor OLAP¹⁰. Es compatible con expresiones multidimensionales y el lenguaje de consulta XML para el análisis
- **Pentaho Dashboards.** Proporcionar información sobre sus datos, donde se pueden ver informes, gráficos interactivos y los cubos creados con las herramientas Pentaho Report Designer.
- **Pentaho Data Minina.** Es el proceso de correr datos en algoritmos completamente sofisticados, relevando significantes patrones y correlaciones que pueden estar escondidos median la herramienta Weka. Esto puede ser usado para ayudar a entender lo mejor para el negocio y explotar el rendimiento de este en un futuro prediciendo completamente en el análisis.

⁹ETL (Extracción, Transformación y Carga)

¹⁰OLAP (Procesamiento analítico en línea)

- **Pentaho para Apache Hadoop.** Es una herramientas de desarrollo visual fáciles de usar y grandes análisis de datos que permiten a los usuarios preparar, modelar, visualizar y explorar conjuntos de datos estructurados y no estructurados en Hadoop. Pentaho simplifica el ciclo de vida de los datos de Hadoop de extremo a extremo proporcionando una plataforma completa desde la preparación de datos hasta el análisis predictivo. Pentaho es único al proporcionar ejecución en Hadoop para un rendimiento extremadamente rápido.[?]

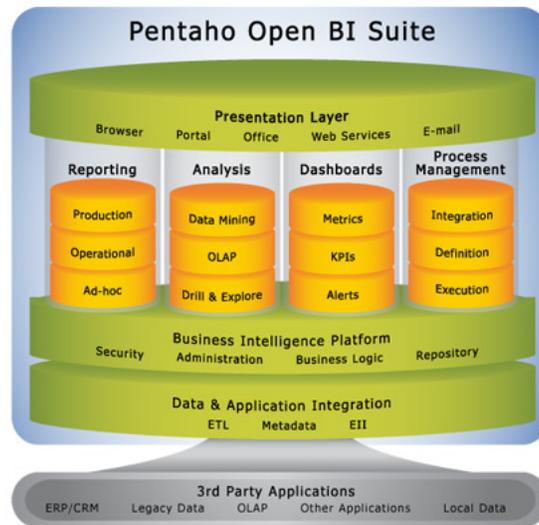


Figura 2.22: Pentaho Arquitectura (Fuente:[Pentaho, 2016])

2.4.2.2. Pentaho Data Integration

Pentaho Data Integration (PDI) es una herramienta que permite extraer, transformar y cargar (ETL- Extract, Transform and load) la información disponible en aplicaciones y bases de datos separadas y Hojas de Cálculos para ponerlas en manos del usuario, proyectando consistencia. También es conocido como kettle y posee las siguientes aplicaciones:

- **Spoon.** Herramienta gráfica que permite diseñar procesos ETL. Soporta conexión con diversas fuentes de datos y permiten transformar los datos necesarios para cargarlos dentro de la Base Datos destino.
- **Pan.** Herramienta que permite ejecutar transformar diseñar con Spoon en XML o en un repositorio de base de datos. Generalmente las transformaciones se programan en modo por lotes para ser ejecutadas en intervalos de tiempo regulares.
- **Chef.** Permite ejecutar trabajos complejos que automatizan los procesos de actualización de la base de datos.

- **Kitchen.** Herramienta que ayuda a ejecutar el trabajo por lotes, permitiendo iniciar y controlar fácilmente procesos ETL.
- **Carte.** Servidor web que permite la supervisión remota de procesos ETL.

De las aplicaciones nombrada anteriormente una de las más usadas es Spoon, cuya interfaz se puede ver . Spoon es fácil de usar y resulta muy útil incluso para realizar migraciones pequeña, como pasar datos de una Hoja de Cálculo a una base de datos.

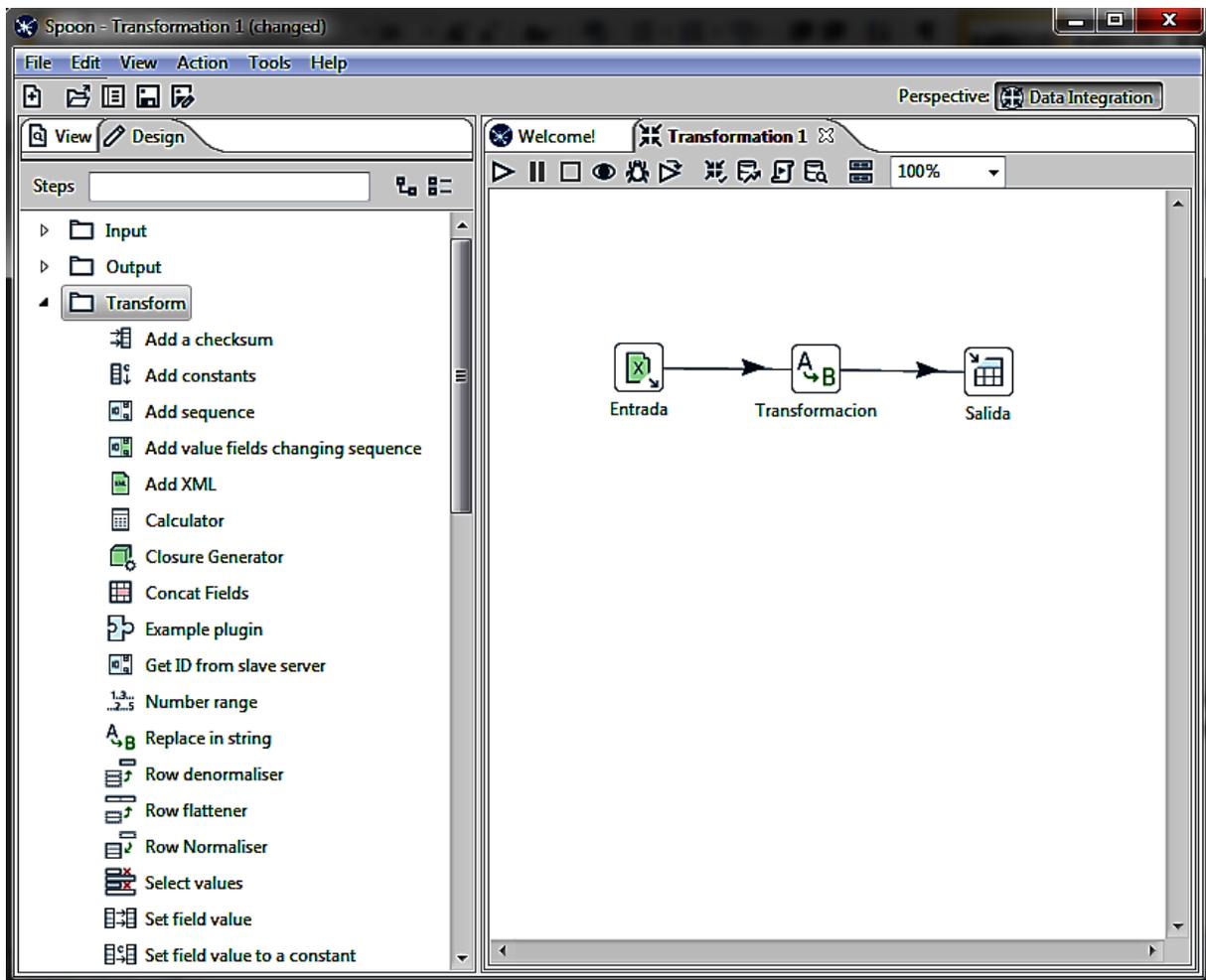


Figura 2.23: Interfaz gráfica de Spoon

2.4.3. Herramienta para crear la bodega de datos (Datamart)

Un bodega de datos es una bases 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 es decir donde los datos están denormalizados basandose en la información que necesite un departamento. Siguiendo la línea de tecnologías de código abierto, se ha seleccionado la herramienta para crear la bodega de datos.

2.4.3.1. PostgreSQL

PostgreSQL es un Sistema Manejador de base de datos objeto-relacional, de código abierto, que cuenta con más de 15 años de desarrollo activo y una arquitectura probada que se ha ganado una sólida reputación por su fiabilidad, integridad de datos y corrección. Además, permite desarrollar procedimientos en diferentes lenguajes de programación como Java, C++, Ruby y Python, lo que hace a PostgreSQL altamente personalizable[PostgreSQL, 1996].

Soporta grandes cantidades de datos y una alta concurrencia de usuarios accediendo a la vez al sistema, por lo que se considera una herramienta que favorece a los usuarios con sistemas empresariales de gran tamaño.

PostgreSQL se ha enfocado tradicionalmente en la fiabilidad, integridad de datos y características integradas enfocadas al desarrollador. Tiene un planificador de consultas extremadamente sofisticado, que es capaz de unir cantidades relativamente grandes de tablas eficientemente [PostgreSQL, 2001].

Se distribuye bajo la Licencia PostgreSQL, que es una licencia similar a la de la Distribución de Software de Berkeley (BSD) y a la del Instituto de Tecnología de Massachusetts (MIT), que permite a los usuarios hacer cualquier cosa que quieran con el código, incluyendo la reventa de los binarios sin el código [PostgreSQL, 2001].

PostgreSQL tiene las siguientes ventajas:

- Es código abierto.
- La velocidad de respuesta se mantiene al aumentar el tamaño de la base de datos, cosa que no sucede con otros programas que se suelen poner lentos.
- Proporciona estabilidad y confiabilidad.
- Tiene una gran capacidad de almacenamiento.
- Soporta gran número de peticiones simultáneas a la base de datos de forma correcta.
- Puede operar sobre distintas plataformas como Linux, Windows, Unix, Solaris y MacOS X.
- Provee un buen sistema de seguridad mediante la gestión de usuarios, grupos de usuarios y contraseñas.
- Soporta los tipos de datos, cláusulas, funciones y comandos de tipo estándar SQL92/SQL99 y extendidos propios de PostgreSQL.

También presenta las siguientes desventajas:

- En comparación con otros Sistemas Manejadores de Base de Datos, como por ejemplo MySQL, es más lento en inserciones y actualizaciones, ya que cuenta con cabeceras de intersección.
- Cuenta con muchos foros oficiales de ayuda, pero no con una documentación de ayuda obligatoria.
- La sintaxis de algunos comandos o sentencias no es tan intuitiva.

2.4.4. Herramientas de Visualización

Se presenta la siguiente herramienta de visualización, para observar de manera más reducida los grandes volúmenes de datos de forma más flexible y que apoye la toma de decisiones.

2.4.4.1. Tableau

Tableau es una herramienta o software de inteligencia de negocio¹¹ que permite visualizar de una manera reducida grandes volúmenes de información, en forma rápida, flexible y gran usabilidad¹². Tableau es en la actualidad la herramienta de BI líder y de mayor velocidad de crecimiento según Gartner, destacando por su facilidad de uso, potencialidad para generar visualizaciones y capacidad de manejo de grandes volúmenes de Datos. A diferencia de las herramientas tradicionales de inteligencia de negocio (BI) desarrolladas pensando en el usuario técnico del área de sistemas, Tableau está orientado a que personas de todos los ámbitos puedan manejar información fácilmente y presentarla en forma atractiva. Así, abogados, periodistas, ingenieros, médicos, entre otros, que trabajen en una organización o en forma independiente, encontrarán en Tableau un poderoso aliado analítico.

Tableau Software se fundó sobre la idea de que el análisis de datos y los informes subsiguientes no deben ser actividades aisladas, sino que deben integrarse en un proceso único de análisis visual: uno que les permita a los usuarios ver rápidamente patrones en sus datos y cambiar las vistas al instante para seguir su línea de pensamiento. Tableau combina la exploración de los datos y la visualización de estos en una aplicación fácil de usar que todos pueden aprender rápidamente. Cualquier persona acostumbrada al uso de Excel puede crear análisis interactivos y enriquecidos, eficaces para compartílos de manera segura en la empresa. Los equipos de TI pueden administrar los datos y metadatos de manera centralizada, controlar los permisos y escalar hasta implementaciones en toda la empresa. Esta descripción general se diseñó para responder las preguntas comunes de los gerentes y administradores de TI y ayudarles a admitir las implementaciones de software de análisis visual de cualquier tamaño.

¹¹Inteligencia de Negocio es poner la información adecuada, en las manos de las personas que toman decisiones

¹²Usabilidad Cualidad de la página web o del programa informático que son sencillos de usar

Arquitectura Tableau Server cuenta con una arquitectura de cliente-servidor de n niveles altamente escalable que presta servicios a clientes móviles, clientes web y software instalado en equipos de escritorio. Las soluciones de Tableau tienen 2 componentes principales: Tableau Desktop y Tableau Server.[TableauPDF, 2016]

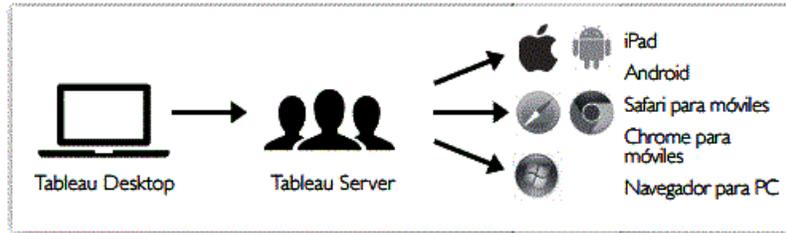


Figura 2.24: Tableau Server (Fuente[TableauPDF, 2016])

Tableau Server es una plataforma de análisis de negocios de clase empresarial que puede escalar verticalmente hasta cientos de miles de usuarios. Ofrece poderosos análisis móviles y basados en navegador, y funciona con la arquitectura de datos, la administración del ciclo de vida y las restricciones de seguridad y gestión existentes en la empresa.[TableauPDF, 2016]

Tableau Server cumple con requisitos empresariales, entre los que se incluyen [TableauPDF, 2016]:

- Escalabilidad Tableau Server puede escalar vertical y horizontalmente para satisfacer las necesidades de su empresa. El servidor puede escalar verticalmente con la incorporación de CPU y RAM adicionales. Todos los componentes de Tableau Server son de multiproceso y pueden configurarse en función de sus patrones de uso. Se puede escalar aún más agregando nodos adicionales que pueden configurarse de manera que cumplan con los requisitos de la organización.
- Alta disponibilidad con administración de clústeres interna y admite equilibradores de carga externos.
- Seguridad, cifra el tráfico interno, admite integración con Active Directory, SAML y OAuth.
- Facilidad de administración directa, desde la gestión de usuarios hasta las actualizaciones.
- Extensibilidad ofrece API eficaces.

El siguiente diagrama muestra la arquitectura de Tableau Server[TableauPDF, 2016] :

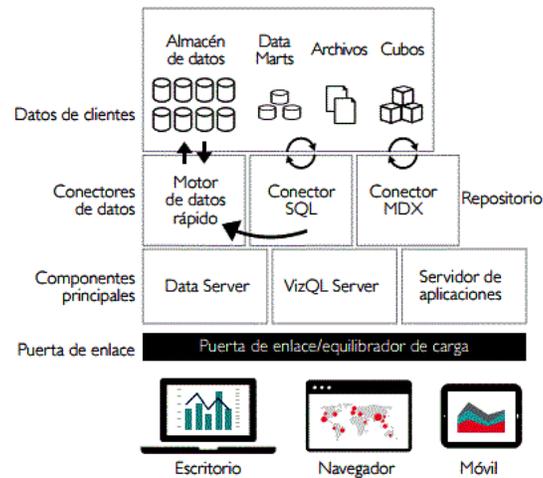


Figura 2.25: Arquitectura de Tableau Server (Fuente [TableauPDF, 2016])

- **Capa de datos.** Una de las características fundamentales de Tableau es que admite la arquitectura de datos que usted elija. Tableau no requiere que sus datos se almacenen en un solo sistema, propietario o de otro tipo. La mayoría de las organizaciones tiene un entorno de datos heterogéneo: almacenes de datos conviven con bases de datos, aunque sean locales o se encuentren en la nube. Los cubos y los archivos planos como los de Excel se siguen usando mucho. Tableau puede trabajar con todos ellos de manera simultánea. No es necesario que reúna todos los datos en la memoria, a menos que así lo decida. Si sus plataformas de datos actuales son rápidas y escalables, Tableau le permite beneficiarse de manera directa de su inversión aprovechando la eficacia de la base de datos para responder preguntas. Si este no es el caso, Tableau proporciona opciones simples para actualizar sus datos a fin de que sean rápidos y respondan con nuestro motor de datos en memoria.
- **Conectores de datos.** Tableau incluye más de 40 conectores de datos optimizados para fuentes de datos como Microsoft Excel, SQL Server, Google BigQuery, Amazon Redshift, Oracle, SAP HANA, Salesforce.com, Teradata, Vertica, Cloudera y Hadoop, y se agregan conectores de datos nuevos regularmente. También hay un conector ODBC genérico para cualquier sistema sin conector nativo. Tableau proporciona dos modos de interacción con los datos: conexión en vivo o en memoria. Los usuarios pueden cambiar entre la conexión en vivo y en memoria, según lo deseen.
- **Componentes de Tableau Server.** El trabajo de Tableau Server se realiza mediante los siguientes cuatro procesos de servidor:

 - **Servidor de aplicaciones.** Los procesos del servidor de aplicaciones (wgserver.exe) controlan la exploración de contenido, la administración del servidor

y los permisos para las interfaces web y móvil de Tableau Server. Cuando un usuario abre una vista en un dispositivo cliente, ese usuario inicia una sesión (`workgroup_session_id`) en Tableau Server. El administrador puede configurar fácilmente el tiempo de expiración predeterminado de esta sesión. El usuario puede ejecutar dos o más procesos de servidor de aplicaciones para satisfacer sus necesidades de escalabilidad y disponibilidad. VizQL Server: una vez que el usuario recibe la autenticación del servidor de aplicaciones, puede abrir una vista. El cliente envía una solicitud al proceso VizQL (`vizqlserver.exe`). A continuación, el proceso VizQL envía las consultas directamente a la fuente de datos y devuelve un conjunto de resultados que se expresa en imágenes y se presenta al usuario. En muchos casos, Tableau Server aprovecha las representaciones y el almacenamiento en caché del cliente para reducir la carga del servidor. Además, cada VizQL Server tiene su propia memoria caché que pueden compartir varios usuarios de manera segura. El usuario puede ejecutar dos o más procesos de VizQL Server para satisfacer sus necesidades de escalabilidad y disponibilidad.

- **Data Server.** A diferencia de los enfoques tradicionales de administración de metadatos, el Data Server de Tableau es un componente clave que permite a los administradores de TI habilitar el monitoreo, la administración de metadatos y el control para los equipos de TI, a la vez que se habilitan los análisis de autoservicio para usuarios profesionales. Permite administrar y almacenar fuentes de datos de Tableau de manera centralizada y proporciona a los usuarios finales acceso seguro a datos confiables por medio de implementaciones de análisis de autoservicio. Usted puede administrar centralizadamente metadatos, como conexiones, controladores y filtros de fuentes de datos, para acceder a los datos. Puede asignar permisos específicos a las fuentes de datos de manera que permita al equipo de TI administrar los permisos a las fuentes de datos en función de grupos de identificación específica. En un entorno administrado, los usuarios que conocen mejor sus datos también cuentan con la flexibilidad de definir y publicar definiciones, cálculos y grupos. Estos se pueden compartir para que los usen todos los miembros de la organización o los usuarios de Tableau Desktop para crear y aprovisionar sus propios cálculos, definiciones y grupos. La fuente de datos publicada se puede basar en:
 - Una extracción del motor de datos de Tableau;
 - Una conexión en vivo (los cubos no se admiten como conexiones en vivo).
- **Componente de segundo plano.** El componente de segundo plano actualiza las extracciones programadas, entrega notificaciones y administra otras tareas de segundo plano. El componente de segundo plano está diseñado para consumir todos los recursos de CPU disponibles a fin de concluir la actividad de segundo plano tan pronto como sea posible.

