UNIVERSIDAD CENFOTEC

Maestría en Tecnologías de Bases de Datos

"Proyecto de Investigación Aplicada 2"

Herramientas de Inteligencia de negocios aplicadas al análisis de la información educativa del sistema PIAD, en Costa Rica.

Elaborado por:

Ing. Marisol Núñez Vásquez

Ing. Ronny Flores Caravaca

Diciembre, 2013

Índice General

Índice de Figuras

Índice de Tablas

1. Introducción

El Proyecto de Informatización para el Alto Desempeño (PIAD), nace de la necesidad de generar indicadores de calidad, para la educación pública en Costa Rica. Al no contar con la información en forma digital, se inició con la idea de desarrollar un sistema que permitiera a los docentes y administrativos, de los centros educativos, realizar sus labores de manera eficiente y eficaz; y a la vez proveer una asesoría que permita la toma de decisiones mejor fundamentadas en materia de educación.

El presente proyecto propone el desarrollo de un Almacén de Datos, que tome la información del sistema PIAD, y permita el traslape de variables; proveyendo otra que permita analizar patrones de comportamiento, y la toma de decisiones inteligentes.

Como elemento innovador, el presente proyecto utiliza la metodología CRISP en el desarrollo de un Almacén de Datos, la cual fue creada para proyectos de minería de datos, con una guía de usuario muy completa compuesta por fases, tareas y actividades. Se desarrollará un Almacén de Datos que comparte, al igual que un proyecto de minería de datos, la necesidad de tomar datos capturados de los sistemas transaccionales, y trabajar con los mismos para generar los resultados deseados. Se valorará en qué medida cada una de las fases, tareas y actividades propuestas en la metodología CRISP, pueden ser aplicadas, y se aportarán las modificaciones realizadas sobre la misma, para cumplir los objetivos de este proyecto.

1.1 Marco Institucional

La Asociación para la Innovación Social (ASIS), constituida, en el año 2002, por miembros de la comunidad educativa de San Rafael Abajo de Desamparados, es la organización que lidera el proyecto PIAD; una nueva forma de cooperación público – privada. Según muestra el organigrama, visualizado en la figura 1, se plasmó el 26 de marzo del 2008, al firmar un acuerdo de cooperación de largo plazo para la implementación del PIAD con la Asociación Empresarial para el Desarrollo (AED), con la Asociación Nacional de Educadores (ANDE) que ha colaborado ampliamente en la difusión de las herramientas entre sus cuarenta y cinco mil agremiados, y con el Ministerio de Educación Pública (MEP), que avala el proyecto mediante directrices y un proceso de colaboración e institucionalización, desde el año 2006.

El PIAD es una herramienta informática que consta principalmente de dos componentes: el registro electrónico y el sistema de información. El registro electrónico es una hoja de cálculo que integra veintiséis funcionalidades relacionadas con notas y asistencia. Por otra parte, el sistema de información es una herramienta que integra los módulos de expediente del estudiante, proceso de matrícula, expediente del funcionario, el Plan Operativo Anual (POA), inventario y equipo. Estos instrumentos han sido adaptados a diferentes entornos educativos: pre-escolar, primaria multidocente, primaria unidocente, secundaria académica y secundaria técnica profesional.

El PIAD nació con la finalidad de producir indicadores para medir la calidad de la educación; sin embargo, en sus inicios no se contaba con la recolección efectiva de los datos y los mismos se encontraban dispersos, lo que impedía el poder generar las estadísticas requeridas.

Actualmente, se trabaja en integrar los diferentes tipos de centros educativos en un ambiente centralizado en la nube; como primer paso para contar con los datos requeridos, y lograr generar los indicadores de calidad.

Este proyecto partirá de la información contenida en los repositorios de datos de los centros educativos, para crear el Almacén de Datos (una copia estructurada de los datos de las transacciones, para la consulta y el análisis) que permitirá combinar la información para analizar los resultados.

1.1.1 Misión

Establecer las condiciones, para generar un alto impacto en la realidad educativa, especialmente en la disminución de la deserción y los efectos de la pobreza en la educación; esto mediante el desarrollo e implementación de herramientas informáticas que automatizan procesos administrativos, y mejoran la toma de decisiones para los docentes, directores(as) y jerarcas del Ministerio de Educación Pública.

1.1.2 Visión

Eliminar las consecuencias de la pobreza en todos los centros educativos del país, y llevar el PIAD a otros países en forma exitosa.

1.1.3 Valores Organizacionales

Los principales valores organizacionales son: la confianza, horizontalidad, trabajo por procesos y resultados, flexibilidad, capacidad de prevención y resolución de conflictos, trabajo en equipo, comunicación asertiva, priorización y delimitación de responsabilidades claras, eficiencia, rendición de cuentas y solidaridad.

1.1.4 Principios Estratégicos

Alianzas público-privadas de largo plazo, trabajar con formadores y tomadores de decisiones estratégicos, orientación permanente hacia los usuarios primarios (trabajadores que digitan, y usan la información), tecnología para la equidad en las áreas de salud y educación pública.

1.1.5 Organigrama

La estructura organizacional de ASIS, mostrada en la figura 1, se distingue por estar apoyada en un convenio con instituciones públicas y privadas, que permiten llevar a cabo la misión de ésta. Además, de un equipo humano que se rige por valores organizacionales, que lo distinguen por romper las normas de las contrataciones laborales típicas, permitiendo el teletrabajo parcial o total.

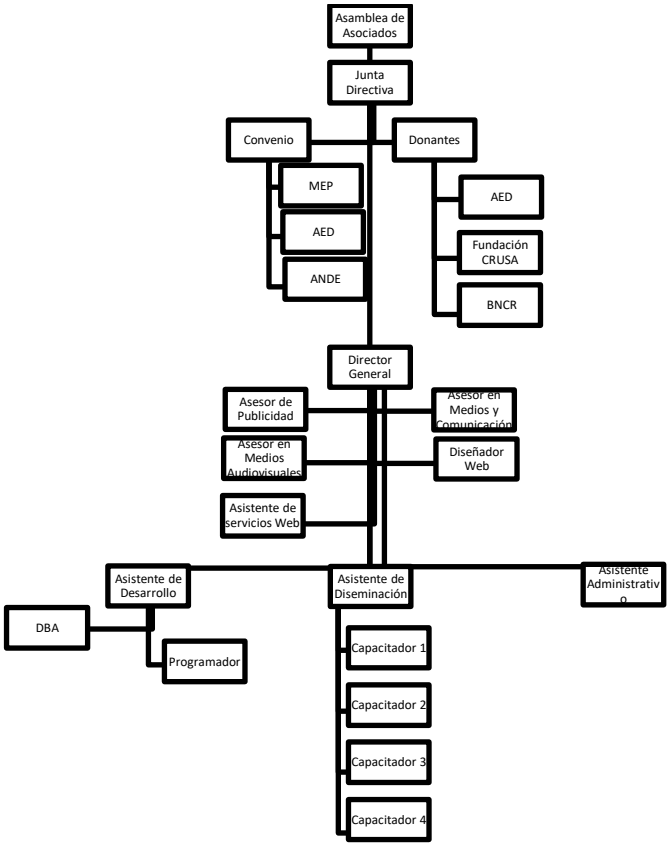


Figura 1: Organigrama de ASIS

1.2 Descripción del Problema

1.2.1 Problema General:

¿Cómo proveer de una herramienta de inteligencia de negocios, que permita el análisis de los datos de rendimiento académico y permanencia de los estudiantes, durante los períodos lectivos del 2008 al 2012, de los centros educativos públicos de Costa Rica que utilizan el sistema PIAD?

1.2.2 Subproblemas:

a. ¿Cómo adquirir el conocimiento del ambiente laboral de la educación pública, y qué datos aportan valor al análisis del rendimiento y la permanencia de los estudiantes en los centros de enseñanza?

b. ¿Cómo proponer una solución de estructuras de datos, capaz de brindar información de los centros educativos públicos de forma unificada, con el propósito de analizar y generar estadísticas?

c. ¿Cómo implementar una solución de inteligencia de negocios, utilizando herramientas de bases de datos, que permitan estructurar la información para el análisis, y generación de estadísticas de la educación pública?

1.3 Justificación

Se plantea esta investigación debido a la necesidad que presenta el país, de contar con información unificada de los centros educativos públicos. Además, de los requerimientos que se tienen actualmente para la generación de estadísticas de rendimiento académico, *repitencia*, ausentismo, deserción, adecuaciones curriculares y situación socioeconómica del estudiante.