- **Puerta de enlace/equilibrador de carga.** La puerta de enlace dirige solicitudes a otros componentes. Las solicitudes de los clientes, primero, se encuentran con un equilibrador de carga externo, si se configura uno, o la puerta de enlace y, de allí, se los dirige al proceso adecuado. En caso de que no haya un equilibrador de carga externo, si varios procesos se configuran para cualquier componente, la puerta de enlace actúa como un equilibrador de carga y distribuye las solicitudes hacia los procesos. En una configuración de un solo servidor, todos los procesos residen en la puerta de enlace o en el servidor primario. Cuando se trabaja en un entorno distribuido, se designa una máquina física como servidor primario, y las otras se designan como servidores de trabajo y pueden ejecutar cualquier cantidad de procesos adicionales. Tableau Server siempre usa una sola máquina como servidor primario.
- **Clientes: navegadores web y aplicaciones móviles.** Tableau Server proporciona dashboards interactivos a los usuarios mediante HTML5 que no deja rastro en navegadores web o móviles, o de manera nativa mediante una aplicación móvil. No se necesita ActiveX, Java ni Flash para ejecutar informes o visualizaciones. No se requieren complementos ni aplicaciones auxiliares. Tableau Server admite:
 - Navegadores web: Internet Explorer, Firefox, Chrome y Safari.
 - Safari para móviles: las vistas optimizadas para la función táctil se presentan automáticamente en Safari para móviles.
 - Aplicación para iPad: aplicación nativa para iPad que proporciona vistas, contenido, navegación y edición optimizados para la función táctil.
 - Navegador de Android: vistas optimizadas para la función táctil que se ofrecen automáticamente en el navegador de Android.
 - Aplicación para Android: aplicación nativa para Android que proporciona vistas, contenido, navegación y edición optimizados para la función táctil.
- **Clientes: Tableau Desktop.** Tableau Desktop es el entorno de creación rápida de análisis de negocios que se usa para crear y publicar vistas, informes y dashboards en Tableau Server. Mediante Tableau Desktop, un autor de informes puede conectarse a varias fuentes de datos, explorar relaciones, crear dashboards, modificar metadatos y, por último, publicar un libro de trabajo completo o una fuente de datos en Tableau Server. Tableau Desktop también puede abrir cualquier libro de trabajo publicado en Tableau Server o conectarse con cualquier fuente de datos publicada, ya sea que se haya publicado como una extracción o como una conexión en vivo. Tableau Desktop es compatible con escritorios de Windows y de Mac.

Capítulo 3

Marco Metodológico

3.1. Metodología Descendente (TOP-DOWN)

Este enfoque Descendente, se recomienda aplicar cuando se tiene un conocimiento previo de la tecnología y las necesidades de la empresa. Se trata de un método sistémico, que disminuye los problemas de integración pero que resulta ser costoso debido a la gran cantidad de datos que debe manejarse y a la poca flexibilidad en esta metodología.

Lo esencial es ir desde lo más general a lo más específico. Es decir que primero debe formularse un resumen del sistema sin especificar detalles para luego redefinir cada parte del sistema con mayor detalle y así sucesivamente hasta que las especificaciones sean lo más detalladas para validar el modelo. Uno de los elementos que nos ayuda en el diseño son las llamadas “cajas negras”, aunque las mismas no expliquen en detalle los componentes individuales.

Una de las principales ventajas de este enfoque radica en que mediante el mismo almacén de datos resultante se enfoca realmente en las necesidades del cliente y por lo tanto su proceso de apoyo al proceso de toma de decisiones es más preciso.

Por otro lado, este enfoque tiene como inconveniente principal que aumenta la complejidad en la obtención de información para la carga de datos, sobre todo en los casos en los que las fuentes no se encuentran automatizadas.

Este enfoque se basa en la visión de Bill Inmon quien considera que el repositorio de datos debe responder a las necesidades de todos los usuarios en la organización y no solo a un pequeño grupo.

3.1.1. Metodología de Bill Inmon

Esta metodología fue definida por Bill Inmon en el año 1992 en el libro “Building the Data Warehouse” [Inmon, 2002]. En él se proponen los mecanismos necesarios para llevar a cabo la correcta realización de un almacén de datos.

B. Inmon presenta dos caminos a seguir para la construcción de un almacén de datos. Una metodología y un plan de migración. La metodología describe actividades específicas, y el orden en el que deben ejecutarse las actividades, sin embargo las dinámicas

no describen los procesos. Mientras que el plan de migración describe actividades generales dinámicamente que deben ejecutarse a lo largo de todo el desarrollo. Juntos forman una imagen completa de lo que se requiere para construir el almacén de datos. [Inmon, 2002].

Sobre la metodología, argumenta que los ambientes del almacén de datos son dirigidos por los datos, en comparación con los sistemas clásicos los cuales tienen un ciclo de vida de desarrollo dirigido por los requerimientos. Mantiene que los requerimientos son el último elemento a ser considerado en el ciclo de vida del proceso de desarrollo del soporte de decisiones, los mismos son finalmente entendidos luego que el almacén de datos ha sido poblado con datos y los resultados de las consultas han sido analizados por los usuarios [Inmon, 2002].

Uno de los aspectos sobresalientes de la metodología dirigida por los datos es que la misma construye a partir de esfuerzos previos, es decir construye tanto sobre los códigos como de los procesos que han sido desarrollados anteriormente. La única forma en que el desarrollo sobre esfuerzos previos pueda ser completado es mediante el reconocimiento de aspectos comunes. Esto quiere decir que antes que el desarrollador inicie su trabajo, él o ella necesita entender y conocer lo que realmente existe y como se relaciona con el desarrollo del almacén de datos. Esta fase es esencial en la metodología [Inmon, 2002].

B. Inmon fue el primero en definir y defender el concepto de almacén de datos, por lo cual se reconoce como el padre del almacén de datos, ve la necesidad de transferir la información de los diferentes Sistemas Transaccionales (OLTP) de las organizaciones a un lugar centralizado donde los datos puedan ser utilizados para el análisis que sería el Corporate Information Factory (CIF). Insiste además en que ha de tener las siguientes características [Kimball, 2010]:

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

La información ha de estar a los máximos niveles de detalle. Los DWH departamentales o bodega de datos son tratados como subconjuntos de este DWH corporativo, que son construidos para cubrir las necesidades individuales de análisis de cada departamento, y siempre a partir de este DWH Central donde se pueden construir las estructuras Operational Data Stores (ODS) o similares.

Los datos son extraídos de los sistemas operacionales por los procesos ETL y cargados en las áreas de stage, donde son validados y consolidados en el DWH corporativo, donde además existen los llamados metadatos que documentan de una forma clara y precisa el contenido del DWH. Una vez realizado este proceso, los procesos de refresco de los bodega de datos departamentales obtienen la información de la organización, y con las consiguientes transformaciones, organizan los datos en las estructuras particulares requeridas por cada uno de ellos, refrescando su contenido.



Figura 3.1: Enfoque Inmon (Fuente [Kimball, 2010])

La metodología propuesta por B. Inmon explica los resultados que deben obtenerse y el orden en el que deben estar ejecutados los pasos. La manera en la que los resultados sean logrados es un elemento que se deja enteramente a juicio del desarrollador. La metodología se dividió en tres partes componentes, incluye actividades las actividades clásicas necesarias para que sea una metodología dirigida por los datos [Inmon, 2002].

A continuación se definirán de manera general los elementos de cada una de las partes, y componentes de la metodología [Inmon, 2002].

3.1.2. Parte 1: Desarrollo de Sistemas Operacionales

- **Actividades iniciales del proyecto.** Se basa en la recolección de cada uno de los requerimientos a través de entrevistas, recopilación de datos.
- **Uso de código y datos existentes.** Como la premisa indica, en este elemento debemos considerar todos los elementos reutilizables posibles.
- **Determinación de tamaño y fases.** En este punto se divide el proyecto en fases lo más pequeñas y manejables posibles.

- **Formalización de los requerimientos.** Asegurar que los requerimientos sean completos, organizados, legibles, comprensibles y a un nivel de detalle que permita ser efectivos.

Sobre el modelado de datos o modelo dimensional

- **Diagrama Entidad Relación.** A partir de los requerimientos ya definidos se definen los elementos constitutivos del sistema y la relación de cardinalidad entre ellos.
- **Conjunto de Elemento de Datos (CED).** Los CED contiene los atributos de los datos, agrupamiento de los atributos, claves, tipos de datos, conectores y agrupamiento secundario de los datos. Solo los datos primitivos se manejan aquí.
- **Análisis de Desempeño.** Este elemento permite pulir los elementos de ingreso y actualización de data haciendo el proceso más eficiente.
- **Diseño físico de la base de datos.** Se obtienen las tablas y bases de datos diseñadas físicamente, luego de transformar todas las consideraciones lógicas de diseño de datos, desempeño, actualización, ingreso, disponibilidad, etc.

Sobre el Especificaciones de Proceso

- **Descomposición Funcional.** Es la descripción de todas las actividades a ser realizadas durante el desarrollo desde el nivel alto hasta un nivel bajo.
- **Nivel de Contexto 0.** Corresponde a D1, Diagrama Entidad-Relación, en el modelado de datos.
- **Nivel de Contexto 1-n.** Los niveles restantes de la descomposición funcional describen más detalladamente las actividades que ocurren, de manera ordenada, organizada, completa y en concordancia con el flujo de actividades.
- **Diagramas de Flujo de Datos (DFD).** Existe un DFD para cada nivel de contexto n, indica la entrada de un proceso, la salida del proceso, el almacenamiento de datos necesario para establecer el proceso y una breve descripción del proceso.
- **Especificaciones Algorítmicas.** Los procesos de cada DFD se dividen en especificaciones algorítmicas detalladas, tomando en cuenta los aspectos de desempeño que deben ser resueltos en el diseño de los programas.
- **Pseudocódigo.** Los algoritmos y especificaciones de programas se refinan en pseudocódigo, el cual debe incluir completitud, orden de ejecución, todos los casos requeridos. Corresponde a D4 en modelado de datos.

- **Codificación.** Construcción del código fuente. La traducción completa y eficiente de pseudocódigo en código, incluyendo la documentación en código.
- **Caminata.** Explicación verbal del código a los colegas, para encontrar y corregir la mayor cantidad de errores posibles antes de las pruebas.
- **Compilación.** El código fuente es compilado y se corrigen todos los errores encontrados.
- **Pruebas unitarias.** pruebas del código a diferentes niveles.
- **Implementación.** esta etapa incluye actividades como: descarga inicial de datos, conversión de datos, escritura de la documentación y establecimiento de procesos de respaldo.

3.1.3. Parte 2: Desarrollo del almacén de datos

Este es el componente de la metodología que se ocupa del desarrollo de sistemas y procesamiento de soporte a la toma de decisiones.

- **Análisis del Modelo de Datos.** Confirmación que el modelo de datos de la organización es sólido y que contiene la identificación de los temas de mayor interés, cada tema tiene separada su propia definición de datos: subtipos de datos, atributos, relaciones, identificación de claves, entre otros.
- **Análisis Breadbox.** Permite la determinación del tamaño –estimación bruta- del entorno de los sistemas de soporte de toma de decisiones. Simplemente proyecta, en términos crudos, qué cantidad de datos mantendrá el almacén de datos.
- **Valoración Técnica.** Contiene definiciones técnicas que tienen la habilidad de manejar grandes cantidades de datos, permitir que los datos sean ingresados flexiblemente, organizar los datos de acuerdo al modelo de datos y tanto recibir como enviar datos a una amplia variedad de tecnologías.
- **Preparación del entorno técnico.** Instalación, ubicación y desarrollo de los componentes técnicos que recibirán los datos.
- **Análisis de los temas del almacén de datos.** Determinación del tema que será el primero en implementarse.
- **Diseño del almacén de datos.** Este subelemento toma en cuenta las siguientes características; acomodación de los niveles de granularidad, orientación de los datos a los principales temas de organización, la ausencia de datos que no apoya al sistema de soporte de las decisiones, desnormalización donde sea aplicable y finalmente adaptar los datos del entorno operacional al entorno analítico.

- **Análisis de los sistemas fuentes.** Identificación del sistema de registro, es decir el mapeo de los datos del ambiente operacional al ambiente analítico.
- **Especificaciones.** Aquí verifica qué datos operacionales deben ser obtenidos y cómo guardarlos.
- **Programación.** Programas de transformación que permitan la extracción, integración y ubicación en perspectiva de tiempo de los datos.
- **Población.** Ejecución de los programas desarrollados en las etapas anteriores. Se deben resolver los aspectos de frecuencia de población, reglas de purga, administración de múltiples niveles de granularidad y refrescamiento. Con este paso final se obtiene un almacén de datos poblado y funcional, accesible y comprensible que sirve las necesidades de las comunidades de los sistemas de soporte a la toma de decisiones.

3.1.4. Parte 3: Uso del Almacén de Datos

Este último componente describe el uso del almacén de datos para propósitos de análisis.

Repetición del desarrollo estándar. Para la obtención de reportes estándares, el procesamiento analítico repetitivo debe seguir el procesamiento normal descrito en la “Parte 1”, exceptuando el modelado de datos, porque la fuente de datos es el mismo almacén de **datos**.

Una vez comentados los elementos constitutivos de cada una de las partes de la metodología vamos a abordar el segundo elemento de la visión de Inmon, el *Migration Path* [Inmon, 2002]

3.1.5. Plan de Migración (Migration Path)

El punto de arranque del plan de migración es el modelo de datos o modelo dimensional. El modelo suele representar las necesidades de la organización, no necesariamente lo que realmente tiene.

El modelo dimensional debe representar lo siguiente [Inmon, 2002]:

- Atributos
- Las claves
- Los grupos repetitivos de atributos y claves
- Conectores entre las áreas de los temas principales
- Relaciones de subtipos

Una vez que se ha definido el modelo de datos es necesario definir el sistema de registros, el cual se define en términos de los sistemas existentes que la organización ya tiene.

La determinación de la mejor fuente de datos se basa en los siguientes criterios [Inmon, 2002]:

- Datos más precisos
- Datos más cercanos a la fuente de entrada
- Datos más cercanos a la estructura del modelado de datos
- Datos más complejos

El siguiente paso consiste en diseñar el almacén de datos. Si el modelo de datos está elaborado correctamente solo se deben cambiar algunos aspectos para hacerlo más fiel a un diseño de almacén de datos.

Luego del diseño de datos, el siguiente paso es diseñar y construir las interfaces entre el sistema de registro y el almacén de datos, las cuales poblarán el almacén de datos en periodos de tiempo regulares. Entre las actividades que realizan las interfaces tenemos [Inmon, 2002]:

- Integración de los datos
- Alteración de las bases de tiempo de los datos
- Condensación de los datos

La etapa final de este plan de migración es iniciar la población del almacén de datos con el primer tema de interés. Es importante que sólo una parte del almacén de datos sea poblada, de esta manera cuando surjan los cambios, a partir de los procesos iterativos de análisis y reconocimiento de la data, los mismos sean fáciles de manejar e implementar. Una vez que el usuario final tiene acercamiento a los datos y proporciona retroalimentación al arquitecto de datos, entonces puede ser sano poblar cantidades mayores de datos.

De esta manera podemos ver la relación que establece Inmon entre los dos componentes de su visión. El componente *Migration Path* es quien dictamina los pasos a ejecutar para el desarrollo del almacén de datos, mientras que el componente de metodología describe cuáles son los pasos que debe ejecutarse dentro de cada paso del *Migration Path*.

Finalmente, la naturaleza global del enfoque de Inmon hace que su visión resulte más atractiva a la hora de desarrollar un almacén de datos a gran escala, pero para desarrollos de almacén de datos más pequeños, en los que el factor tiempo también es importante, esta metodología no suele resultar tan atractiva.

3.2. Metodología Ascendente (BOTTOM-UP)

En este enfoque el objetivo, es entregar valor de negocio mediante la implementación del modelo dimensional lo más rápido posible. Los datos se modela en un diseño de esquema en estrella para optimizar la facilidad de uso y rendimiento de las consultas.

Se van construyendo bodega de datos relativamente independientes para ir midiendo las ventajas a medida que se va avanzando. A medida que los bodega de datos resultan exitosos se comienzan a enlazar con otros bodega de datos previamente realizados y así sucesivamente hasta que se tiene la completitud del sistema. Las estrategias en la que funcionan el enfoque Ascendente se basan en el conocimiento de todas las variables que traten del sistema

El almacén de datos es un conglomerado de todos los bodega de datos dentro de una empresa, siendo una copia de los datos transaccionales estructurados de una forma especial para el análisis, de acuerdo al Modelo Dimensional (no normalizado), que incluye, como ya vimos, las dimensiones de análisis y sus atributos, su organización jerárquica, así como los diferentes hechos de negocio que se quieren analizar. Por un lado tenemos tablas para las representar las dimensiones y por otro lado tablas para los hechos (facts tables). Los diferentes bodega de datos están conectados entre si por la llamada bus structure, que contiene los elementos anteriormente citados a través de las dimensiones conformadas (que permiten que los usuarios puedan realizar queries conjuntos sobre los diferentes bodega de datos, pues este bus contiene los elementos en común que los comunican). Una dimensión conformada puede ser, por ejemplo, la dimensión cliente, que incluye todos los atributos o elementos de análisis referentes a los clientes y que puede ser compartida por diferentes bodega de datos (ventas, pedidos, gestión de cobros, etc)[Kimball, 2010].

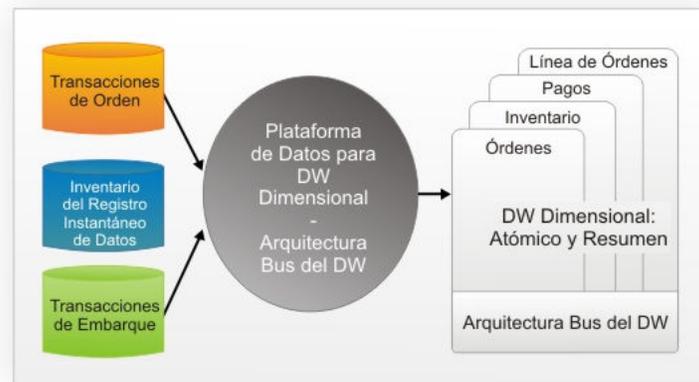


Figura 3.2: Enfoque R. Kimball (Fuentes [Kimball, 2010])

Pues al final el almacén de datos Corporativo no es mas que la unión de los diferentes bodega de datos, que están estructurados de una forma común a través de la bus

structure. Esta característica le hace más flexible y sencillo de implementar, pues podemos construir un Bodega de Dato como primer elemento del sistema de análisis, y luego ir añadiendo otros que comparten las dimensiones ya definidas o incluyen otras nuevas. En este sistema, los procesos ETL extraen la información de los sistemas operacionales y los procesan igualmente en el área stage, realizando posteriormente el llenado de cada uno de los Bodega de Dato de una forma individual, aunque siempre respetando la estandarización de las dimensiones (dimensiones conformadas) [Kimball, 2010].

Uno de los beneficios de este enfoque, es que se centra en la creación de estructuras de datos flexibles, fáciles de usar, utilizando modelos de esquema dimensional, estrella. También ofrece un valor rápidamente por que no ofrece una infraestructura pesada en la delantera. Permite automatizar los procesos de carga de data para así simplificar procesos de mantenimiento y administración de la información.

Un problema con un enfoque, es que se requiere que las organizaciones para hacer cumplir el uso de dimensiones y hechos estándar para garantizar la integración y entregar una única versión de la verdad. Cuando los bodega de datos se ordenan lógicamente dentro de una sola base de datos física, esta integración se realiza fácilmente. Pero en una organización distribuida, descentralizada, puede ser demasiado para pedir a los departamentos y unidades de negocio a que se adhieran y reutilizar las referencias y las reglas para el cálculo de los hechos. No puede haber una tendencia a que las organizaciones crear mercados de datos "independiente" o no integrados.

Este enfoque se adapta a la visión de R. Kimball quien considera que el factor tiempo a la hora de explotar las bondades de un almacén de datos es un elemento determinante a la hora de decidirse por esta tecnología. Este enfoque parte de los requerimientos de negocio, mientras que el enfoque Descendente deja la validación de requerimientos para el final del proceso.

3.2.1. Metodología de Ralph Kimball

La implementación de una solución de BI requiere seguir una metodología con un conjunto de actividades relacionadas entre sí, las cuales han de tener un punto específico de inicio y de fin.

Se tiene que tomar en cuenta en todo momento que la construcción de una solución de BI debe enfocarse directamente en las necesidades de la organización y que los datos presentados al usuario final deben ser consistentes.

Esta metodología se basa en la interpretación de la frase "Ciclo de Vida", que suele usarse en el argot computacional para referirse a todos los pasos del proceso de desarrollo de software pero adaptados a la visión de desarrollo de almacén de datos, en ella se describe paso a paso cómo diseñar, desarrollar y desplegar almacén de datos y bodega de datos. El libro también aborda el proceso de construcción de inicio a fin de forma iterativa y también se muestra la utilización de las técnicas del modelado dimensional. En el mismo, R. Kimball se muestra mucho más técnico que los autores anteriores logrando descripciones más detalladas.