La facilidad de contar con un Almacén de Datos de la educación pública, será una herramienta de toma de decisiones muy importante para las entidades de educación, permitiéndoles evaluar los datos históricos, y realizar proyecciones a futuro del comportamiento del rendimiento y la permanencia del estudiantado, en los centros educativos públicos del país, así como otros indicadores de calidad.

Una vez implementada la herramienta, y alimentada con la información histórica del rendimiento académico y la permanencia de los estudiantes en su lugar de estudio, se puede permitir, a futuro, realizar estudios de patrones de comportamiento, por medio de proyectos de minería de datos.

1.4 Objetivo General

Desarrollar una herramienta de inteligencia de negocios, que permita el análisis de datos y apoye el proceso de toma de decisiones, basado en el historial de rendimiento académico y permanencia de los estudiantes, durante los períodos lectivos del 2008 al 2013, de los centros educativos públicos de Costa Rica, que utilizan el sistema PIAD; aplicando la metodología CRISP, para la generación de estadísticas nacionales.

1.5 Objetivos Específicos

1.5.1 Documentar los requerimientos y los datos que aportan valor para el diseño del Almacén de Datos, basado en la información del Reporte de Variables Múltiples del PIAD.

1.5.2 Diseñar el Almacén de Datos, mediante el modelado de las estructuras requeridas, para contener los datos resultados del análisis.

1.5.3 Implementar un Almacén de Datos e instrumentos de análisis de datos y apoyo al proceso de toma de decisiones, aplicando la metodología CRISP, y utilizando herramientas de bases de datos que se ajusten al diseño propuesto.

1.5.4 Realizar una evaluación de la aplicabilidad de las actividades propuestas en la Metodología CRISP para el desarrollo de un Almacén de Datos.

2 Marco Teórico

El presente capítulo tiene el propósito de proveer a la investigación, conceptos y proposiciones que permitan abordar el problema, dentro de un ámbito donde éste cobre sentido; incorporando conocimientos previos sobre el rendimiento académico, el ausentismo y la deserción de los estudiantes del sistema educativo, la teoría sobre Almacén de Datos y los antecedentes que fundamentan la metodología CRISP.

2.1 Educación pública: rendimiento académico y permanencia del estudiante.

La contribución de la educación al desarrollo nacional, fue fuertemente reconocida en los países del tercer mundo después de los años cincuenta. Durante todos estos años, los gobiernos y los grupos organizados subrayaron la importancia de desarrollar la cantidad de la enseñanza (crear más aulas, responder a la demanda y reducir el déficit de atención escolar, construir más escuelas, contratar más maestros). Sin embargo, la calidad de la enseñanza y el aprendizaje, en la mayoría de las escuelas, se considera baja; lo que preocupa cada día más a los gobiernos y a la opinión pública. La educación, en sus aspectos cuantitativos, se ha visto particularmente afectada a causa de la explosión demográfica, la crisis económica, social y política. Situaciones que han dado una nueva dimensión al problema de la calidad de la educación, frente a la demanda y a los déficits.

2.1.1 Aspectos educativos.

Lorenzo Guadamuz (2008) expone que para asegurar su continuidad histórica, toda sociedad utiliza la educación como instrumento de transmisión de los valores e ideales propios de su cultura. Los conocimientos y las acciones reconocidas como útiles socialmente, se recrean de unas generaciones a otras, a través, de los procesos de enseñanza y aprendizaje (formales e informales) con el fin de fortalecer en el alumno su capacidad de respuesta a las necesidades de su propio desarrollo, y a las de su colectividad. La organización social de la educación y sus contenidos son el reflejo consecuente de la sociedad misma. Si una sociedad es desigual e injusta, la educación o cualquier otro aspecto del sector social lo será también (p. 27).

Según Lorenzo Guadamuz (2008), la educación no está, y no puede estar separada de los problemas de la sociedad costarricense. Se ve afectada por los valores (a los contravalores) generados por una sociedad de consumo, por la violencia cotidiana (difundida por los medios de comunicación colectivos), y por fenómenos de desintegración familiar y social. Hoy más que nunca, la educación es afligida por el impacto de la crisis económica, y los problemas sociales y políticos; influenciando la formación de niños y jóvenes: la interacción familiar, la comunidad, la escuela y los docentes, y sus relaciones internas en la institución (p. 27).