Figura 3.3: Ciclo de Vida de la Metodología de Ralph Kimball (Fuente [Kimball, 2002])

El diagrama no refleja una línea temporal absoluta. Mientras que las cajas son igualmente amplia, hay una gran diferencia en el tiempo y esfuerzo que se requiere para cada actividad principal. Sin embargo a partir de la figura: 3.3 anterior se puede entonces dividir la metodología en [Kimball, 1998]:

3.2.2. Planificación del Proyecto

En su metodología R. Kimball plantea que los pasos iniciales para el desarrollo de un almacén de datos son la planeación del proyecto y la obtención de requerimientos. Cada elemento cumple funciones determinadas [Kimball, 2002].

- **Planeación y Gestión del Proyecto.** Este es el primer paso que debe llevarse a cabo para poder iniciar el desarrollo del almacén de datos. En este paso se establece la situación actual de la empresa y se elabora un plan para el proyecto, definiendo el alcance del mismo.
- **Obtención de Requerimientos.** Este paso se refiere a la implementación de los mecanismos necesarios para poder obtener datos de la información que la empresa desea que se manejen en el almacén de datos. Dichos datos son los que le darán sentido al almacén de datos. Los mecanismos a emplear para obtener dicha información van desde entrevistas hasta sesiones con un facilitador.

3.2.3. Definición de los Requerimientos del Negocio

Consiste en plasmar y expresar los requerimientos de los usuarios basándose en las necesidades de la organización. Este levantamiento de información se puede llevar a cabo realizando reuniones, encuestas, entrevistas, o cualquier otro método que permita expresar de una forma clara y no ambigua los requerimientos del cliente.

Los requerimientos obtenidos han de documentarse, dado a que no sólo servirá para certificar que se contempló todo lo que el cliente necesita, sino que también servirá para el grupo de desarrollo conozca las funcionalidades que ha de tener el sistema, lo que se

necesita para cumplir con los requerimientos especificados, los usuarios involucrados en el proyecto, el léxico que se emplea, entre otras cosas.

Como podemos observar 3.3, después de definir los requerimientos, se realiza tres actividades de forma concurrente, las cuales se enfocan en la tecnología, los datos y las aplicaciones analíticas.

3.2.4. Diseño de la Arquitectura

El diseño técnico sirve como un marco organizativo que soporta la integración de los elementos tecnológicos necesarios para el desarrollo de la solución de inteligencia de negocio.

Este diseño permite identificar los componentes más importantes y minimizar imprevistos al momento de desarrollar e implementar la solución especificada, por medio del análisis de las problemáticas que podrían afectar el proceso de desarrollo. Adicionalmente la realización del diseño permite coordinar la ejecución de diversos procesos en paralelo, acelerando de esta forma el desarrollo mediante la utilización de componentes modulares.

Con los requerimientos documentados se procede a realizar una estructura básica que soporte las necesidades de la implementación de la arquitectura para así llevar a cabo el desarrollo de la solución de inteligencia de negocio.

Una vez especificada la estructura elemental para la solución se procede a realizar el diseño específico de cada uno de los subsistemas que se encuentran implicados con nuestra solución, es decir, se definen los requisitos por cada uno de los componentes con el mayor nivel de detalle posible. También se deben de considerar los requerimientos de seguridad, de la infraestructura física y de las necesidades con respecto a la configuración.

Además de considerar los requerimientos funcionales de la solución, se necesita tomar en cuenta aspectos como la capacidad que debe tener el almacén de datos, la escalabilidad, la flexibilidad y el desempeño de la solución.

Es importante establecer las fases del plan de implementación, darle prioridad a los aspectos técnicos de la arquitectura debido a que usualmente no se pueden implementar todos los elementos deseados desde el inicio del proyecto, por esta razón, se deben definir cuáles son los aspectos necesarios para la implementación del mismo y cuáles pueden obtenerse luego para aportar un valor agregado a la solución.

Para concluir, es necesario documentar el diseño técnico de la arquitectura incluyendo las fases del plan de implementación. Este documento debe incluir información pertinente para que los profesionales calificados procedan con la construcción de la solución.

En la siguiente tabla se resume el modelo de arquitectura que propone R. Kimball. La misma recibe el nombre de marco de trabajo de la arquitectura como vemos en la 3.1 donde las columnas muestran las principales área de la arquitectura: datos, técnica e infraestructura y las filas representan los niveles de detalle en orden creciente: requerimientos del negocio, modelo de arquitectura, modelo detallado, implementación.

Nivel de Detalle	Datos (qué)	Técnica (cómo)		Infraestructura (dónde)
		Back room	From room	
Requerimientos de Negocio y Auditoría	Datos que dirigen el negocio	Cómo se tratan los datos	Formas de análisis de los datos	Hardware y Software necesario
Modelos de arquitectura y documentos	El modelo dimensional	Capacidades especificadas para tratar los datos	Servicios para que los usuarios utilicen la información	Fuente-Destino de datos. Capacidad de almacenamiento
Modelos detallados y especificaciones	El modelo físico y lógico	Que estándares productos proveen dichas capacidades	Especificaciones de reportes. Usuarios posibles. Frecuencia	Interacción con las capacidades y utilidades del sistemas
Implementación	Creación de la Base de Datos, índices, etc.	Programar, automatizar, documentar	Construir el ambiente de reporte y análisis	Instalar y probar los componentes nuevos de la infraestructura

Cuadro 3.1: Marco de trabajo de la Arquitectura

- **Arquitectura de Datos**

Este nivel abarca tanto el diseño físico y lógico de los modelos de datos. Todo lo relacionado a dicho diseño ya fue explicado en puntos anteriores.

- **Arquitectura Técnica**

El área de la arquitectura técnica cubre los procesos y herramientas que se aplican a los datos. Esta área está dividida a su vez en dos elementos con sus requerimientos y mecanismos propios [Kimball, 1998]:

- **Back Room:** responsable de la obtención y preparación de los datos. “Es el cuarto de máquinas del almacén de datos” . La función principal de este elemento es resolver los problemas de migración y transformación de datos desde las fuentes de datos hasta el entorno almacén de datos.

En la arquitectura del back room, se debe analizar los aspectos relativos a: los sitios donde se almacenan los datos (sistemas fuente), los servicios que proporcionará (extracción, transformación, carga) y la administración de los activos del back room (respaldo y recuperación, entre otros).

- **Front Room:** responsable de entregar los datos a los usuarios determinados. La función principal de este elemento es de servir de intermediario entre los usuarios y la información, de manera tal de poder ofrecerle a dichos usuarios lo que necesiten pero sin que los mismos vean las complejidades latentes en dichos procesos.

En la arquitectura del front room se deben analizar los aspectos relativos a: los sitios donde se almacenan los datos (herramientas de acceso a los almacenes, entre otros) y los servicios que proporcionarán para el acceso de datos (navegación, acceso y seguridad, administración de consultas, entre otros).

- **Arquitectura de Infraestructura y Metadatos**

Este apartado se refiere a las plataformas sobre las cuales se va a cimentar el

almacén de datos. La infraestructura incluye el hardware, la red y elementos de bajo nivel que no suelen ser competencia de las otras áreas. Los elementos a tomar en consideración para esta área son:

- **Los requerimientos del negocio:** los cuales son los que determinan la dirección a tomar por parte del almacén de datos.
- **Los factores de la infraestructura del back room:** tamaño de datos, volatilidad de la data, hardware, sistema operativo, entre otros.
- **Los factores de infraestructura del front room:** servidores de aplicación, entre otros.
- **Factores de redes y conectividad:** ancho de banda, acceso remoto, conectividad a la base de datos, entre otros.

3.2.5. Selección de productos e instalación

El producto a seleccionar debe acoplarse al marco definido en el diseño de la arquitectura, adaptándose a los procesos de la organización.

Para seleccionar el producto más apropiado para la solución se debe desarrollar una matriz de evaluación, la cual debe contener los diversos criterios que se deseen analizar. Mientras más específicos y detallados los criterios mejor se adapta a los requerimientos necesarios para el desarrollo de la solución. Además de la matriz, es necesario conocer la mayor cantidad de productos posibles que se encuentren en el mercado, para así poder tener más fundamentos al momento de seleccionar un producto en específico.

Después de realizar dicha evaluación se selecciona el producto que (según los análisis realizados) se adapte mejor para el cumplimiento de los requerimientos funcionales, así como también, de las especificaciones del diseño técnico de la arquitectura.

3.2.6. Diseño de Datos o modelado dimensional

El modelado dimensional, según su creador R. Kimball, es el diseño tanto físico como lógico que se va a encargar de transformar las fuentes de datos, previamente definidas en el paso anterior, en estructuras aptas para el almacén de datos. Cada modelo dimensional está compuesto de una tabla que tiene una clave compuesta llamada tabla de hechos y un conjunto de tablas más pequeñas llamadas dimensiones. Cada tabla dimensión tiene una clave primaria simple, que a su vez forma parte de la clave compuesta de la tabla de hechos. Esta descripción se refiere al esquema estrella explicado en el Capítulo 2.

Los pasos necesarios para convertir un Diagrama Entidad-Relación (ERD) a un conjunto de diagramas que se basan en el modelado dimensional son [Kimball, 2002]:

- Separar el Diagrama Entidad-Relación en procesos finitos y discretos e ir los modelando por separado.

- Determinar cuáles serán las tablas de hechos mediante la selección de las relaciones muchos a muchos del modelo que no tengan claves primarias simples.
- Determinar cuáles serán las tablas de hechos mediante la selección de las relaciones muchos a muchos del modelo que no tengan claves primarias simples.
- Crear las tablas dimensionales desnormalizando todas las tablas y añadiéndoles claves primarias simples.
- Definir también las tablas dimensionales conformadas que son aquellas que sirven de dimensión a más de una tablas de hechos.

Se fundamenta en las siguientes ventajas para proponer el modelado dimensional [Kimball, 2002]:

- Soporta cambios inesperados en el comportamiento del usuario.
- Es capaz de crecer y extenderse para aceptar nuevos elementos de datos y de diseño.

Otros elementos que se toman en cuenta dentro del aspecto del modelado dimensional son [Kimball, 2002]:

- **Arquitectura de Bus del almacén de datos**

La arquitectura del bus de almacén de datos se basa en la idea que un almacén de datos está compuesto por muchos bodega de datos. Cada Bodega de Dato a su vez debe estar representado dentro de un modelo dimensional y compuesto por tablas de hechos y tablas dimensionales [Kimball, 2002].

Por lo tanto el autor plantea que es necesario al principio del proyecto definir una arquitectura estricta y finita de datos sobre la cual se desarrollará el almacén de datos. Posteriormente, y basándose en la arquitectura previamente definida, deben irse construyendo los bodega de datos. R. Kimball explica que “cada Bodega de Dato para que sea práctico debe ser basado en los datos más granulares (atómicos) que sea posible coleccionar y almacenar” [Kimball, 2002].

La correcta adherencia entre un Bodega de Dato y otro se consigue a través de las tablas dimensionales conformadas, ya que a través de ellas se logra [Kimball, 2002]:

- Una tabla dimensional solo puede ser usada contra múltiples tablas de hechos en el mismo espacio de base de datos.
- Las interfaces de usuario y los datos que contienen son consistentes en cualquier instante en que se utilice la dimensión.
- Existe una interpretación consistente de los atributos en cada Bodega de Dato.
- A partir de lo anteriormente expuesto podemos entender que las dimensiones conformadas son las que hacen de bus en el almacén de datos. Por lo tanto, al definir una interfase estándar de bus (mediante dimensiones conformadas) un nuevo Bodega de Dato puede añadirse al almacén de datos y coexistir con los otros sin ningún inconveniente.

A partir de lo anteriormente expuesto podemos entender que las dimensiones conformadas son las que hacen de bus en el almacén de datos. Por lo tanto, al definir una interfase estándar de bus (mediante dimensiones conformadas) un nuevo Bodega de Dato puede añadirse al almacén de datos y coexistir con los otros sin ningún inconveniente.

■ **Elementos del Modelado Dimensional**

Los siguientes son elementos esenciales para el Modelado Dimensional:

- **Hechos:** un hecho es una observación del mercado de trabajo.
- **Atributos:** usualmente son campos de texto que describen las características de algo tangible, como las descripciones del producto.
- **Dimensiones:** se refiere a los atributos que describen al objeto y que a su vez se correlacionan entre ellos.

■ **Método de Diseño de cuatro pasos para diseñar una tabla de hechos**

Los siguientes son pasos necesarios para un diseño lógico de un esquema dimensional:

- **Escoger el bodega de datos.** Como se desea en primera instancia mantener un nivel de simplicidad adecuado entonces es recomendable en lo posible escoger una única fuente de datos para el bodega de datos.
- **Declarar la granularidad de la tabla de Hechos.** Consiste en definir qué se considera una tabla de hechos para el diseño dimensional que se propone.
- **Escoger las dimensiones.** escoger, en lo posible, aquellas dimensiones que sean capaces de tomar un valor único en el contexto de un conjunto dado de medidas del negocio.
- **Escoger los hechos.** Los hechos deben ser siempre específicos a la granularidad de la tabla de hechos. No se deben hacer manipulaciones convenientes a otros cálculos porque eso podría afectar la naturaleza del hecho.

■ **Construcción de Modelos Dimensionales**

R. Kimball concibe cada bodega de datos como un modelo dimensional separado y el conjunto de modelos dimensionales constituye el diseño lógico del almacén de datos. Para la construcción de los bodega de datos R. Kimball propone emplear un enfoque Ascendente al que llama Matriz de Arquitectura del almacén de datos, en el cual se identifican todos los bodega de datos que se puedan construir y a su vez todas las dimensiones que implican a cada bodega de datos, posteriormente se deben relacionar ambos elementos para así definir cuáles dimensiones podrán ser utilizadas por más de un bodega de datos. Dichas dimensiones pasarán a formar el denominado Bus de almacén de datos.

En la siguiente etapa, una vez que se han identificado los bodega de datos a construir junto con sus respectivas dimensiones, se procede a realizar los diseños físicos

y lógicos de las tablas individuales usando el método de cuatro pasos explicado anteriormente.

Con todos estos elementos se podrá elaborar entonces un borrador inicial del sistema el cual permitirá identificar los hechos base y los hechos derivados que deberán ser incluidos y así pulir el modelo para adaptarlo mejor a las características y exigencias del negocio. En este proceso se podrán realizar todos los cambios necesarios, los cuales deben estar debidamente documentados, junto a todas las decisiones que se realicen, desde elegir las fuentes de datos hasta determinar las fórmulas de cálculo de los hechos derivados y demás.

3.2.7. Especificación y desarrollo de aplicaciones analíticas

Como se puede apreciar en la 3.3, la especificación y el desarrollo de las aplicaciones analíticas se realiza después de obtener los requerimientos del negocio.

Para realizar las especificaciones de las aplicaciones analíticas debemos llevar a cabo un análisis de los diversos productos existentes en el mercado, buscando siempre aquel que se adapte mejor a lo que se requiere para satisfacer las necesidades de la organización.

En la fase de desarrollo de las aplicaciones analíticas es necesario establecer convenciones con respecto a los nombres, cálculos, librerías y codificaciones para evitar rehacer trabajo a futuro. Esta actividad puede iniciarse una vez que el diseño de la base de datos esté completo, las herramientas de acceso a los datos y metadatos estén instaladas y los datos históricos hayan sido cargados al almacén.

3.2.8. Implementación

En esta fase de la metodología es necesario contemplar todos aquellos factores que durante fases iniciales y de diseño no fueron contempladas. A nivel del diseño lógico se debe realizar una última revisión sobre el diseño lógico para determinar cuáles hechos deben ser agregados con respecto a qué dimensiones para mejorar así el rendimiento del almacén de datos.

A nivel del diseño físico el mismo debe culminarse. La secuencia básica es iniciar con elementos de planeación, que se refieren a establecer estándares de nomenclatura, de bases de datos, estrategias de seguridad, entre otros. Luego, a partir del estado del diseño físico, realizar un corte preliminar en el tamaño de la base de datos y su respectiva tasa de crecimiento. Al mismo tiempo, cuando se tienen ya definidas todas las tablas, se debe hacer un corte para la estrategia de establecimiento de índices. A partir de este momento se puede dar por concluido el diseño físico y se puede iniciar realmente el proceso de implementación pues se cuenta con suficiente información acerca de lo que ocurre en la base de datos en sus mínimos detalles.

Posteriormente quedarían pendientes dos elementos dentro de esta fase de la metodología: la consolidación de datos y construir aplicaciones al usuario final.

Sobre la consolidación de datos R. Kimball propone dividir este proceso en un plan de 10 pasos para conseguir con éxito el proceso de consolidación de los Datos en el almacén de datos [Kimball, 1998]:

- **Plan**

- Crear un esquema de alto nivel, muy consolidado, del flujo de fuente a destino.
- Probar, escoger e implementar la herramienta de consolidación de datos.
- Bosquejar gráficamente cualquier reestructuración o transformación de datos compleja.

- **Carga de Dimensiones**

- Construir y probar la carga de una tabla de dimensión estática.
- Construir y probar el proceso de cambio lento para una dimensión.
- Construir y probar la carga de las dimensiones faltantes.

- **Tablas de hechos y automatización**

- Construir y probar la carga de las tablas de hechos históricas.
- Construir y probar el proceso de carga incremental.
- Construir y probar la carga de las tablas de agregaciones.
- Diseñar, construir y probar las aplicaciones de automatización del proceso de conciliación.

Debido a que en esta etapa deben cuidarse todos los detalles de data y los inherentes a la implementación entonces suele considerarse que la etapa de conciliación de datos es una de las más complicadas en el proceso de desarrollo de un almacén de datos.

Sobre la construcción de aplicaciones de usuario final, este proceso está relacionado a la herramienta que se vaya a emplear para el acceso a los datos y por lo tanto deberá brindar soporte a los usuarios con requerimientos de análisis tanto definidos como improvisados.

3.2.9. Despliegue y Crecimiento

Un despliegue exitoso de un almacén de datos requiere:

- **Planeación del despliegue**

“Despliegue es la convergencia de tecnología, datos y aplicaciones en los escritorios de los usuarios de negocios, junto con la necesaria educación y estructura de soporte al usuario” [Kimball, 1998].

La planeación del despliegue se puede dividir en varias partes. En una de ellas se determina la disponibilidad de los escritorios de los usuarios para su instalación, en otra parte se desarrolla una estrategia para educar al usuario final sobre la mejor manera de explotar las bondades del almacén de datos. El desarrollo de un plan para brindar soporte al usuario final sobre el uso de la herramienta es fundamental para cimentar el proceso de aprendizaje y así garantizar el éxito del proyecto.

Un último elemento a ser tomado en consideración es el desarrollo de un marco de trabajo para la emisión del despliegue. Cada emisión estará definida dentro de nuevos requerimientos que vayan añadiéndose al almacén de datos o nuevos grupos de usuarios que vayan a hacer uso del mismo. Es importante que cada emisión de despliegue esté debidamente documentada.

- **Mantenimiento y Crecimiento del almacén de datos**

Este es el último componente dentro de la metodología propuesta por R. Kimball. Este apartado se basa en la constante retroalimentación que deberá existir entre los usuarios finales y el equipo técnico de forma de poder medir y proyectar el éxito del proyecto y también gestionar adecuadamente las operaciones del almacén de datos.

3.3. Metodología a seguir

Desde un estudio de las dos metodologías, se puede establecer una serie de comparaciones, descritas en la tabla 3.2

Metodología Inmon	Metodología Kimball
Los requerimientos no se conocen al principio.	Los requerimientos definen la implementación.
Método descendente.	Método ascendente.
Se define modelo, arquitectura e implementación.	Se define de forma paralela arquitectura, modelo e implementación.
Implementación lenta.	Implementación rápida.
Ideal para soluciones complejas.	Ideal para soluciones escalables.
Parte de que no todos los requerimientos están establecidos.	Parte de que los requerimientos están establecidos.
Adaptada a necesidades universales.	Adaptada a necesidades universales.
No contempla de forma directa la visualización de los indicadores.	No contempla de forma directa la visualización de los indicadores.
Es una metodología para el desarrollo de un almacén de datos.	Es una metodología para el desarrollo de un almacén de datos.

Cuadro 3.2: Comparación de las metodologías

Bill Inmon plantea que el desarrollo de la solución tiene de ser descendente es decir que se debe realizar primero el almacén de datos que abarque toda la estructura de

la organización, pasando por un proceso de integración y purificación de los datos. Posteriormente se generan la bodega de datos que serán subconjuntos del almacén de datos que se crean a partir de agrupaciones lógicas existentes, como por ejemplo agrupar los elementos que tienen que ver con el área de ventas y generar una bodega de datos.

Ralph Kimball plantea que el desarrollo de la solución se tiene que realizar de manera inversa a la ofrecida por Inmon, puesto que se basa en la construcción de las bodegas de datos departamentales, para luego hacer su integración en un almacén de datos que abarque toda la organización, este tipo de diseño se denomina “Ascendnete”. Ralph establece que la complejidad de las organizaciones dificulta la realización de una solución inicial que abarque toda la organización. Por lo que la realización de bodega de datos que apoyen los procesos organizacionales específicos, es un mejor comienzo para el desarrollo de una solución. Eventualmente, a medida que se van construyendo más bodega de datos, estos pueden conectarse por elementos en común (atributos, dimensiones, etc) por lo que se va generando una solución que va abarcando íntegramente la organización.

La metodología Kimball es idónea cuando se está empezando a desarrollar una infraestructura de inteligencia de negocio, porque se puede desarrollar mercado de datos los cuales tienen una complejidad menor e instaurar una solución de negocios completa sobre un área determinada. Esto ayuda a que el proyecto pueda ser puesto en producción de forma rápida que si se compara con Inmon, lo cual ayuda a la organización a recuperar su inversión más rápido.

La metodología a seguir es una adaptación a la metodología de Kimball que se muestra en la figura 3.4 en cuando a una solución de inteligencia de negocio, ya que la metodología busca es el desarrollo de un almacén de datos, cuando se realiza su adaptación es incorporarle un componente de manipulación de la información.

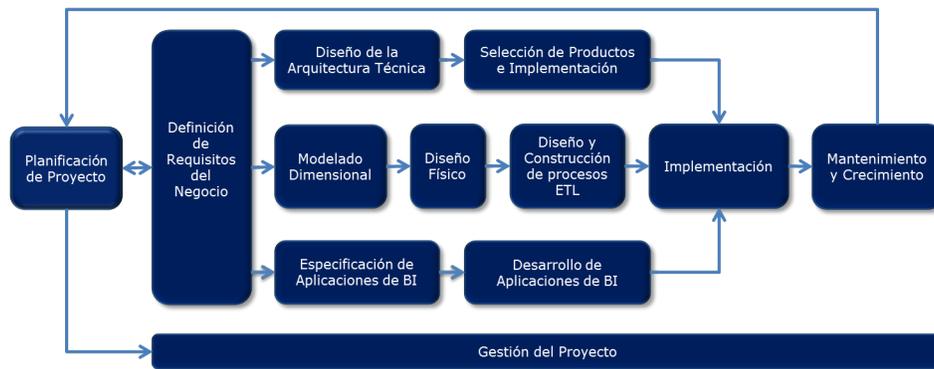


Figura 3.4: Ciclo de vida dimensional adaptado a inteligencia de negocio

Capítulo 4

Marco Aplicativo

Para cumplir con los objetivos planteados, la solución de BI para la *Coordinación de Investigación* se realizó empleando una adaptación a la metodología de Ralph Kimball.

La descripción de las actividades realizadas a lo largo del proceso de desarrollo de la solución de inteligencia de negocio, usando la metodología seleccionada, se exhiben a continuación.

4.1. Planificación del Proyecto

Como en todo proyecto software, es necesario realizar una planificación del mismo para tratar de garantizar que el desarrollo, se realizara según las siguientes actividades previstas:

- Levantamiento de la información.
- Definición de indicadores de gestión.
- Diseño de modelo dimensional.
- Implementación del almacén de datos.
- Implementación de las consultas analíticas y reportes.
- Elaboración de la implementación
- Pruebas a la aplicación.

El mayor tiempo del proyecto se destina al modelo dimensional, fase que agrupa cuatro subtarefas que son: Definir el modelo de negocio, Definir el grano, Elegir las dimensiones, Identificar los hechos.

En la tabla 4.1 se define con detalle cada una de las fases del proyecto que se desarrolla.

Actividades	Duración
Definición de Requisitos del Negocio	2 semanas
Diseño de la Arquitectura Técnica	1 semana
Selección de Productos e Implementación	1 semana
Modelado Dimensional	4 semanas
Diseño Físico	2 semanas
Diseño y Construcción de procesos ETL	2 semanas
Implementación	4 semanas

Cuadro 4.1: Planificación del proyecto

Se realizó un análisis de la base de datos de los proyectos de investigación de la *Facultad de Ciencias* para poder determinar los elementos de los proyectos, productos, eventos, área de conocimiento, la formación de los investigadores, entre otros. Es decir lo que se desea conocer.

Con la investigación realizada se logró establecer un modelo de constelación del proceso estudiado, así como también, se logró definir una serie de reportes y de indicadores, facilitar la toma de decisiones en la *Coordinación de Investigación de la Facultad de Ciencias*.

4.2. Definición de los Requerimientos

Bajo la perspectiva de seguimiento de proyectos de investigación, abstrayéndonos de las variantes que pudieran implementarse en el mismo, para poder generar beneficios a la organización necesitan poder monitorear los proyectos, productos, eventos y el nivel académico en términos generales, entre otros.

Para facilitar a la *Coordinación de Investigación* la obtención de información rápida, precisa y de calidad para saber de ellos. Por esta razón basándonos en lo que se desea saber, se contempló una serie de reportes y indicadores de la coordinación para apoyar en la toma de decisiones. Se muestran a continuación:

1. Conteo de productos de los investigadores, por área de conocimiento, área de aplicación, institución, y la fecha de publicación de dicho producto.
2. Conteo de proyectos de los investigadores, por área de conocimiento, área de aplicación, institución, y la fecha que inicia y finaliza un proyecto.
3. Conteo de eventos de los investigadores, por área de conocimiento, institución, y la fecha en que inicia y finaliza el evento.
4. Saber el Nivel académico de los investigadores, por institución, y fecha de titulación.

Para poder desarrollar los indicadores y reportes referentes a los requerimientos contemplados para la solución de inteligencia de negocio propuesta para un modelo de gestión de los proyectos de la *Coordinación de Investigación*, es necesario especificar cómo se puede obtener lo deseado. Se realiza un análisis de los indicadores, especificando las fórmulas utilizadas para calcularlos.

4.3. Diseño de la Arquitectura Técnica

El diseño de la arquitectura implementado para la solución de inteligencia de negocio se expone en la figura 4.1.



Figura 4.1: Arquitectura de Diseño

La arquitectura implementada está conformada por cuatro (4) componentes y diversos procesos y elementos que permiten llevar a cabo el flujo de trabajo de la misma:

1. Se presentan las fuentes de datos que es la base de datos de la coordinación, que se considera para la solución de inteligencia de negocio. Estos datos pueden provenir de fuentes internas y/o externas, usualmente se emplean aquellos almacenados en los sistemas transaccionales y en archivos utilizados por la organización.
2. Es el proceso de extracción, transformación y carga que se ejecutará periódicamente, se procede a almacenar los datos de las fuentes de datos en el almacén de datos.
3. La estructura del almacén de dato que permite llevar a cabo la obtención de los datos requeridos por el usuario con un alto rendimiento.
4. El ultimo componente se refiere a la capa de presentación, es decir, a las herramientas de acceso a los datos que hacen uso del almacén de datos para permitir a los usuarios visualizar y analizar los distintos indicadores y reportes contemplados para la solución de inteligencia de negocio

4.5. Modelo Dimensional

El diseño del modelo dimensional se realizó basándose en los requerimientos definidos, se definió en un modelo tipo estrella para los proyectos de investigación, éstos fueron analizados para determinar qué elementos podrían considerarse como hechos o dimensiones que pudieran formar parte de la solución de inteligencia de negocio.

Para poder saber que hechos y dimensiones son necesarios para poder realizar la solución planteada, se desglosó uno a uno los requerimientos haciendo hincapié en determinar lo que se quería medir y bajo qué perspectivas se deseaba ver.

Los hechos medibles que se tomaron en cuenta para la solución planteada son los elementos o indicadores que se quieren saber en la *Coordinación de Investigación*, éstos son:

- Conteo de proyectos.
- Conteo de productos.
- Conteo de Eventos.
- El Nivel Académico de los investigadores

Así como se contemplaron los hechos, el análisis indicó las diversas perspectivas con las cuales se desearía o se pudiera agrupar los datos. Indicó como se identificaron las distintas dimensiones que se contemplan en la solución y se mencionan a continuación:

- Tiempo.
- Investigador.
- Área de Conocimiento.
- Área de Aplicación.
- Institución.
- Proyectos.
- Producto.
- Evento.
- Nivel Académico.

Con los elementos expuestos anteriormente se construyó el modelo dimensional tipo estrella como se muestra en la figura 4.3.



Figura 4.3: Modelo Dimensiona tipo estrella

Una vez definidas las perspectivas a utilizar y los hechos a medir en la solución propuesta, se procedió a identificar por cada dimensión los atributos y las jerarquías que se consideraron para poder cumplir con los objetivos planteados.

Una vez establecida las dimensiones, puede establecerse las distintas jerarquías asociadas a ella. Entendiendo que las jerarquías van a permitir navegar sobre los distintos niveles lógicos existentes. En la figura 4.4 se observan las jerarquías establecidas para cada una de las dimensiones.

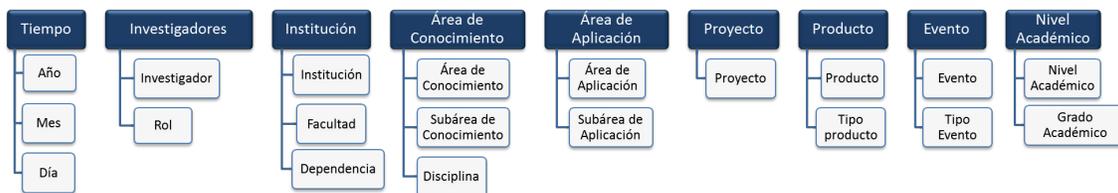


Figura 4.4: Jerarquía de las dimensiones

La jerarquía en cada dimensión se define dependiendo de lógica del negocio. Una vez definidos los niveles jerárquicos para cada dimensión contemplada para la solución de inteligencia de negocio propuesta, se definen los atributos necesarios para cumplir con los requerimientos planteados.

A continuación se describen todas las dimensiones y la tabla de hecho contempladas para la solución de inteligencia de negocio para un modelo tipo estrella de los proyectos de investigación:

- Dimensión Tiempo.** Permite ver los proyectos de investigación en función del momento en que se crea, termina un proyecto. Por esto, dicha dimensión tomara distintos roles como la fecha de inicio y fecha fin en el modelo dimensional. Una característica particular de la dimensión tiempo es que son generadas todas las fechas acorde a un período de tiempo en el proceso de ETL, para luego asociar las fechas específicas a las ocurrencias en las instancias. Esta forma de trabajo permite generar en los cuadros de mandos calendarios porque se poseen todas las fechas, para posteriormente solo obtener los proyectos a partir de un rango de fechas establecido. La tabla 4.2 muestra los atributos de la dimensión tiempo.

Atributo	Tipo de Dato	Descripción	Procedencia
TIEMPO_SK.	SERIAL (clave primaria) .	Identificador univoco de una instancia.	Generado en el proceso ETL.
FECHA	DATE	Una fecha del calendario	Generado en el proceso ETL.
FECHA_NOMBRE	VARCHAR(255)	Fecha del calendario	Generado en el proceso ETL.
ANO	INTEGER	Año de la fecha a partir del 2015.	Generado en el proceso ETL.
ANO_NOMBRE	VARCHAR(2550)	Año de la fecha	Generado en el proceso ETL.
MES	INTEGER	Mese del ano del 1 al 12	Generado en el proceso ETL.
MES_NOMBRE	VARCHAR(255)	Nombre de los meses	Generado en el proceso ETL.
DIA_SEMANA	INTEGER	Día de la semana .	Generado en el proceso ETL.
DIA_SEMANA_NOMBRE	VARCHAR(255)	Nombre de la semana , Domingo hasta Sábado.	Generado en el proceso ETL.

Cuadro 4.2: Dimensión Tiempo

- Dimensión Investigador.** Permite ver el investigador con sus atributos que lo definen como se muestra en la figura 4.5.

Atributo	Tipo de Dato	Descripción	Procedencia
INVESTIGADOR_SK	SERIAL (clave primaria)	Identificador univoco de una instancia.	Generado en el proceso ETL.
INVESTIGADOR_ID	INTEGER	Identificador unívoco utilizado en el ambiente transaccional de solo los Investigadores.	Base de Dato Sigic. usuario.id
INVESTIGADOR_TIPO_DOC_CI	CHAR(1),	Tipo de documento V o E	Base de Dato Sigic. usuario. tipo_documento_identidad.
INVESTIGADOR_CI	INTEGER	Numero de cedula o pasaporte del usuario	Base de Dato Sigic. usuario. documento_identidad.
INVESTIGADOR_NOMBRE	VARCHAR(50)	Primer nombre del usuario	Base de Dato Sigic. usuario. primer_nombre
INVESTIGADOR_APELLIDO	VARCHAR(50)	Primer apellido del usuario	Base de Dato Sigic. usuario. primer_apellido
USUARIO_ROL_ID	INTEGER	Identificador unívoco utilizado en el ambiente transaccional .	Base de Dato Sigic. usuario_rol.id.
INVESTIGADOR_ROL_ID	INTEGER	Identificador unívoco utilizado en el ambiente transaccional .	Base de Dato Sigic. rol.id.
INVESTIGADOR_ROL_NOMBRE	VARCHAR(255)	Nombre de los roles (Administrador, Administrador de dependencia, Investigador)	Base de Dato Sigic. rol.nombre.

Cuadro 4.3: Detalle de la Dimensión Investigador

- **Dimensión Institución.** Permite ver la institución con sus atributos que los definen como se muestra en la figura 4.5.

Atributo	Tipo de Dato	Descripción	Procedencia
INSTITUCION_SK	SERIAL (clave primaria) .	Identificador univoco de una instancia.	Generado en el proceso ETL.
INSTITUCION_SIGLAS	INTEGER	Identificador de la institución	Base de Dato Sigic. institucion.siglas
INSTITUCION_NOMBRE	VARCHAR(255)	Nombre de la institución.	Base de Dato Sigic. Institucion.nombre.
FACULTAD_SIGLAS	VARCHAR(25)	Sigla de la facultad.	Base de Dato Sigic. facultad.codigo
FACULTAD_NOMBRE	VARCHAR(255)	Nombre de la facultad.	Base de Dato Sigic. facultad
DEPENDENCIA_ID	INTEGER	Identificador de la dependencia.	Base de Dato Sigic. Dependencia.id
DEPENDENCIA_SIGLAS	VARCHAR(25)	Siglas de la s dependencia.	Base de Dato Sigic. Dependencia.siglas
DEPENDENCIA_NOMBRE	VARCHAR(255)	Nombre de la s dependencia	Base de Dato Sigic. Dependencia.nombre
DEPARTAMENTO_ID	INTEGER	Identificador de los departamento.	Base de Dato Sigic. departamento_investigacion.id
DEPARTAMENTO_SIGLAS	VARCHAR(25)	Sigla de los departamento.	Base de Dato Sigic. departamento_investigacion.codigo
DEPARTAMENTO_NOMBRE	VARCHAR(205)	Nombre de los departamentos.	Base de Dato Sigic. departamento_investigacion.nombre
UNIDAD_ID	INTEGER	Identificador de la unidad.	Base de Dato Sigic. Unidad.id
UNIDAD_SIGLAS	VARCHAR(25)	Sigla de las unidades	Base de Dato Sigic. Unidad.codigo
UNIDAD_NOMBRE	VARCHAR(255)	Nombre de las unidades	Base de Dato Sigic. Unidad.nombre

Cuadro 4.4: Dimensión de Institución

- **Dimensión de Área de Conocimiento.** Permite ver el área de conocimiento con sus atributos que lo definen como se muestra en la figura 4.5 .

Atributo	Tipo de Dato	Descripción	Procedencia
AREA_CONOCIMIENTO_SK	SERIAL (clave primaria) .	Identificador univoco de una instancia.	Generado en el proceso ETL.
DISCIPLINA_ID	INTEGER	Identificador de la disciplina.	Base de Dato Sigic. Disciplina.id.
AREA_CONOCIMIENTO_NOMBRE	VARCHAR(255)	Nombre del área de conocimiento.	Base de Dato Sigic. area_conocimiento.nombre.
SUBCONOCIMIENTO_NOMBRE	VARCHAR(255)	Nombre de la subarea de conocimiento.	Base de Dato Sigic. subarea_conocimiento.nombre.
DISCIPLINA_NOMBRE	VARCHAR(255)	Nombre de la disciplina.	Base de Dato Sigic. Disciplina.nombre.

Cuadro 4.5: Dimensión de Área de Conocimiento

- **Dimensión de Área de aplicación.** Permite ver el área de aplicación con sus atributos que lo definen como se muestra en la figura 4.5 .

Atributo	Tipo de Dato	Descripción	Procedencia
AREA_APLICACION_SK	SERIAL (clave primaria) .	Identificador univoco de una instancia.	Generado en el proceso ETL.
SUBAPLICACION_ID	INTEGER	Identificador de la subarea de aplicación.	Base de Dato Sigic. subarea_aplicacion.id.
AREA_APLICACION_NOMBRE	VARCHAR(255)	Nombre del área de aplicación.	Base de Dato Sigic. area_aplicacion.nombre.
SUBAPLICACION_NOMBRE	VARCHAR(255)	Nombre de la subarea de aplicación	Base de Dato Sigic. subarea_aplicacion.nombre.

Figura 4.5: Dimensión de Área de aplicación

- **Dimensión Proyecto.** Permite ver el proyecto con sus atributos que lo definen como se muestra en la figura 4.6 .

Atributo	Tipo de Dato	Descripción	Procedencia
PROYECTO_SK	SERIAL (clave primaria) .	Identificador univoco de una instancia.	Generado en el proceso ETL.
PROYECTO_ID	INTEGER	Nombre del área de aplicación.	Base de Dato Sigic. proyecto.nombre.
PROYECTO_TITULO	TEXT	Título del proyecto	Base de Dato Sigic. Proyecto.titulo
PROYECTO_OBJETIVO	TEXT	Objetivo del proyecto	Base de Dato Sigic. proyecto. objetivo
PROYECTO_CONDUCE_GRADO	CHAR(2)	Campo que indica si el proyecto conduce a grado o no: S(SI), N(NO)	Base de Dato Sigic. Proyecto. conduce_grado
PROYECTO_MODALIDAD	VARCHAR(10)	Campo que indica si el proyecto puede es individual o grupal: INDIVIDUAL, GRUPAL	Base de Dato Sigic. Proyecto. modalidad
PROYECTO_ESTADO	VARCHAR(13)	Campo que indica el estado del proyecto: CONCLUIDO, EN DESARROLLO	Base de Dato Sigic. Proyecto. estado_proyecto
PROYECTO_RESUMEN	TEXT	Resultado del proyecto	Base de Dato Sigic. Proyecto. resumen

Cuadro 4.6: Dimensión de Proyecto

- **Dimensión Producto.** Permite ver el producto con sus atributos que lo definen como se muestra en la figura 4.7 .

Atributo	Tipo de Dato	Descripción	Procedencia
PRODUCTO_SK	SERIAL (clave primaria) .	Identificador univoco de una instancia.	Generado en el proceso ETL.
PRODUCTO_ID	INTEGER	Identificador de los productos.	Base de Dato Sigic. proyecto.id.
PRODUCTO_TITULO	TEXT	Titulo del producto	Base de Dato Sigic. producto.titulo
PRODUCTO_TIPO_ID	INTEGER	Identificador del tipo de producto	Base de Dato Sigic. tipo_producto. id
PRODUCTO_TIPO_NOMBRE	VARCHAR(255)	Nombre del tipo de producto	Base de Dato Sigic. tipo_producto. nombre
PRODUCTO_MODALIDAD	VARCHAR(10)	Campo que indica si el proyecto puede es individual o grupal: INDIVIDUAL, GRUPAL	Base de Dato Sigic. producto.. modalidad
PRODUCTO_RESUMEN	TEXT	Resumen del producto	Base de Dato Sigic. producto.resumen

Cuadro 4.7: Dimensión de Producto

- **Dimensión Evento.** Permite ver el evento con sus atributos que lo definen como se muestra en la figura 4.8 .