2.1.2 Principales problemas de la educación costarricense.

De acuerdo a Lorenzo Guadamuz (2008), los problemas principales de la educación en Costa Rica pueden resumirse de la siguiente manera:

a. El sistema formal educativo no llega a satisfacer las demandas en educación preescolar, de la población residente en las zonas alejadas y población atípica.

b. La eficiencia interna del sistema educativo básico es poco importante, y refleja un elevado porcentaje de deserción, aumentando el número de analfabetos funcionales.

c. El contenido de la enseñanza general básica, a pesar de la renovación de los programas, sigue siendo inadecuado para las necesidades de las zonas rurales y marginales urbanas, así como lo es la preparación de los alumnos para integrarse al circuito productivo.

d. El sistema de educación, después del nivel básico, exceptuando a las escuelas de enseñanza diversificada de carácter técnico, se centra excesivamente en la realización de estudios superiores y la obtención de un diploma. Presenta falta de flexibilidad para ofrecer "salidas laterales".

e. Se acentúa el desequilibrio entre el sistema educativo y el mercado del trabajo. Por un lado, se produce una oferta de trabajo poco calificada; y por otro, la llegada de profesionales y personas con una preparación básica, lo que conlleva al subempleo de los recursos humanos.

f. La duplicación de las profesiones, en las instituciones de educación superior, implica elevar los costos de éstas.

g. La distribución equitativa de las posibilidades educativas, frenada por la estratificación social, aún está muy marcada en nuestra sociedad. El origen socio-económico del alumno condiciona su acceso en el sistema, su permanencia e incluso la salida con certificación, para los menos favorecidos.

h. Existe una desigualdad en la distribución regional de las posibilidades de educación, en todos los niveles del sistema, lo que contribuye a acentuar el desequilibrio de las regiones en el desarrollo nacional.

i. El sistema educativo no favorece la participación de la comunidad en el proceso educativo.

j. La educación se limita a transmitir los modelos científicos y tecnológicos provenientes de los países desarrollados, y a la crítica de estos modelos favorece el espíritu de investigación, que podría crear una capacidad propia generalizada para todos los alumnos, de desarrollo científico y tecnológico.

k. La transmisión de cultura, que realiza el sistema educativo, es limitada, y muchas veces, anulado por la "contra cultura" de las prácticas sociales de diferentes grupos; dadas a conocer ampliamente por los medios de comunicación.

l. En el plano administrativo, no hay una estructura clara, ni una especificación de las funciones y responsabilidades; aun cuando la centralización es muy importante, y hace de las Direcciones Regionales de Educación, entes útiles. Situación agravada, por el hecho de que estas direcciones no tienen su propio presupuesto para comprar el material de oficina, el combustible, o el pago de viáticos, para realizar las visitas a la supervisión de los centros educativos.

2.1.3 Rendimiento Académico

Schiefelbein y Simmons realizan un trabajo, publicado en los años 70, que consta de una revisión de 26 estudios sobre el rendimiento escolar, y que consideraba 20 países sub-desarrollados de Asia, África y América Latina. (Schiefelbein, 1981).

Este estudio combina las características siguientes:

a. Utilización de una encuesta a escala nacional, para identificar los posibles determinantes del rendimiento escolar.

b. Empleo del análisis de los sistemas y de las variables múltiples, para establecer la importancia relativa de los indicadores.

c. Aplicación del concepto económico de "función de producción", en el estudio de los efectos de los indicadores.

A partir del análisis de los datos realizado, según las hipótesis precedentes, los autores concluyeron que los indicadores con mayor efecto en el rendimiento de los estudiantes eran: las características del centro educativo, del educador y los rasgos personales del alumno.

Las investigaciones brindan resultados evidentes: 16 dimensiones sobre las categorías centro educativo, educadores y 15 sobre los alumnos. Solamente fueron retenidas las evidencias estadísticamente significativas, y se llegó a un conjunto de diez variables. Su descripción es la siguiente:

a. Características de la escuela:

- Dimensión de la clase (número de alumnos en cada clase): Una clase de 25 a 50 alumnos no afecta mucho el rendimiento escolar; sin embargo, si el número de estudiantes aumenta, habrían reacciones negativas por parte de los educadores. Además, la utilización de métodos y nuevas técnicas demanda la disminución de la dimensión de la clase.

- Libros de texto: En general, no existen relaciones bien definidas entre éstos y el rendimiento escolar. Las variaciones podrían surgir ya sea, cuando los textos son mal utilizados, o no son necesarios, porque existen mejores fuentes de información; e inclusive cuando no contribuyen ni al aprendizaje, ni a la adquisición de altos niveles de habilidades (por ejemplo: el juicio evaluativo).