Atributo	Tipo de Dato	Descripción	Procedencia
EVENTO_SK	SERIAL (clave primaria) .	Identificador univoco de una instancia.	Generado en el proceso ETL.
EVENTO_ID	INTEGER	Identificador del evento	Base de Dato Sigic. Evento.id.
EVENTO_NOMBRE	VARCHAR(255)	Nombre del evento	Base de Dato Sigic. producto. nombre
EVENTO_TIPO_EVENTO_ID	INTEGER	Identificador del tipo de evento al que pertenece el evento	Base de Dato Sigic. tipo_evento. id
EVENTO_TIPO_EVENTO_NOMBRE	VARCHAR(255)	Nombre del tipo de evento	Base de Dato Sigic. tipo_evento. nombre

Cuadro 4.8: Dimensión de Evento

- **Dimensión Nivel Académico.** Permite ver el nivel de académico con sus atributos que lo definen como se muestra en la figura 4.9 .

Atributo	Tipo de Dato	Descripción	Procedencia
NIVEL_ACADEMICO_SK	SERIAL (clave primaria)	Identificador univoco de una instancia.	Generado en el proceso ETL.
NIVEL_ACADEMICO_ESTUDIO_GRADO_ACADEMICO_ID	INTEGER	Identificador del grado académico	Base de Dato Sigic. grado_academico. id
NIVEL_ACADEMICO_NOMBRE	VARCHAR(255)	Nombre del nivel académico	Base de Dato Sigic. nivel_academico. nombre
NIVEL_ACADEMICO_JERARQUIA	INTEGER	Jerarquía del nivel académico	Base de Dato Sigic. nivel_academico.jerarquia
NIVEL_ACADEMICO_GRADO_ACADEMICO_NOMBRE	VARCHAR(255)	Nombre del nivel académico	Base de Dato Sigic. grado_academico.nombre
NIVEL_ACADEMICO_GRADO_ACADEMICO_TITULO	CHAR(2)	Campo que indica si tiene título o no: S(SI), N(NO)	Base de Dato Sigic. grado_academico.titulo

Cuadro 4.9: Dimensión Nivel Académico

- Tabla de Hecho investigación.** El modelo dimensional tipo estrella busca la representación de los proyectos de investigación. Por tanto la medida en el modelo dimensional descrito anteriormente es la incidencia de un hecho (tabla de hechos sin hechos) es llamado FACTLESS y no medidas específicas. Por tanto la tabla de hechos, como se observa en la tabla 4.10 está conformada por un conjunto de claves foráneas que enlazan al hecho con las dimensiones, una clave que identifica unívocamente ese hecho y un atributo medida con valor por defecto 1, que ayuda a comprender que cada instancia tiene el mismo valor.

Atributo	Tipo de Dato	Descripción	Procedencia
HECHO_PROYECTO_SK	SERIAL (clave primaria)	Identificador univoco de una instancia.	Generado en el proceso ETL.
FECHA_INICIO_SK	INTEGER	Asociación con la dimensión tiempo.	Generado en el proceso ETL.
FECHA_FIN_SK	INTEGER	Asociación con la dimensión tiempo.	Generado en el proceso ETL.
FECHA_PUBLICACION_SK	INTEGER	Asociación con la dimensión tiempo.	Generado en el proceso ETL.
FECHA_TITULACION_SK	INTEGER	Asociación con la dimensión tiempo.	Generado en el proceso ETL.
INVESTIGADOR_SK	INTEGER	Asociación con la dimensión investigador.	Generado en el proceso ETL.
AREA_CONOCIMIENTO_SK	INTEGER	Asociación con la dimensión de área de conocimiento.	Generado en el proceso ETL.
AREA_APLICACION_SK	INTEGER	Asociación con la dimensión área de aplicación.	Generado en el proceso ETL.
INSTITUCION_SK	INTEGER	Asociación con la dimensión institución.	Generado en el proceso ETL.
PROYECTO_SK	INTEGER	Asociación con la dimensión proyecto.	Generado en el proceso ETL.
EVENTO_SK	INTEGER	Asociación con la dimensión evento.	Generado en el proceso ETL.
NIVEL_ACADEMICO_SK	INTEGER	Asociación con la dimensión nivel académico.	Generado en el proceso ETL.
MEDIDA	INTEGER	Forma de esclarecer que la medida de un evento es igual para cualquier instancia (1 por defecto).	Generado en el proceso ETL.

Cuadro 4.10: Tabla de Hecho Gestión de los proyectos

Ya con la definición del modelo dimensional tipo estrella, contempla la solución de inteligencia de negocio y con los hechos necesarios para cumplir con los objetivos trazados,

se realizó la elaboración del modelo dimensional definitivo que puede apreciarse.

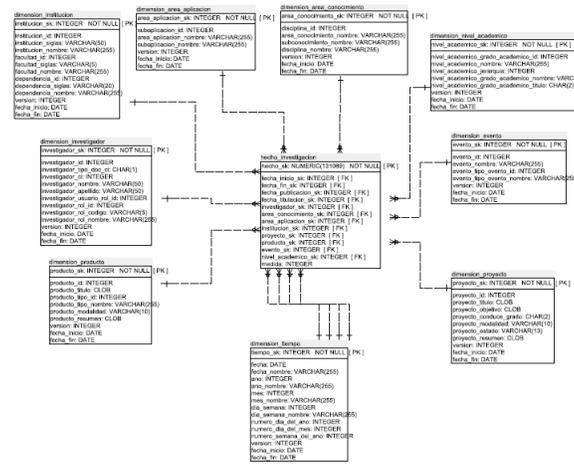


Figura 4.6: Modelo definitivo tipo estrella

4.6. Diseño físico

Una vez que se definió la arquitectura, las herramientas y el modelo dimensional a utilizar en la solución de inteligencia de negocio se continuó con la realización del diseño físico.

El almacén de datos se desarrolló físicamente usando PostgreSQL y basándose en el modelo dimensional realizado y descrito con anterioridad.

En la primera instancia, usando PostgreSQL, se definieron cada una de las dimensiones contempladas en el modelo dimensional, estableciendo inicialmente el nombre de la dimensión, luego sus atributos y por último sus niveles jerárquicos.

Para crear la dimensión tiempo se utilizó Data Integration de Pentaho que permite generar está dimensión. En este caso se especificó el nombre que se deseaba para dicha dimensión y los atributos a considerar (ano, mes y día) a partir del año 2000, de esta forma la herramienta se generó las fechas con su respectiva jerarquía como vemos en la figura 4.7.

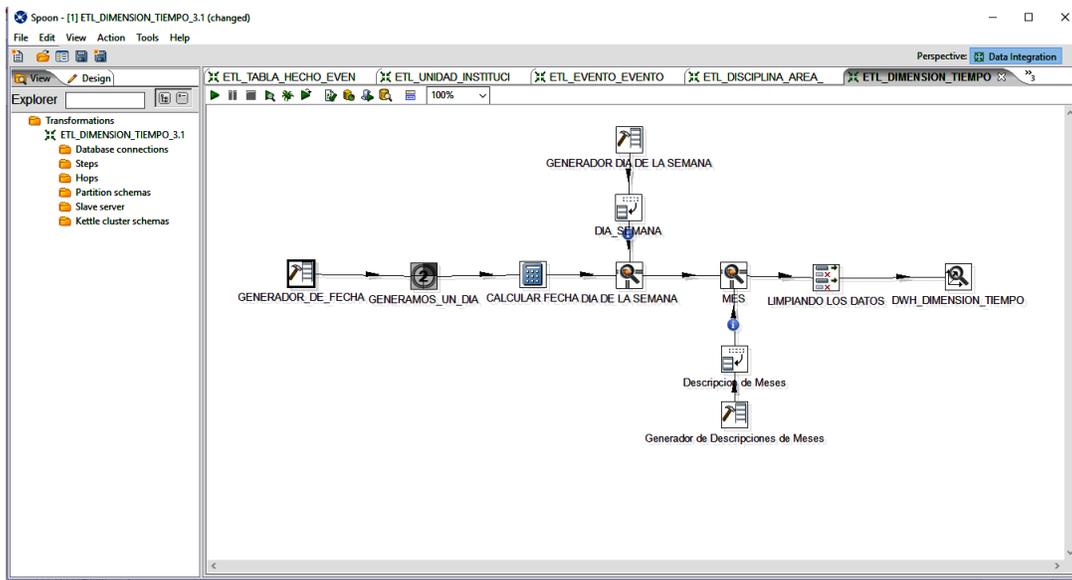


Figura 4.7: Dimensión Tiempo

Una vez completado dicho diseño se realizó el despliegue del mismo usando PostgreSQL herramienta seleccionada, de esta forma se crearon todas tablas y elementos que componen el almacén de datos propuesto para la solución de inteligencia de negocio.

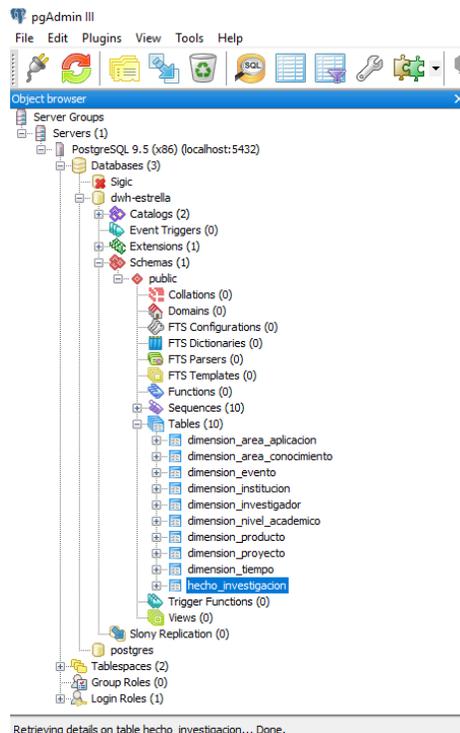


Figura 4.8: Diseño lógico de las dimensiones

Cabe destacar que al crear las tablas para el almacén de datos, se agregaron una serie de campos o columnas necesarias para el correcto funcionamiento de las distintas aplicaciones.

Estos campos adicionales no interfieren de ninguna manera con el modelo dimensional propuesto para la solución de inteligencia de negocio, tanto los requerimientos como los objetivos establecidos pueden cumplirse aún con las modificaciones de las dimensiones, que son los atributo que tienen por nombre (versión, fecha inicio, fecha_fin). Como vemos en figura 4.9

The screenshot shows the Spoon ETL tool interface. The main window displays a job design with three steps: 'BD_TABLA_INVESTIGADOR', 'INVESTIGADOR', and 'DWH_DIMENSION_INVESTIGADOR'. A 'Dimension Lookup / Update' dialog box is open, showing configuration for a dimension table. The dialog includes fields for 'Version field' (set to 'version'), 'Stream Datefield', 'Date range start field' (set to 'fecha_inicio'), 'Use an alternative start date?' (checked, set to 'System date'), and 'Table date range end' (set to 'fecha_fm'). Below the dialog, the 'Execution Results' window shows a table of data rows.

#	usuario_id	usuario_tipo_doc_ci	usuario_ci	usuario_nombre	usuario_apellido	usuario_rol_id	rol_id	rol_codigo	rol_nombre
1	1	V	20227714	VIVIANA	PÉREZ	1	1	I	INVESTIGADOR
2	2	V	3182096	CARLOS	AVESTA	4	1	I	INVESTIGADOR
3	3	V	13067646	JOHANNA	BALLESTERO	5	1	I	INVESTIGADOR
4	4	V	4350664	DANIEL	CARJEVSCI	6	1	I	INVESTIGADOR
5	5	V	3850023	SANTIAGO	GÓMEZ	7	1	I	INVESTIGADOR
6	6	V	13600453	ROSCHMAN	GONZÁLEZ	8	1	I	INVESTIGADOR

Figura 4.9: ETL de la Dimension Investigador

El resultado del despliegue del modelo dimensional realizado puede apreciarse en la figura 4.10

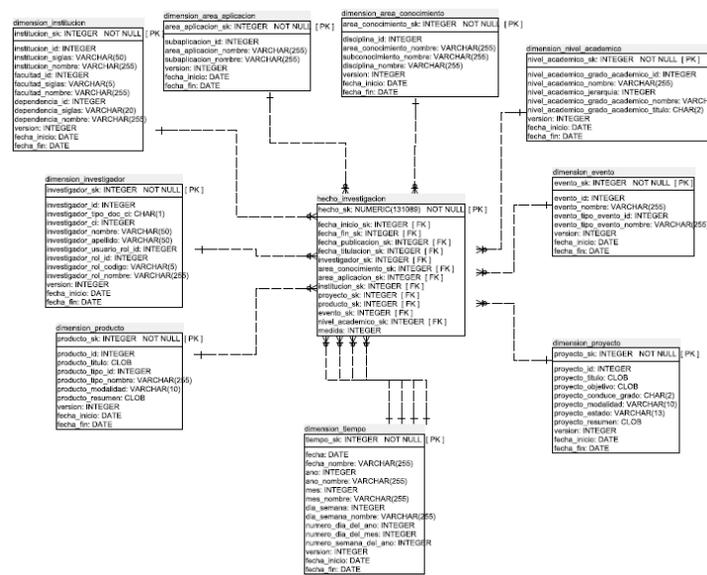


Figura 4.10: Modelo dimensional

4.7. Diseño y Construcción de procesos ETL

Dado a la arquitectura propuesta se contempla (la base de datos, y el almacén de datos), fue necesario elaborar procesos ETL para las extraer los datos de la base de datos de la coordinación.

Un proceso capaz de permitir el llenado del almacén de datos usando como fuente de datos la base de datos. Dichos procesos se muestra a continuación.

4.7.1. Diseño de Procesos de ETL y Desarrollo (almacén de datos)

El diseño de los procesos de extracción, transformación y carga para el almacén de datos se realizó utilizando la herramienta Data Integration de Pentaho (kettle). El proceso de ETL para cada una de la dimensiones, es similar al mostrado en la figura 4.11.



Figura 4.11: ETL de Dimensión Investigador

El proceso de construcción de los ETL, para el llenado de la única tabla de hecho llamada hechos investigación, conllevan a un conjunto de elementos adicionales que

serán descritos a continuación. Es requerido usar las dimensiones del almacén de datos, y extraer los hechos que se quieren analizar de la fuentes de datos, en nuestro caso la fuente es solo una base de datos.

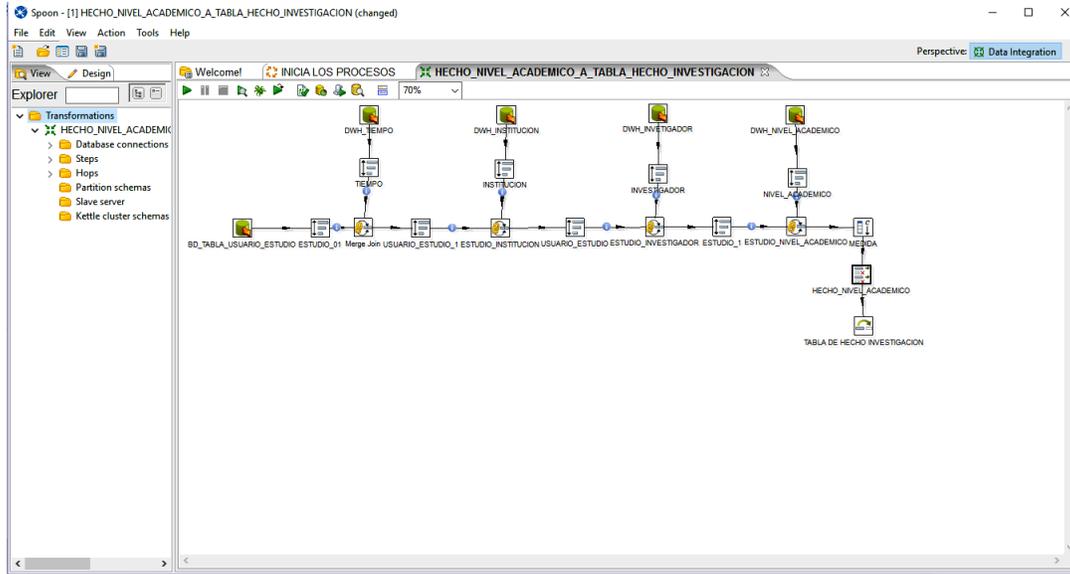


Figura 4.12: ETL de los Hechos del nivel académico a la Tabla de Hecho de investigación

Como vemos en la figura 4.12, a partir del uso de la tabla relacional `BD_TABLA_USUARIO_ESTUDIO`, se van realizando una serie de uniones para asociar el conjunto de dimensiones. Esto porque una instancia de la tabla `BD_TABLA_USUARIO_ESTUDIO`, representa el nivel de granularidad asociado al modelo dimensional tipo estrella, lo que implica que se requiere ir agregando las claves asociadas a las dimensiones que caracterizan esa acción correctiva en particular.

Una vez realizado todas las transformaciones o ETL, se creara el flujo del proceso es decir el job que ejecutara todos los procesos según el flujo se crea como vemos en la figura 4.13

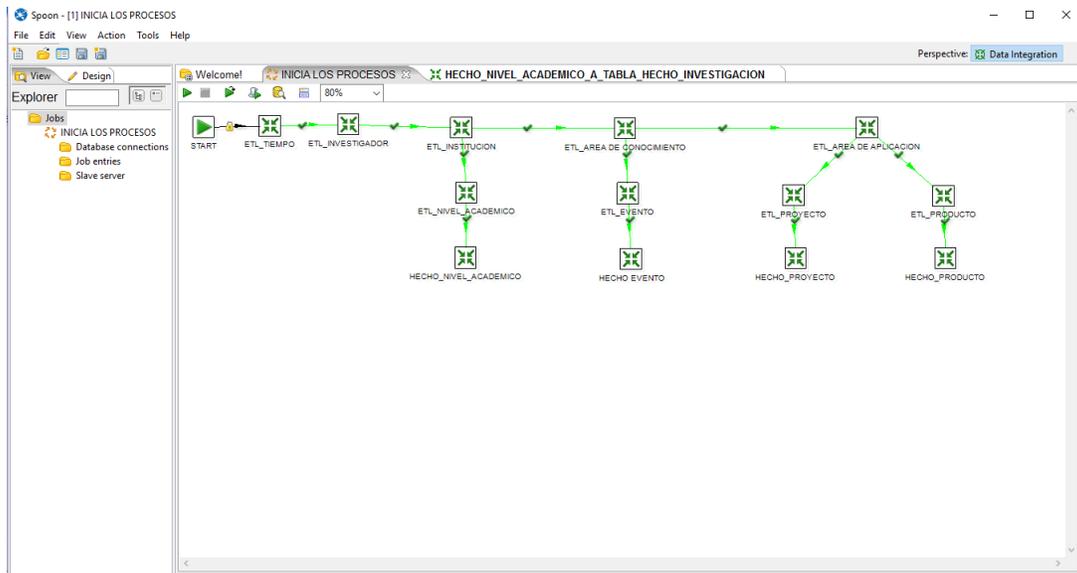


Figura 4.13: El llenado del almacén de datos

4.7.2. Verificar la calidad de los datos

Una vez se ha realizado la inserción de los datos en el almacén de datos, es necesario comprobar que la carga ha sido correcta y que el almacén de datos representa de forma acertada los datos del modelo dimensional, para esto se realiza la verificación consistente en la aplicación de un conjunto de consultas en ambos ambientes y que estén cargados de forma adecuados los datos del modelo relacional de la base de datos fuente. Se muestra la comparación de cada elemento del modelo dimensional que se encuentre relacionado con la base de datos.

- **Dimensión investigador.** Se extraer los datos del usuario que solo tienen el rol investigador, se muestra en la figura 4.14 los primeros 19 registros de ambas tablas y la consulta de la cantidad de valores existentes en cada una de las tablas como vemos en la figura 4.15.

investigador_id	investigador_nombre	investigadorRol_nombre
1	1 VIVIANA	INVESTIGADOR
2	2 CARLOS	INVESTIGADOR
3	3 JOHANNES	INVESTIGADOR
4	4 DANIEL	INVESTIGADOR
5	5 SAMUEL	INVESTIGADOR
6	6 ROCHONDO	INVESTIGADOR
7	7 MELISSA	INVESTIGADOR
8	8 WYCOL	INVESTIGADOR
9	9 MARIALIA	INVESTIGADOR
10	10 MARIELA	INVESTIGADOR
11	11 PEDRO	INVESTIGADOR
12	12 JERON	INVESTIGADOR
13	13 PAOLA	INVESTIGADOR
14	14 CARIBAY	INVESTIGADOR
15	15 TERESA	INVESTIGADOR
16	16 CARLOS	INVESTIGADOR
17	17 HERIBERTO	INVESTIGADOR
18	18 DANIE	INVESTIGADOR
19	19 ADRIAN	INVESTIGADOR

Figura 4.14: Verificación de la dimensión investigador



Figura 4.15: Verificación de la cantidad de los datos de la dimensión investigador

- Dimensión área de conocimiento.** Esta compuesta por tres tablas en la base de datos, que es la tabla disciplina, subárea de conocimiento y área de conocimiento. Donde podemos ver las primeras nueve filas de ambas tablas y la cantidad de valores existentes en cada una de las tablas.