- Tareas: Esta variable se encuentra en 6 u 8 estudios. En sí, son importantes para las políticas de educación, ya que no representan ningún gasto y permiten un buen resultado. Hay que tomar en cuenta el modo en que, el tiempo de dichos trabajos es utilizado.

b. Características de los educadores:

- Certificaciones, formación y estudios de perfeccionamiento. Estas variables tienen un peso diferente en el rendimiento académico, para los diferentes países.

- Experiencia en la enseñanza: esta característica es un determinante significativo del rendimiento académico, en 7 de los 19 estudios realizados. En general, entre más grande sea la experiencia, menos se utilizan métodos y técnicas nuevas. Sin embargo, los educadores con mucha experiencia se ven a menudo, asignados a clases de alto rendimiento.
- Formación: No existe una relación evidente entre una larga formación del educador y un mejor rendimiento escolar. Si tal fuera el caso, habría importantes implicaciones en las demandas de enseñanza.

c. Rasgos de los estudiantes:

- Estatus socio económico de la familia (SES): Es la variable más estudiada. En 10 de 13 casos, hubo una fuerte incidencia del rendimiento académico.
- Malnutrición y salud: Son factores predicativos importantes del rendimiento académico, ya que tienen una considerable correlación con el estatus socio-económico, y por lo tanto, deberían considerarse en otra dimensión.
- Repetición: Para determinar su peso, hay que tomar en cuenta como las habilidades del estudiante, fueron juzgadas por el maestro.
- Desarrollo precoz: La importancia del desarrollo bien logrado es reconocido en el primer periodo, pero los estudios sobre el rendimiento escolar solo consideran los años posteriores al maternal.

2.1.4 Ausentismo

a. Definición

Se puede definir el absentismo como la situación de inasistencia a clases, por parte del estudiante en la etapa obligatoria de manera permanente y prolongada. En determinadas ocasiones, esto tiene lugar por causas ajenas al propio alumno, como pueden ser la aparición de una enfermedad o un traslado familiar; en otras, se debe a una "elección" por parte del estudiante, que no encuentra en el centro educativo la respuesta a sus problemas e intereses, que acumula retrasos en relación con su grupo de edad, o que desea buscar otra opción al margen del sistema educativo (Uruñuela 2005).

Sin embargo, la anterior consideración del absentismo estudiantil resulta completamente insuficiente; primero, porque se refiere únicamente a la última etapa del problema, cuando se ha producido la ruptura definitiva con el sistema educativo; segundo, porque olvida el proceso que conlleva el fenómeno del absentismo, que se desarrolla poco a poco, y se concreta en diversas manifestaciones; por último, porque hace casi imposible una respuesta adecuada a este problema, ya que, cuando se presenta está completamente desarrollado, y apenas ha dejado margen para actuar.

El absentismo debe ser conceptualizado como una respuesta de rechazo, hacia el sistema educativo, por parte del alumno, y que adopta varias manifestaciones y grados: en algunos casos, son ausencias a clase que deben ser contempladas como una travesura infantil, más que como un problema; en otros, son ausencias mucho más preocupantes, e incluyen: el absentismo pasivo del estudiante, desligado de las explicaciones y actividades normales de las clases; a las faltas de puntualidad; la inasistencia a clase, de forma especial, que tienen lugar en ambos extremos horarios; las ausencias intermitentes a unas clases o asignaturas; el abandono esporádico del centro, a determinadas horas; hasta llegar al abandono definitivo.

La consecuencia inmediata de todas estas conductas, en función de su mayor o menor grado de desarrollo, suele ser la alteración del ritmo de aprendizaje que puede llevar a la repetición de curso, y al fracaso educativo; pero, a su vez, hay que tomar en cuenta que en otras ocasiones el fenómeno puede ser el contrario, y sea precisamente el fracaso de la enseñanza, el que termine expulsando al alumno de la escuela, dando lugar a un serio problema.