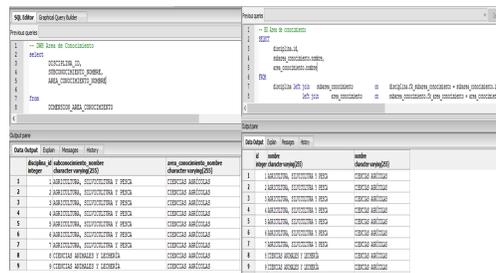


Figura 4.16: Verificación de la dimensión área de conocimiento

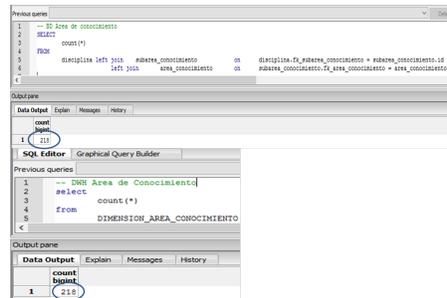


Figura 4.17: Verificación de la cantidad de datos de la dimensión área de conocimiento

- Dimensión área de aplicación.** Dicha dimensión esta compuesta de dos tablas de las base de datos, que es la tabla de subárea de aplicación, y área de aplicación.

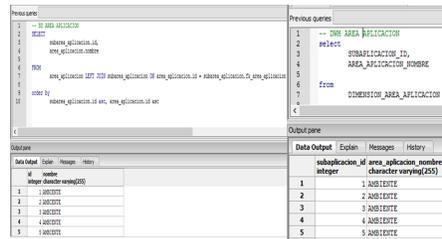


Figura 4.18: Verificación de la dimensión área de aplicación

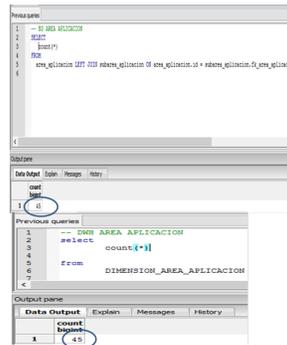


Figura 4.19: Verificación de la cantidad de datos de área de aplicación

- **Dimensión institución.** Esta compuesta de tres tabla a partir de la dependencia, facultad, y institución.

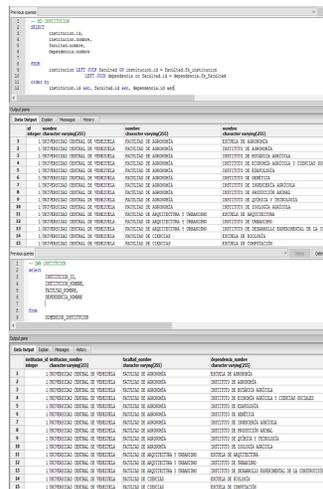


Figura 4.20: Verificación de la dimensión institución

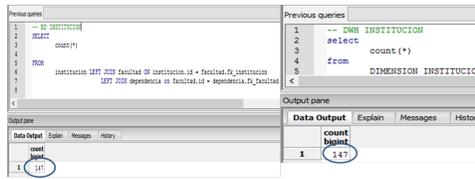


Figura 4.21: Verificación de la cantidad de datos de la dimensión institución

- **Dimensión proyecto.** La validación de los datos se ha realizado observando la tabla de proyecto.

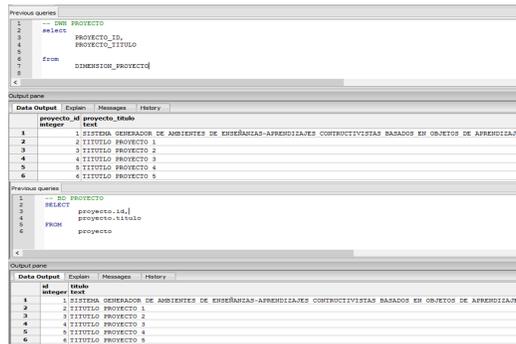


Figura 4.22: Verificación de la dimensión proyecto

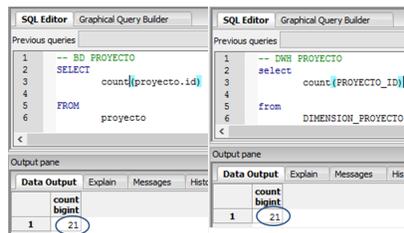


Figura 4.23: Verificación de la cantidad de datos de la dimensión proyecto

- **Dimensión producto.** La validación de los datos se ha realizado observando los productos y su tipo.

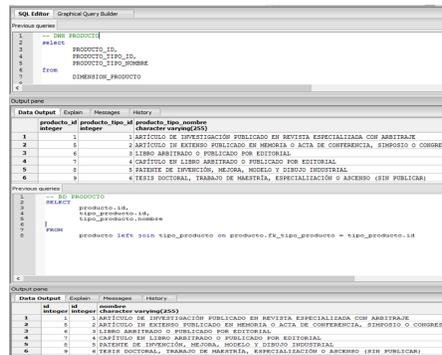


Figura 4.24: Verificación de la dimensión producto

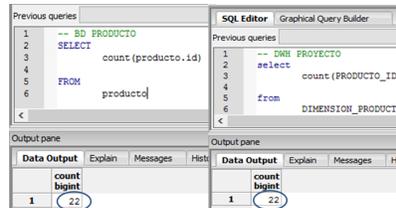


Figura 4.25: Verificación de la cantidad de datos de la dimensión producto

- **Dimensión evento.** La validación de datos se ha realizado observando los eventos en función de su nombre y tipo de evento.

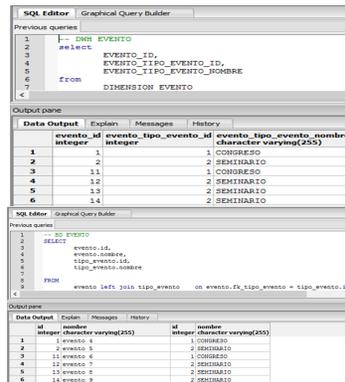


Figura 4.26: Verificación de la dimensión evento

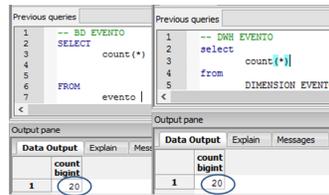


Figura 4.27: Verificación de la cantidad de datos de la dimensión evento

- **Dimensión nivel académico.** La validación de los datos se ha realizado observando la tabla de grado académico, y nivel académico

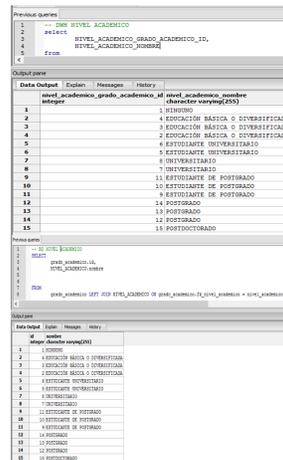


Figura 4.28: Verificación de la cantidad de datos de la dimensión nivel académico

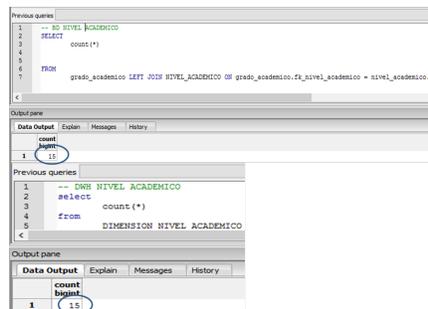
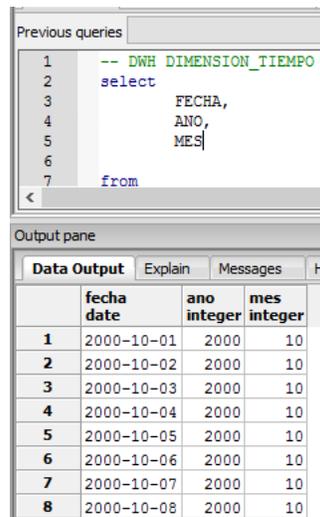


Figura 4.29: Verificación de la cantidad de datos de la dimensión nivel académico

- **Dimensión tiempo.** La dimensión no requiere una verificación de datos con respecto a la base de datos, sin embargo puede confirmarse que los datos son consistentes en función de una fecha.



```

1  -- DWH DIMENSION_TIEMPO
2  select
3      FECHA,
4      ANO,
5      MES
6
7  from

```

	fecha date	ano integer	mes integer
1	2000-10-01	2000	10
2	2000-10-02	2000	10
3	2000-10-03	2000	10
4	2000-10-04	2000	10
5	2000-10-05	2000	10
6	2000-10-06	2000	10
7	2000-10-07	2000	10
8	2000-10-08	2000	10

Figura 4.30: Verificación de datos de la dimensión tiempo

4.7.3. Verificación de la calidad de los datos de los componentes del modelo implementado

Adicionalmente a la verificación de los datos por tabla, se ha realizado una verificación para comprobar que efectivamente los datos del modelo relacional han sido transformados y cargados en el modelo dimensional tipo estrella, ya que existe una única tabla de hechos.

Para esto hay que comprender que la realización del modelo de estrella, la tabla de hecho y dimensiones, generan nueva clave primaria llamada clave subrogada para el registro de cualquier elemento. Por lo que, la inserción de dicho elemento en la tabla de hechos conlleva a que se haga una relación a partir de estas claves generadas en el proceso ETL. La verificación del modelo estrella implica comprobar que las distintas claves subrogadas en la tabla de hecho, corresponden efectivamente a los datos que se quiere enlazar en las distintas dimensiones. Se realiza del modo que la tabla de hecho se relaciona con una dimensión, y está se encuentra relacionada con ella en la base de datos.

4.7.3.1. Verificación de la tabla de hecho investigación con sus dimensiones

Como es un modelo dimensional tipo estrella, todas las dimensiones se relacionan con la tabla de hechos. Así que se realiza la verificación de esas tablas con la tabla de hecho investigación.

- Dimensión tiempo.** Para la comprobación de que efectivamente se están representando los datos del modelo relacional con respecto a la dimensión tiempo. Se muestra un registro de la tabla de proyecto de la base de datos de la fecha de inicio y de fin del proyecto y la tabla de hecho producto con eso mismo valor

asociado una clave subrogada de la dimensión tiempo.

fecha de proyecto en la Base de datos

id	fecha_inicio	fecha_fin	dwh_fecha	dwh_fecha_sk	dwh_fecha_1	dwh_fecha_sk_1
1	2009/01/01 04:30:00.000	2012/01/01 04:30:00.000	2009/01/01 04:30:00.000	3015	2012/01/01 04:30:00.000	4110
2	2010/01/01 04:30:00.000	2010/03/01 04:30:00.000	2010/01/01 04:30:00.000	3380	2010/03/01 04:30:00.000	3439
3	2010/01/02 04:30:00.000	2010/03/02 04:30:00.000	2010/01/02 04:30:00.000	3381	2010/03/02 04:30:00.000	3440
4	2010/01/03 04:30:00.000	2010/03/03 04:30:00.000	2010/01/03 04:30:00.000	3382	2010/03/03 04:30:00.000	3441

fecha de proyecto en la tabla de hecho investigación

Figura 4.31: Comparación de los dato del modelo con la dimensión tiempo

- **Dimensión proyecto.** Para la comprobación de dicha dimensión se realiza dado unos registros como en la figura 4.33, sus valores en la dimensión y su veracidad como en la figura 4.32 en la base de datos.

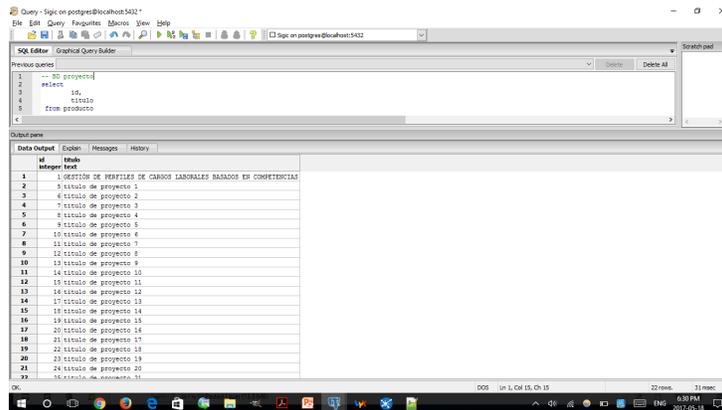


Figura 4.32: Proyecto en la base de datos

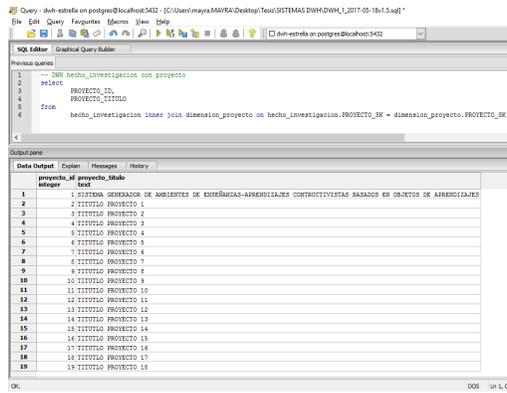


Figura 4.33: Dimensión proyecto con la tabla de hecho investigación

- **Dimensión área de conocimiento.** Para la comprobación de dicha dimensión se realiza unos registros como en la figura 4.34, sus valores en la dimensión y su veracidad como en la figura 4.35 en la base de datos.

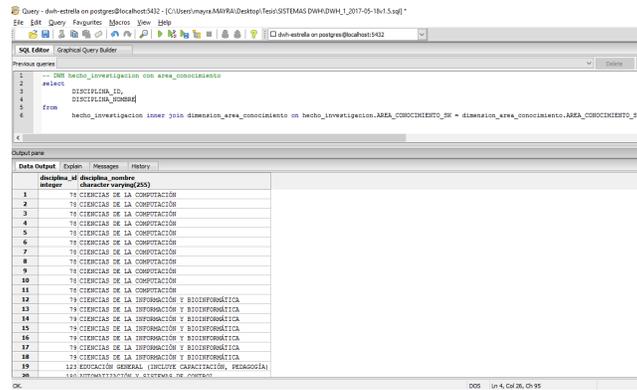


Figura 4.34: Dimensión área de conocimiento con la tabla de hecho investigación

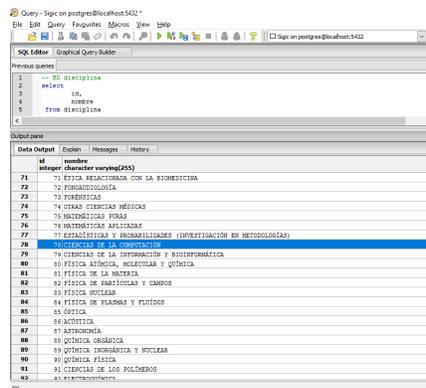


Figura 4.35: Área de conocimiento en el almacén de datos

- **Dimensión área de aplicación.** Para la comprobación de dicha dimensión se realiza unos registros como en la figura 4.37, sus valores en la dimensión y su veracidad como en la figura 4.36 en la base de datos.

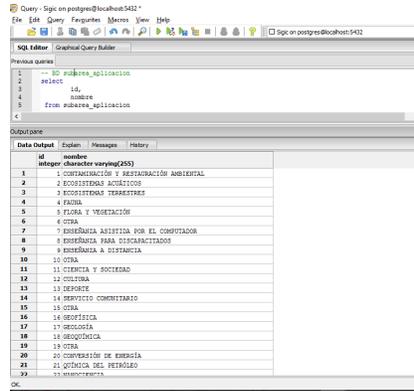


Figura 4.36: Subárea de aplicación en la base de datos

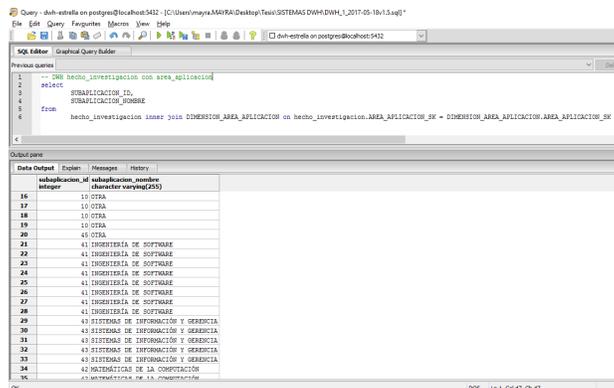


Figura 4.37: Dimensión área de aplicación con la tabla de hecho investigación

- **Dimensión institución.** Para la comprobación de dicha dimensión se realiza unos registros como en la figura 4.39 , sus valores en la dimensión y su veracidad como en la figura 4.38 en la base de datos.


```

1 -- id usuario
2 select
3 id,
4 documento_identidad,
5 primer_nombre
6 from usuario

```

id	documento_identidad	primer_nombre	apellido
199	199	14379213	RODRIGUEZ
200	200	14379213	RODRIGUEZ
201	201	7033950	KELIO
202	202	3984619	BRUNSON
203	203	15918114	ZOHMSTRAM
204	204	15912761	RODRIGUEZ
205	205	14300614	ABRILANA
206	206	17148139	LOPEZ
207	207	15168693	RODRIGUEZ
208	208	3948474	RICARDO
209	209	14220265	OSHEE
210	210	15150494	ISHERE
211	211	18030914	COMTELLA
212	212	6162614	FRANCISCO
213	213	1591444	RODRIGUEZ
214	214	8907514	OSHEE
215	215	6144831	RODRIGUEZ DE LOS ANGELES
216	216	1020195	RODRIGUEZ
217	217	4974679	RODRIGUEZ
218	218	4616714	RODRIGUEZ

Figura 4.41: Dimensión investigador con la tabla de hecho investigador

4.7.3.2. Detallar los requerimientos según el modelo

Desde el estudio del modelo dimensional, que plasma los datos relevantes de la base de datos, específicamente los proyectos, productos, eventos, y el nivel académico, y utilizando los gráficos que se estaban generando en la base de datos para buscar satisfacer las necesidades analíticas, se han propuesto un conjunto de requerimientos analíticos presentados a continuación:

1. Indicadores

- Conteo de proyectos por investigador.
- Conteo de proyectos por área de conocimiento.
- Conteo de proyectos por área de aplicación.
- Conteo de proyectos por institución.
- Conteo de producto por investigador.
- Conteo de producto por área de conocimiento.
- Conteo de producto por área de aplicación.
- Conteo de producto por institución.
- Conteo de evento por investigador.

2. Reportes

- Los proyectos del investigador.
- Los proyectos por área de conocimiento.
- Los proyectos de la institución.
- Los productos del investigador
- Los productos del área de aplicación
- Los eventos de los investigadores
- Nivel académico del investigador.

4.8. Desarrollo de los requerimientos de información

Esta fase consiste en la creación de los indicadores de gestión usando la herramienta de inteligencia de negocio (Tableau), de forma tal que puedan satisfacerse los requerimientos analíticos planteados previamente. Para esto se debe configurar desde la herramienta a utilizar la conexión a la base de datos en este caso es la conexión a PostgreSQL, realizar las consultas que satisfacen los requerimientos analíticos para luego de forma sencilla poder visualizar los distintos indicadores dispuesto de una forma ordenada. Por último se realiza una verificación de los datos para confirmar la exactitud de los indicadores.

4.8.1. Configuración de la conexión

Implica integrar la implementación del modelo dimensional ubicado en el almacén de datos PostgreSQL en Tableau, de forma que se puedan manipular los datos dentro de la herramienta de inteligencia de negocio.

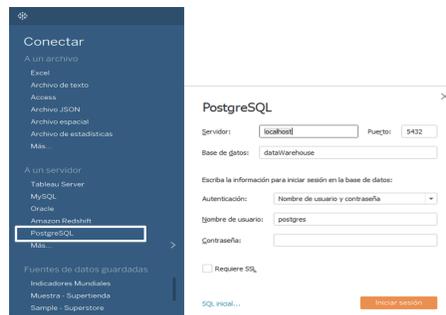


Figura 4.42: Creación de la conexión con la fuentes de dato desde Tableau

Tableau enlaza las distintas tablas a partir de atributos clave, un atributo clave es aquel que su nombre que se repite en dos tablas, Tableau realiza una unión entre dichas tablas por lo que se puede navegar entre los atributos de dichas tablas.