Una de las preocupaciones, a la hora de describir el fenómeno del absentismo, viene dado por la ausencia de estadísticas fiables respecto a este, ni a nivel nacional, ni a nivel local o autonómico; nada hay previsto en los planes estadísticos sobre el absentismo escolar, sabiendo que una buena información sobre este tema, llevaría a una comprensión más adecuada de este asunto.

b. Factores y causas para la explicación del absentismo

En numerosas ocasiones, se ha tratado de explicar el absentismo estudiantil recurriendo a factores psicológicos del estudiante (baja autoestima, ausencia de habilidades sociales, entre otras.), o a causas sociológicas, tales como: la pertenencia a un determinado grupo, a una minoría étnica, u otras características sociales. Si bien es cierto, que el

absentismo afecta a sectores de población, que sufren situaciones de marginación o degradación social y/o económica; no puede concluirse que éste sea el único factor explicativo de este problema. Lo mismo habría que decir del entorno familiar que, por varios motivos, se puntualiza en no prestar la atención necesaria, tanto al cumplimiento de la educación obligatoria, como a la evolución del aprendizaje.

Siguiendo a Rué y colaboradores, hay que considerar que el absentismo escolar, es la respuesta de un determinado alumno a una situación de aprendizaje, ofrecida por ofrecida el centro educativo; y que, en dicha determinación, puede presentarse; en primer lugar, una serie de factores predeterminantes, entre los que cabe mencionar a los de tipo sociológico; en segundo lugar, habría que considerar otra serie de agentes detonantes, tales como: el sentimiento de pérdida de la autoestima, el desencuentro entre los intereses del alumno y los del centro educativo, y un cierto grado de complicidad por parte de la familia o del grupo de iguales; complicidad que viene a reforzar la resolución absentista, por parte del estudiante.

Tres son los elementos que deben revisarse desde la propia institución, para analizar su repercusión en las conductas absentistas: el curriculum, la organización del centro y el tipo de relaciones que se establecen en el mismo. No cabe duda, que un plan de estudios marcado por el academicismo y la abstracción, muy alejado de los intereses vitales de muchos de los alumnos, y sobrecargado de contenidos y materias; incide directamente en las actitudes de los estudiantes hacia el lugar de aprendizaje. Lo mismo se puede decir de la rígida organización de los Institutos: su inflexibilidad horaria, reglamentos "de régimen interior"; la falta de relaciones humanas, que pueden darse entre los estudiantes y sus profesores (muchas veces por falta de tiempos y de espacios que las hagan posibles); o de las relaciones basadas en el modelo "dominio-sumisión" que caracterizan la disciplina en los centros. Todo esto, por no hablar de la violencia entre iguales, que poco a poco, va saliendo a la luz en todas las instituciones educativas.

2.1.5 Deserción

En nuestro país existe la obligatoriedad de la Educación Pública, hasta el noveno año de secundaria; como lo menciona el Undécimo Informe del Estado de La Nación "Costa Rica ha apostado por la educación como elemento clave para promover el desarrollo humano. El decidido impulso que se dio a la educación primaria a finales del siglo XIX fue uno de los factores que marcó la diferencia en materia de alfabetización que exhibió

el país frente al resto de Centroamérica, pues ya para inicios del siglo XX había superado su rezago".(2006, p. 275)

Los esfuerzos han sido grandes en lo que se refiere a primaria, lo cual lleva a "una población alfabetizada que, según el Censo del 2000, registraba en esa fecha un promedio de 7,6 años de escolaridad". (Undécimo Informe Estado de la Nación, 2006, p. 275).

Definición de Exclusión, Expulsión, Repulsión y Abandono Escolar.

El Doctor Leonardo Garnier, actual Ministro de Educación de Costa Rica (2010-2013), sostiene, en su artículo llamado "Ética, estética y ciudadanía: una educación para saber vivir y saber convivir", que el abandono de la educación formal, se puede explicar con tres fenómenos distintos: exclusión, expulsión y repulsión.

a. La exclusión

Obedece a factores económicos, de modo que, aquellos jóvenes, pertenecientes a sectores en desventaja socioeconómica, se ven obligados a abandonar las aulas, para realizar el trabajo formal e informal como su actividad diaria, no permitiendo a la educación formal ser una actividad prioritaria, contrario a lo que corresponde a su edad, y a lo que establece nuestra constitución y los diferentes tratados internacionales de protección a la infancia y la adolescencia. Considera el señor Ministro, que este problema trasciende las funciones de su Ministerio, y compete a otras instituciones que están obligadas a resolver y evitar la exclusión.

Complementando esta definición, se tiene lo expuesto por Duschatzky y Corea (2002), cuando mencionan la exclusión como un estado en que se encuentra un individuo. La persona excluida es un mero producto: "un dato, un resultado de la imposibilidad de integración, el expulsado es resultado de una operación social, una producción, tiene un carácter móvil" (p. 18).

b. La expulsión

Obedece a factores que incumben fundamentalmente al sistema educativo, y su deber de sostener a jóvenes que presentan dificultades, para adaptarse y responder a las exigencias y requerimientos académicos y formales, de los procesos de educación.

Para Duschatzky y Corea (2002), la expulsión social es la relación entre ese estado de exclusión, y lo que hizo esto posible. Al no considerar la exclusión como estado o determinación, sino como una condición, se le da un carácter productivo. La expulsión, entendida como serie de operaciones, ofrece la oportunidad de observar el funcionamiento y la producción de la situación, de la persona a quien se expulsa. La expulsión social alude una forma de constitución de lo social, y ocasiona alguien que ya ha desaparecido. El expulsado o expulsada pierde visibilidad, palabra, nombre, pierde presencia en la vida pública; entra al mundo de la indiferencia, ya que anda en una sociedad que pareciera no esperar nada de él o ella.

c. La repulsión

Vargas (2006), menciona que el problema de la repulsión escolar, en nuestro país, se centra en la saturación que se hace a los niños y niñas en el sistema educativo, donde, tanto docentes como estudiantes, son bombardeados con lecciones académicas, que desmotivan y son aburridas.

La repulsión, también, es un problema que atañe directamente a la institución educativa, ya que se encuentran jóvenes para quienes el centro educativo deja de ser atractivo, su oferta académica y su promesa de preparación no es creíble, un sector de este grupo resulta más vulnerable, con menor tolerancia y abandona las aulas.

Otra situación que puede favorecer este rechazo al sistema educativo es, como lo plantea Rojas (2000):

"la educación es un elemento homogeneizador que olvida la existencia de diferencias sociales y económicas en los cuales los individuos están inmersos. La población desertora experimenta la exclusión social, porque la educación es transmisora de los valores socialmente aceptados, además los cambios que sufren los adolescentes pueden ir desde los propios de su crecimiento y adaptación al medio, hasta aquellos que pueden modificar negativamente el destino de su vida, como son la delincuencia, las drogas, la prostitución, entre otros" (p. 2)

d. El abandono escolar

Espíndola (2002, p. 9) cita: “rara vez es un evento inesperado; se presenta más bien como una cadena de hechos que van elevando el riesgo de deserción a medida que se avanza en edad y se experimentan crecientes dificultades de rendimiento y de adaptación, especialmente cuando se transita del ciclo primario al secundario”.

Richard (2006, p.6) agrega: “quien hace abandono del sistema escolar se ve enfrentado a poner en juego estilos de sobrevivencia o bien habilidades laborales iniciales para asegurar su propia manutención”.

Por su parte, Vargas (2006) menciona, que es urgente revisar la política educativa, considerando la deserción escolar como una bomba de tiempo, donde las pandillas, la delincuencia y la drogadicción serán quienes la alimenten, si no se realizan prontas mejoras.

e. Salida anticipada

Se utiliza este término, para hacer referencia a lo que comúnmente se ha englobado bajo el nombre de deserción escolar. De manera, que salida anticipada, para efectos de esta aproximación teórica, se define como: la salida de la adolescente y el adolescente del sistema educativo formal costarricense, sin haber concluido sus estudios de secundaria; esto como producto de un proceso de alejamiento paulatino y progresivo, influenciado por una serie de factores, tanto internos como externos.

Factores internos y externos que intervienen en la Salida anticipada de estudiantes del sistema educativo.

Factores internos

Intereses

Según Holland (1987), el término interés se puede definir como: atracción por una actividad, que puede ser espontánea y de carácter intelectual; en donde no se debe ver como un deseo, pues tienen objetivos diferentes, ni tampoco debe de confundirse con el agrado. Es aquella preferencia que se manifiesta mediante las actividades, que la persona emprende durante su tiempo libre, el contenido de sus lecturas favoritas, el éxito y el gusto por ciertas disciplinas escolares y en el trabajo.

Este se muestra en diferentes situaciones, y es un factor de motivación que ayuda a los individuos a sentirse bien (Casullo y Cayssials, 1994).