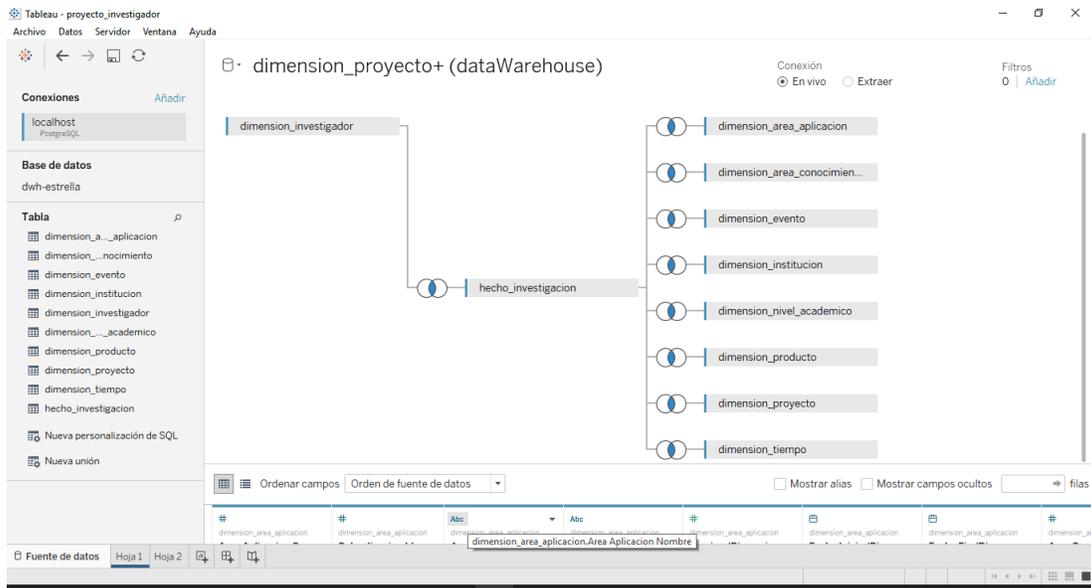


Figura 4.43: La tabla de hecho investigación con sus dimensiones

A continuación se detallara los elementos que conforman la fuente de datos.

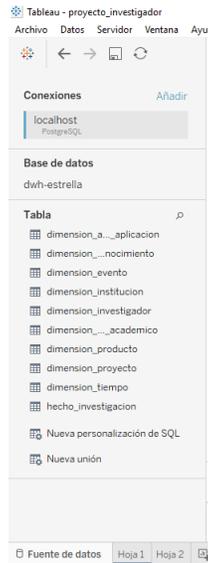


Figura 4.44: Elemento de almacén de datos

4.8.2. Generador de los indicadores de gestión

Disponibles los datos en la herramienta de inteligencia de negocio se procede a desarrollar los indicadores de gestión son indicadores eficiente, es decir miden los resultados en fin que se cumplan con la propuesta de los requerimientos analíticos señalados

previamente. Y mostrar los reportes que fueron solicitados

La construcción de los indicadores de gestión implicó la evaluación de las distintas formas como pueden observarse los mismos y la forma de distribuir los indicadores de gestión dentro de la fuente de datos de forma que se tuviera una lógica al ser observado.

4.8.3. Desarrollo y distribución de consultas

Para la realización de las consultas, se ha optado por seleccionar la tabla de hecho investigación con sus dimensiones asociadas, para luego observar las dimensiones con sus hechos.

Una de las ventajas que ofrece Tableau es que permite las agrupaciones de las dimensiones a las tabla de hecho siempre que esta tenga una relación con esas tablas. Se muestran a continuación.

4.8.3.1. Tabla de hecho de investigación

Se presentan todas las tablas con que se relaciona la tabla de hecho investigación, que se desea analizar para responder los indicadores y mostrar los reportes nombrado anteriormente.

1. Indicadores

- a) **Proyectos por investigador.** Se puede observar desde la fuente de datos la figura 4.44, permite seleccionar la unión que se desea tener. lo que se quiere analizar y observar.

Conteo proyecto por investigador				SUMA(Número de registr...	
Proyecto M.	Apellido	Investigado..	Nombre		
GRUPAL	MORGADO	8	MYLOA	1	
	ORTEGA	9	NATALIA	1	
	PETRALANDA	10	MARÍA	1	
	RODRIGUEZ	11	PEDRO	1	
	ROMERO	12	JESÚS	1	
	SANDOVAL	253	FRANKLIN	1	
	TONINO	13	PAULA	1	
	URBINA	14	CARIBAY	1	
	INDIVIDUAL	ACOSTA	16	CARLOS	1
		ASTUDILLO	17	REINALDO	1
AYESTA		2	CARLOS	1	
BALLESTERO		3	JOHANNA	1	
BLANCO		18	JAIIME	1	
CARJEVSCI		4	DANIEL	1	
GÓMEZ		5	SANTIAGO	1	
GONZÁLEZ		6	ROSCHMAN	1	
MARQUES		7	SHEILA	1	
PÉREZ		1	VIVIANA	1	
VARGAS	15	TERESA	1		
Total general				19	

Figura 4.45: Conteo de proyectos del investigador

- b) **Proyectos por área de conocimiento.** Se provee las área de conocimiento para la clasificación de los proyectos según las áreas.

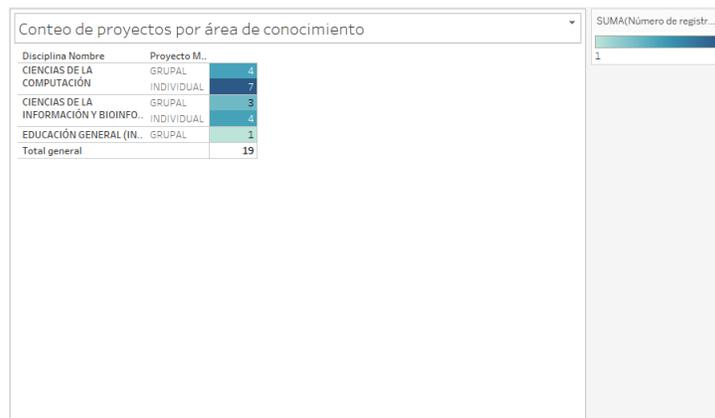


Figura 4.46: Conteo de proyectos por área de conocimiento

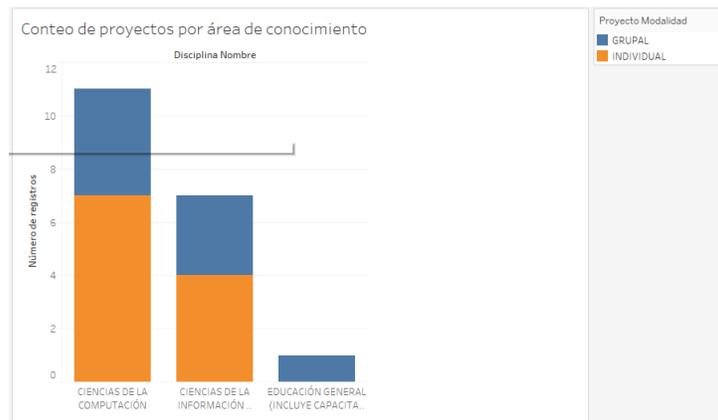


Figura 4.47: Gráfica de conteo de proyectos del investigador

- c) **Proyectos por área de aplicación.** Provee las áreas de aplicación para saber las distintas áreas que siguen los proyectos.



Figura 4.48: Conteo de proyectos por área de aplicación

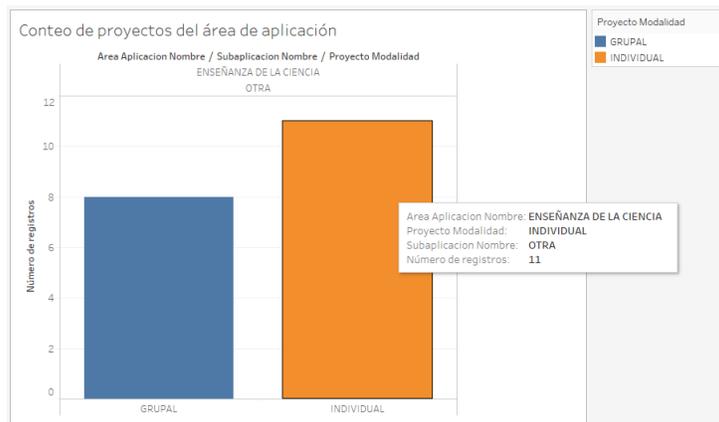


Figura 4.49: Gráfica de conteo de proyectos por área de aplicación

- d) **Proyectos por área de institución.** Provee las instituciones a la que un proyecto puede pertenecer.



Figura 4.50: Conteo de proyectos por institución

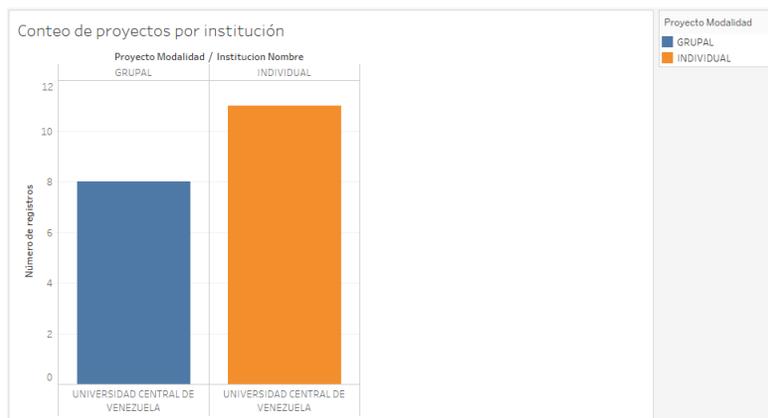


Figura 4.51: Gráfica de conteo de proyectos por institución

- e) **Producto por investigador.** Se puede observar desde la fuente de datos la figura 4.44, permite seleccionar la unión que se desea tener. lo que se quiere analizar y observar.

Institucion Nombre	Institucion ..	Producto M..	
UNIVERSIDAD CENTRAL DE VENEZUELA	UCV	GRUPAL	11
		INDIVIDUAL	11
Total general			22

SUMA(Número de registr...
11

Figura 4.53: Cantidad de productos por institución

Investigado..	Producto M..	Apellido	Investigado..	Nombre	
2	GRUPAL	AYESTA	2	CARLOS	1
3	GRUPAL	BALLESTERO	3	JOHANNA	1
4	GRUPAL	CARJEVSCI	4	DANIEL	1
5	GRUPAL	GÓMEZ	5	SANTIAGO	1
6	GRUPAL	GONZÁLEZ	6	ROSCHMAN	1
7	GRUPAL	MARQUES	7	SHEILA	1
8	GRUPAL	MORGADO	8	MYLOA	1
9	GRUPAL	ORTEGA	9	NATALIA	1
10	GRUPAL	PETRALANDA	10	MARÍA	1
11	INDIVIDUAL	RODRIGUEZ	11	PEDRO	1
12	INDIVIDUAL	ROMERO	12	JESÚS	1
13	INDIVIDUAL	TONINO	13	PAULA	1
14	INDIVIDUAL	URBINA	14	CARIBAY	1
15	INDIVIDUAL	VARGAS	15	TERESA	1
16	INDIVIDUAL	ACOSTA	16	CARLOS	1
17	INDIVIDUAL	ASTUDILLO	17	REINALDO	1
18	INDIVIDUAL	BLANCO	18	JAIIME	1
19	INDIVIDUAL	BOTTINI	19	ADRIÁN	1
20	INDIVIDUAL	CARBALLO	20	YUSNEYI	1
21	INDIVIDUAL	CARMONA	21	RHADAMES	1
253	GRUPAL	SANDOVAL	253	FRANKLIN	1
Total general					22

SUMA(Número de registr...
1

Figura 4.52: Conteo de producto del investigador

- f) **Producto por área de institución.** Provee las instituciones a la que un producto puede pertenecer.
- g) **Producto por área de conocimiento.** Se provee las área de conocimiento para la clasificación de los productos según las áreas.



Figura 4.54: Conteo de productos por área de conocimiento

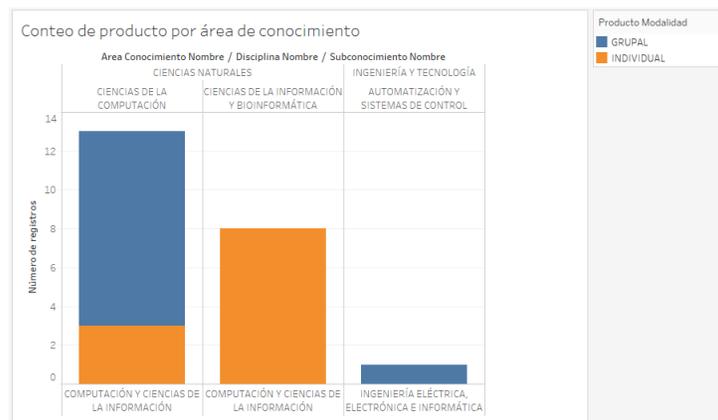


Figura 4.55: Gráfica de conteo de productos por área de conocimiento

- h) **Producto por área de aplicación.** Provee las área de aplicación para saber las distintas áreas que siguen los productos.

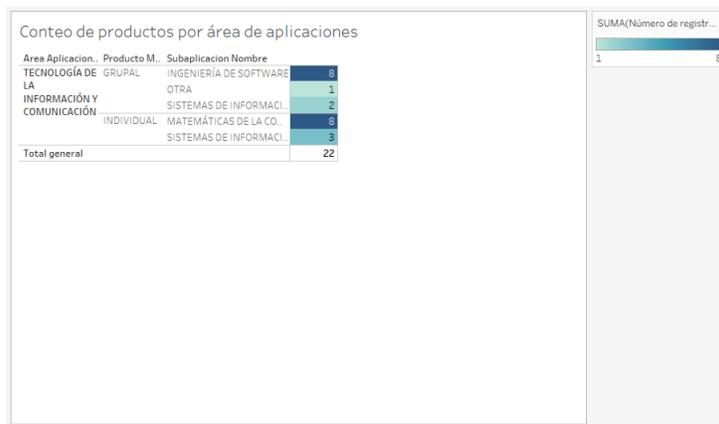


Figura 4.56: Conteo de productos por área de aplicación

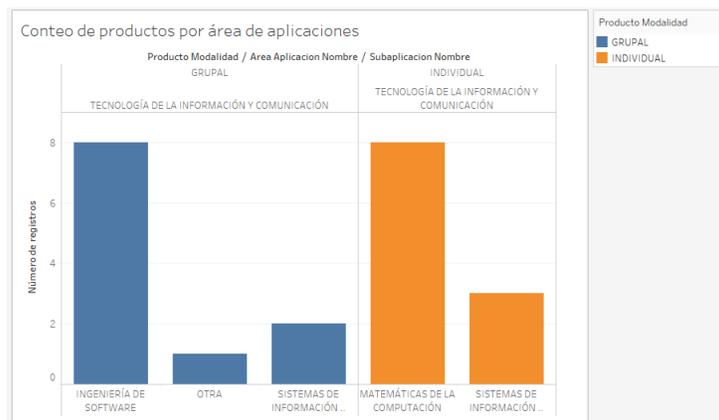


Figura 4.57: Gráfica de conteo de productos por área de aplicación

- i) **Eventos por investigador.** Se puede observar los eventos de los investigadores y el conteo de los eventos.

Conteo de eventos del investigador				SUMA(Número de registr...
Investigado..	Apellido	Nombre	Evento No..	
14	URBINA	CARIBAY	evento_15	1
15	VARGAS	TERESA	evento_6	1
			evento_11	1
16	ACOSTA	CARLOS	evento_4	1
17	ASTUDILLO	REINALDO	evento_5	1
18	BLANCO	JAIME	evento_5	1
19	BOTTINI	ADRIÁN	evento_5	1
20	CARBALLO	YUSNEYI	evento_5	1
			evento_6	1
21	CARMONA	RHADAMES	evento_4	1
22	CARRERO	EUDES	evento_5	1
23	CORREA	ELIEZER	evento_6	1
241	CASOTTO	MERIS	evento_7	1
			evento_12	1
242	GARCÍA	MARÍA	evento_8	1
			evento_13	1
243	GARCÍA	RUTH	evento_9	1
			evento_14	1
244	RAMÍREZ	JESÚS	evento_10	1
			evento_15	1
245	GIBIAQUI	MÓNICA	evento_11	1
Total general				38

Figura 4.58: Conteo de eventos por investigador

2. Reportes

a) **Proyectos por investigador.** Se puede observar en la figura 4.59, los proyectos del investigador

Proyecto del investigador				Proyecto Modalidad		SUMA(Número de registr...
Investigado..	Apellido	Nombre	Proyecto Titulo	GRUPAL	INDIVIDUAL	
1	PÉREZ	IVIANA	TITULO_PROYECTO_1		1	1
2	AYESTA	CARLOS	TITULO_PROYECTO_2		1	1
3	BALLESTERO	JOHANNA	TITULO_PROYECTO_3		1	1
4	CARJEVSCHI	DANIEL	TITULO_PROYECTO_4		1	1
5	GÓMEZ	SANTIAGO	TITULO_PROYECTO_5		1	1
6	GÓNZÁLEZ	ROSCHMAN	TITULO_PROYECTO_6		1	1
7	MARQUES	SHEILA	TITULO_PROYECTO_7		1	1
8	MORGADO	MYLOA	TITULO_PROYECTO_8	1		1
9	ORTEGA	NATALIA	TITULO_PROYECTO_9	1		1
10	PETRALAN	MARÍA	TITULO_PROYECTO_10	1		1
11	RODRIGUEZ	PEDRO	TITULO_PROYECTO_11	1		1
12	ROMERO	JESUS	TITULO_PROYECTO_12	1		1
13	TONINO	PAULA	TITULO_PROYECTO_13	1		1
14	URBINA	CARIBAY	TITULO_PROYECTO_14	1		1
15	VARGAS	TERESA	TITULO_PROYECTO_15		1	1
16	ACOSTA	CARLOS	TITULO_PROYECTO_16		1	1
17	ASTUDILLO	REINALDO	TITULO_PROYECTO_17		1	1
18	BLANCO	JAIME	TITULO_PROYECTO_18		1	1
253	SANDOVAL	FRANKLIN	SISTEMA GENERADOR DE AMBIENTES DE ENSEÑANZAS.	1		1

Figura 4.59: Los proyectos del investigador

b) **Proyectos por área de conocimiento.** Se puede observar los proyectos que hay por área de conocimiento.

Proyectos por área de conocimiento		
Area Conoci.	Disciplina Nombre	Subconocimiento Nombre
CIENCIAS NATURALES	CIENCIAS DE LA COMPUTACIÓN	COMPUTACIÓN Y CIENCIAS DE LA INFORMACIÓN
	CIENCIAS DE LA INFORMACIÓN Y BIOINFORMÁTICA	COMPUTACIÓN Y CIENCIAS DE LA INFORMACIÓN
CIENCIAS SOCIALES	EDUCACIÓN GENERAL (INCLUYE CAPACITACIÓN, PEDAGOGÍA)	CIENCIAS DE LA EDUCACIÓN

Figura 4.60: Los proyectos del área de conocimiento

- c) **Proyectos de la institución.** Se puede observar los proyectos de las instituciones.

Proyectos de la institución	
Proyecto Titulo	Institucion Nombre
SISTEMA GENERADOR DE AMBIENTES DE ENSEÑ.	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_1	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_2	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_3	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_4	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_5	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_6	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_7	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_8	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_9	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_10	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_11	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_12	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_13	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_14	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_15	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_16	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_17	UNIVERSIDAD CENTRAL DE VENEZUELA
TITULO_PROYECTO_18	UNIVERSIDAD CENTRAL DE VENEZUELA

Figura 4.61: Los proyectos de las instituciones

- d) **Productos del investigador.** Se puede observar los proyectos del investigador.

Producto del investigador

Investigado..	Apellido	Nombre	Producto M..	Producto Titulo
1	PÉREZ	VIVIANA	GRUPAL	titulo de proyecto_1
2	AYESTA	CARLOS	GRUPAL	titulo de proyecto_2
3	BALLESTERO	JOHANNA	GRUPAL	titulo de proyecto_3
4	CARJEVSCI	DANIEL	GRUPAL	titulo de proyecto_4
5	GÓMEZ	SANTIAGO	GRUPAL	titulo de proyecto_5
6	GONZÁLEZ	ROSCHMAN	GRUPAL	titulo de proyecto_6
7	MARQUES	SHEILA	GRUPAL	titulo de proyecto_7
8	MORGADO	MYLOA	GRUPAL	titulo de proyecto_8
9	ORTEGA	NATALIA	GRUPAL	titulo de proyecto_9
10	PETRALANDA	MARÍA	GRUPAL	titulo de proyecto_10
11	RODRIGUEZ	PEDRO	INDIVIDUAL	titulo de proyecto_11
12	ROMERO	JESÚS	INDIVIDUAL	titulo de proyecto_12
13	TONINO	PAULA	INDIVIDUAL	titulo de proyecto_13
14	URBINA	CARIBAY	INDIVIDUAL	titulo de proyecto_14
15	VARGAS	TERESA	INDIVIDUAL	titulo de proyecto_15
16	ACOSTA	CARLOS	INDIVIDUAL	titulo de proyecto_16
17	ASTUDILLO	REINALDO	INDIVIDUAL	titulo de proyecto_17
18	BLANCO	JAIIME	INDIVIDUAL	titulo de proyecto_18
19	BOTTINI	ADRIÁN	INDIVIDUAL	titulo de proyecto_19
20	CARBALLO	YUSNEYI	INDIVIDUAL	titulo de proyecto_20
21	CARMONA	RHADAMES	INDIVIDUAL	titulo de proyecto_21
253	SANDOVAL	FRANKLIN	GRUPAL	GESTIÓN DE PERFILES DE

Figura 4.62: Los productos del investigador

e) **Productos del área de aplicación.** Se puede observar los proyectos del área de aplicación.

Productos del área de aplicación

Producto M..	Producto Titulo	Area Aplicacion Nombre	Subaplicacion Nombre
GRUPAL	GESTIÓN DE PERFILES..	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	OTRA
	titulo de proyecto_1	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	INGENIERÍA DE SOFTWARE
	titulo de proyecto_2	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	INGENIERÍA DE SOFTWARE
	titulo de proyecto_3	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	INGENIERÍA DE SOFTWARE
	titulo de proyecto_4	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	INGENIERÍA DE SOFTWARE
	titulo de proyecto_5	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	INGENIERÍA DE SOFTWARE
	titulo de proyecto_6	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	INGENIERÍA DE SOFTWARE
	titulo de proyecto_7	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	INGENIERÍA DE SOFTWARE
	titulo de proyecto_8	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	INGENIERÍA DE SOFTWARE
	titulo de proyecto_9	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	SISTEMAS DE INFORMACIÓN Y GERENCIA
	titulo de proyecto_10	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	SISTEMAS DE INFORMACIÓN Y GERENCIA
INDIVIDUAL	titulo de proyecto_11	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	SISTEMAS DE INFORMACIÓN Y GERENCIA
	titulo de proyecto_12	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	SISTEMAS DE INFORMACIÓN Y GERENCIA
	titulo de proyecto_13	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	SISTEMAS DE INFORMACIÓN Y GERENCIA
	titulo de proyecto_14	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	MATEMÁTICAS DE LA COMPUTACIÓN
	titulo de proyecto_15	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	MATEMÁTICAS DE LA COMPUTACIÓN
	titulo de proyecto_16	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	MATEMÁTICAS DE LA COMPUTACIÓN
	titulo de proyecto_17	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	MATEMÁTICAS DE LA COMPUTACIÓN
	titulo de proyecto_18	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	MATEMÁTICAS DE LA COMPUTACIÓN
	titulo de proyecto_19	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	MATEMÁTICAS DE LA COMPUTACIÓN
	titulo de proyecto_20	TECNOLOGÍA DE LA INFORMACIÓN Y COMUN..	MATEMÁTICAS DE LA COMPUTACIÓN

Figura 4.63: Los productos del área de aplicación

f) **Evento de los investigadores.** Se puede observar los eventos de los investigadores

Investigado.	Apellido	Nombre	Evento No.
14	URBINA	CARIBAY	evento_7
			evento_15
15	VARGAS	TERESA	evento_6
			evento_11
16	ACOSTA	CARLOS	evento_4
17	ASTUDILLO	REINALDO	evento_5
18	BLANCO	JAIIME	evento_5
19	BOTTINI	ADRIÁN	evento_5
20	CARBALLO	YUSNEYI	evento_5
			evento_6
21	CARMONA	RHADAMES	evento_4
22	CARRERO	EUDES	evento_5
23	CORREA	ELIEZER	evento_6
241	CASOTTO	MERIS	evento_7
			evento_12
242	GARCIA	MARÍA	evento_8
			evento_13
243	GARCIA	RUTH	evento_9
			evento_14
244	RAMÍREZ	JESÚS	evento_10
			evento_15
245	GIBIAQUI	MÓNICA	evento_11

Figura 4.64: Evento de los investigador

g) **Nivel académico por investigador.** Se puede observar desde la fuente de datos la figura 4.44, permite seleccionar la unión que se desea tener. lo que se quiere analizar y observar.

Nivel Académico	Nombre	Investigado.	Nombre	Nivel Académico Grado Ac.	Apellido
EDUCACIÓN BÁSICA O DIVERSIFICADA		201	ELIO	PRIMARIA	MÉNDEZ
		202	NELSON	SECUNDARIA	MERENTES
		203	JONNATHAN	TÉCNICO MEDIO	OTERO
ESTUDIANTE DE POSTGRADO		206	LUÍS	MAESTRÍA	PEREDES
		207	ANGIE	DOCTORADO	PINEDA
		214	JUAN	MAESTRÍA	ÁLVAREZ
POSTGRADO		200	ARGENIS	MAESTRÍA	MÉNDEZ
		208	RICARDO	ESPECIALIZACIÓN	RÍOS
		209	JOSÉ	MAESTRÍA	SÁNCHEZ
		210	IRENE	DOCTORADO	SANTOS
		211	DANIELA	DOCTORADO	TORREALBA
		212	FRANCISCO	MAESTRÍA	TOVAR
		213	MARA	ESPECIALIZACIÓN	VALERA
		253	FRANKLIN	DOCTORADO	SANDOVAL
				MAESTRÍA	SANDOVAL
UNIVERSITARIO		1	VIVIANA	LICENCIATURA O EQUIVAL.	PÉREZ
		204	ÁNGEL	LICENCIATURA O EQUIVAL.	PADILLA
		205	ADRIANA	LICENCIATURA O EQUIVAL.	PADRÓN
		215	MARÍA DE LOS ÁNGELES	LICENCIATURA O EQUIVAL.	ÁLVAREZ
		216	ROSA	LICENCIATURA O EQUIVAL.	AMARO

Figura 4.65: El nivel académico de los investigador

4.8.4. Verificar la calidad de los datos en la herramienta de manipulación y visualización de datos

Una vez generado la visualización de los datos, se realiza la verificación de los datos. En este caso fueron analizadas las distintas dimensiones lo cual implica la realización de consultas sobre la base de datos relacional ajustando los indicadores de gestión a partir de los filtros planteados las fuentes de datos para comprobar su igualdad.

- **Cantidad de proyecto.** Como en la figura 4.66, la consulta al sistema de la base de datos muestra el mismo resultado que en la consulta en Tableau.

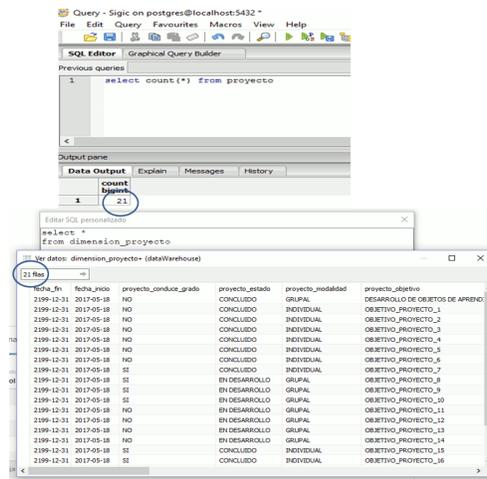


Figura 4.66: Verificación de la tabla de hecho investigación con base de datos

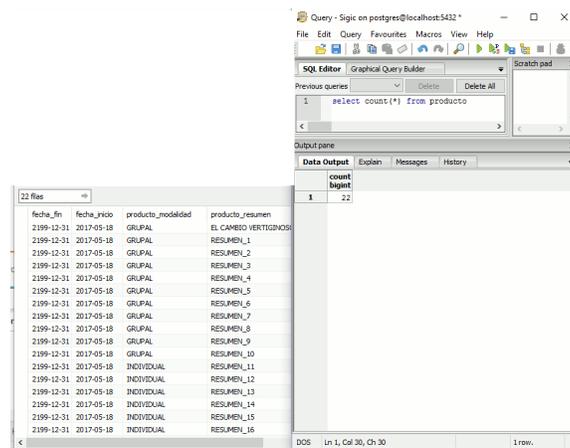


Figura 4.67: Verificación de la tabla de hecho investigación con base de datos

- **Cantidad de producto.** Como en la figura 4.67, la consulta al sistema de la base de datos muestra el mismo resultado que en la consulta en Tableau.
- **Cantidad de eventos.** Como en la figura 4.68, la consulta al sistema de la base de datos muestra el mismo resultado que en la consulta en Tableau.
- **Nivel Académico de los investigadores.** Como se observa en la figura 4.69, la consulta al sistema de la base de dato muestra el mismo resultado que la consulta en Tableau.

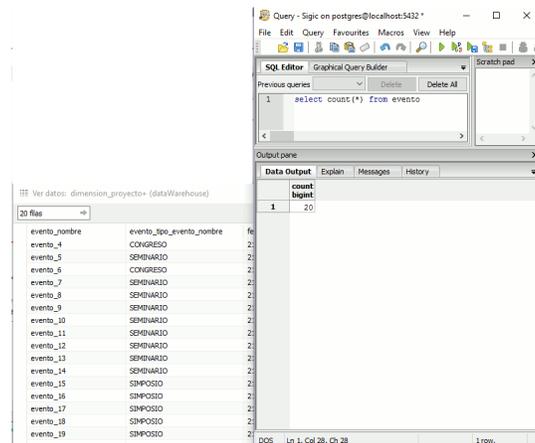


Figura 4.68: Verificación de la tabla de hecho investigación con base de datos

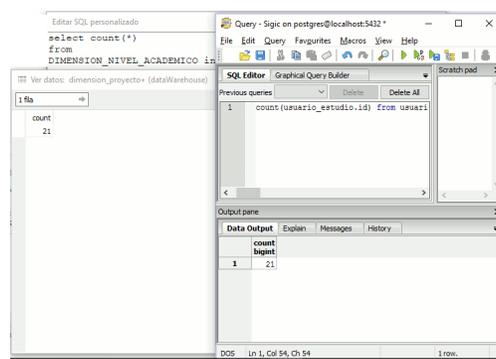


Figura 4.69: Verificación de la tabla de hecho de nivel académico con base de datos

4.9. Implementación

Debido a que el trabajo presentado forma parte de una solución de inteligencia de negocio para un trabajo especial de grado, su implementación dentro de la coordinación no trabajo con datos reales, por tanto las pruebas de aceptación no fue realizada. Esto genera que dicha fase de desarrollo de la solución no sea ejecutada de forma total.

Con estas herramientas, tal y como se explicó en el capítulo anterior, se integró la capa física con el modelo del negocio, y éste a su vez con la capa de presentación que es la que se exhibe a los usuarios.

4.9.1. Ajuste de las consultas

Para el ajuste de consultas se realizaron pruebas de visualización de los indicadores para comprobar la vialidad en la forma como se estaban desplegando los indicadores de gestión.

De estas reuniones con el tutor se plateo los indicadores y reportes, para así diseñar el modelo dimensional.

Conclusiones y recomendaciones

Los fundamentos teóricos aprendidos en la carrera y los estudios realizados para el desarrollo del seminario y la tesis fueron importantes para el cumplimiento de los objetivos planteados, dado a que de esta forma, se obtuvieron las bases sobre los conocimientos y experiencia necesarios para el desarrollo de la solución de inteligencia de negocio.

Este desarrolló es un prototipo de una solución de inteligencia de negocio para la *Coordinación de Investigación*, ya que no se logró trabajar con los datos reales.

En la construcción de la solución de negocio se desarrolló un modelo dimensional que permite observar los proyectos, productos, eventos, y nivel académico a partir de distintas perspectivas.

La utilización de la metodología propuesta por Ralph Kimball para el desarrollo de soluciones de inteligencia de negocio facilitó el cumplimiento de los objetivos planteados. Se logró desarrollar una solución adaptada a las necesidades y requerimientos de la *Coordinación de Investigación*. Sin embargo, debido a que el trabajo de investigación se efectuó en un ambiente de pruebas y no en un ambiente de producción, algunas tareas se realizaron parcialmente.

La flexibilidad a la hora de realizar informes o visualizar indicadores usando gráficos permite el análisis de datos de forma eficiente. La realización de nuevas consultas y gráficos a partir de necesidades presentadas pueden ser atendidas rápidamente. La facilidad de uso se traduce en que el sistema es capaz de resolver necesidades analíticas sin necesidad de que el usuario final recurra a programadores que tengan que generar un nuevo elemento para la obtención de información.

Como recomendaciones generales tenemos:

- Se plantea realizar un proceso de adiestramiento a los usuarios en el uso de las herramientas.
- Ajustar las consultas en función a necesidades adicionales de los usuarios e implementar efectivamente la solución en la organización.
- Se plantea establecer la conexión con la base de datos real para la implementación de los indicadores y que sea de gran utilidad para la toma de decisiones.
- Hacer uso extensivo de la herramienta Tableau, ya que es una herramienta de fácil manejo por el usuario, y entrenar a los usuarios en su uso para que puedan generar mucha más información que la mostrada en este proyecto.

Bibliografía

- [Andrade, 2005] Andrade, S. (2005). *Diccionario de Economía (3era ed.)*. Andrade.
- [CoorInv, 2011] Coordinación de Investigación. (2011). *¿Quiénes Somos?*. Caracas, Venezuela: Facultad de Ciencias de la Universidad Central de Venezuela. Recuperado en julio 2016 de: <http://www.coordinv.ciens.ucv.ve/investigacion/quienes.php>
- [Pérez, 2016] Pérez V. (2016). *GENCI-2 Sistema de Gestión de Información Científica para la Coordinación de Investigación de la Facultad de Ciencias de la UCV*. Caracas, Facultad de Ciencias de la Universidad Central de Venezuela (UCV).
- [Borges, 2006] Borges C. & Rivero A. (2006). *Generador de Sitios Web de Centros de Investigación. Centro de Ingeniería de Software y Sistemas, Facultad de Ciencias, Universidad Central de Venezuela, Caracas, Venezuela*. Recuperado en julio 2015, de: <http://www.coordinv.ciens.ucv.ve/investigacion/genci/index.php>
- [Prior, 2007] Prior V. (2007). *Glossary of Terms used in Competitive Intelligence and Knowledge Management*. Knowledge-Board.
- [Thomas, 2005] Thomas M. & Carolyn E. (2005). *“Sistemas de bases de datos”*. Pearson Education.
- [Michael, 2007] Michael, V. (2007). *Administración de bases de datos Diseño y desarrollo de aplicaciones*. México: McGRAW-HILL/INTERAMERICANA EDITORES, S.A. DE C.V.
- [Date, 2001] Date, C. (2001). *Introducción a los sistemas de Base de datos*. México: Pearson Educación

- [Ramaskrishnan, 1999] Ramakrishnan, R. & Gehrke, J. (1999). *Database Management Systems*. US: McGRAW-HILL Higher Education
- [CCM, 2016] CCM. (2016). *Administración de bases de datos (DBMS)*. Recuperado en Julio 2016 de: <http://es.ccm.net/contents/66-introduccion-bases-de-datos>
- [Peña, 2006] Peña A. (2006). *Inteligencia de negocio: Una Propuesta para su Desarrollo en las organizaciones. Primera Edición*. Instituto Politécnico Nacional. México: INSTITUTO POLITÉCNICO NACIONAL
- [Cano, 2007] Cano, J. (2007). *Business Intelligence: competir con información*. Madrid: Fundación Cultural Banesto .
- [1] Planeaux, D., & Alvin, D. (2007). *Oracle Business Intelligence Standard Edition One*. California: Oracle Parkway.
- [sinnexus, 2016] Sinnexus. (2016). *Data Warehouse*. Recuperado en Julio 2016 de: http://www.sinnexus.com/business_intelligence/datamart.aspx
- [DWH, 2010] Buigues, A. (2010). *Arquitectura de un Data Warehouse*. Recuperado en Julio 2016 de: <http://anabuigues.com/2010/03/05/arquitectura-de-un-data-warehouse/>
- [DWH, 2015] Pacheco, O. (2015). *Características de un Data Warehouse*. Recuperado en Julio 2016 de: <http://dwhucv.blogspot.com/p/caracteristicas-de-un-datawarehouse.html>
- [Inmon, 2002] Inmon, W. (2002). *Building the Data Warehouse*. Canada: Technical Publishing Group.
- [Inteligencia de Negocios, 2017] Inteligencia de Negocios. (2017). *Inteligencia de Negocios*. Recuperado en Enero de 2017, de: <http://www.idensa.com/>
- [Kimball, 2010] Espinosa, R. (2010). *15.2. Kimball vs Inmon. Ampliación de conceptos del Modelado Dimensional*. Recuperado en Julio 2016 de: <https://churriwifi.wordpress.com/2010/04/19/15-2-ampliacion-conceptos-del-modelado-dimencional>

- [Eckerson, 2012] Eckerson, W. (2012). *For Payés to Build a Data Warehouse*. Recuperado en Julio 2016 de: <http://www.bi-bestpractices.com/view-articles/4770>
- [Kimball, 2002] Kimball R. & Ross M. (2002). *The Data Warehouse Toolkit*. Canada: Wiley Computer Publishing.
- [Kimball, 1998] Kimball R. (1998). *The Data Warehouse Life Cycle Toolkit*. Canada: Wiley Computer Publishing.
- [Bernabeu, 2009] Bernabeu, R..(2009).Datawarehouse manager. Recuperado en Julio 2016 de: <http://www.dataprix.com/datawarehouse-manager#x1-520003.4.5.3>
- [PostgreSQL, 1996] The PostgreSQL Global Development Group. (1996-2015). *PostgreSQL*. Recuperado en julio 2015, de: <http://www.postgresql.org/about/>
- [2nsPostgreSQL, 2001] 2ndQuadrant Ltd. (2001-2015). *2ndQuadrant Professional PostgreSQL*. Recuperado en julio 2015, de: <http://2ndquadrant.com/es/postgresql/>
- [Rouda, 2014] Rouda, N. & Anlyst, S. (2014). *Getting Real About Big Data : Build Versus Buy*. Recuperado en Julio 2016 de: <http://www.oracle.com/us/corporate/analystreports/esg-getting-real-bigdata-2228170.pdf>
- [Tableau, 2016] f 2003-2016 Tableau Software. Todos los derechos reservados. Recuperado en Octubre de: <http://www.tableau.com/es-es/about/leadership#aselipsky>
- [TableauPDF, 2016] Kamkolkar, N. & Fields, E. & Rueter, M.(2016). *Tableau para la empresa: descripción general de TI*. tableau software.
- [Pentaho, 2016] bligoo. (2016). Pentaho Open BI. Recuperado en Noviembre de: <http://cognus2.bligoo.cl/content/view/1020771/Pentaho-Open-BI.html>
- [OracleBI,2016] GoLive. (2016). Oracle Business Intelligence. Recuperado en Noviembre de: <http://onegolive.com/es/portfolio-golive-soluciones-consultoriaerp-crm-bi/oracle-bi>

- [Gartner, 2014] Gartner Magic Quadrant. Recuperado en Noviembre de 2016 de: http://www.gartner.com/technology/research/methodologies/research_mq.jsp.
- [Indicador,2014] ALTAG. (2014). 85% de las empresas no analiza correctamente los recursos. Recuperado en Marzo de 2017 de: <http://www.altag.net/wp-content/uploads/2014/03/indicadores1.jpg>
- [Planeaux, 2007] Planeaux, D. & Alvin, D. (2007). *Oracle Business Intelligence Standard Edition One*. California: Oracle Parkway.