Motivación

A lo largo de la vida, las personas viven situaciones, las cuales le hacen enfrentarse con sus propias capacidades y fortalezas; pero no en todo momento, reaccionan con igual nivel de estimulación para cumplir tareas, o alcanzar algún propósito que ellas mismas se hayan propuesto; de ahí la importancia de la motivación. Lakobson (citado por Abarca, 1995) señala, que hablamos de motivación, como el conjunto de aspectos que organizan y orientan la actividad del ser humano, y que se manifiestan en los impulsos a realizarla.

Así, podría apuntarse que la motivación es la fuerza que orienta las acciones humanas, e impulsa a realizar las cosas que se hacen. Es un proceso dinámico, está en constante cambio, es un aspecto que acompaña al sujeto, durante todo su desarrollo; por lo anterior, es que el nivel de motivación que se tenga dependerá, en gran medida, de las experiencias que cada persona haya obtenido a lo largo de su vida (Abarca, 1995).

Factores externos:

Familia

Según Jiménez (1997), la familia es un sistema social, y un núcleo dentro de la sociedad; conformado por un grupo de individuos en interacción constante, que viven bajo un mismo techo, y en donde cada miembro tiene un papel definido, y cada situación que se presente afecta, positiva o negativamente, a sus miembros.

Para Nassif (citado por Jiménez, 1997), la familia es un agente educador de primer orden, en donde se le presentan a las nuevas generaciones los primeros elementos educativos: costumbres, valores, tradiciones, moral, entre otras.

Jiménez (1997) menciona como, ante los constantes cambios de la sociedad, la familia se ha expuesto a una serie de transformaciones, que han puesto en peligro la estabilidad y el bienestar de sus miembros. De manera que, cada día es más frecuente la desintegración familiar, donde los menores de edad son los principales afectados, debido a la desestabilización a que son sometidos, en diversas áreas de sus vidas.

Económicos

La situación socioeconómica es uno de los principales factores influyentes en la vida del individuo (Irola, 2002). Tanto que, en muchos casos, define el que pueda alcanzar o no, los objetivos educativos; de modo que, para muchos padres y madres de familia, el mantener a sus hijos e hijas en los centros de estudio se torna un privilegio, del que no pueden gozar.

Nilo (citado por Jiménez, 1997) considera que una causa de la salida anticipada del centro educativo, son los factores asociados a la situación socioeconómica, como la posición social y status de los padres; quienes muchas veces debido a sus dificultades, retiran a sus hijos e hijas de sus centros de enseñanza, negándoles ese derecho a la educación.

El factor económico se manifiesta, no sólo en diferencias de ingreso, sino también en las desiguales posibilidades de acceder a los sistemas públicos o privados de educación, y a circuitos educacionales de muy distinta calidad. Además, a esto se suma, la disparidad en cuanto al clima educacional del hogar (educación de los padres y madres); factor tanto o más influyente en los logros educativos de los niños y adolescentes, que los propios recursos económicos familiares. Con ello, tiende a reproducirse la desigualdad de oportunidades, de una generación a la siguiente, permitiendo que factores de carácter adscriptivo, primen decisivamente en los logros durante las distintas etapas de la vida escolar. (CEPAL, 2002, p.95).

Grupo de pares:

Según Krauskopf (2002), el grupo de iguales adquiere gran importancia en la vida del y la adolescente, en su proceso de elaboración de su identidad. De manera que, dentro de este grupo pueda probar sus nacientes recursos, experimentar alternativas de roles, y alcanzar posiciones que satisfagan sus necesidades de autonomía.

Además, agrega que las amistades adquieren en la vida del y la adolescente, ese papel de enriquecimiento en sus relaciones interpersonales, el reconocimiento de sus destrezas, la puesta en marcha de sus valores, y así, ampliación de sus opciones y participación en la sociedad.

De manera que, si bien estas relaciones, pueden ser un factor muy importante en la motivación del joven y la joven hacia el estudio, también se pueden convertir en causa de su fracaso y retiro del centro educativo.

Institución Educativa:

Como parte de los factores externos que pueden inducir el abandono escolar, se encuentra lo que puede llamar mala praxis educativa, que contempla los errores en los cuales ha incurrido el sistema educativo, y que promueven la deserción de estudiantes.

Kaplún (2004), menciona que existe un tipo de educador o educadora, cuya visión promueve que "quien aprende, lo hace", dejando a su suerte a personas que presentan alguna dificultad para aprender. Bajo esta misma dirección, Gutiérrez (2007) considera que el sistema educativo puede inducir el abandono escolar en la medida que: "se reduce a obligaciones e instrucciones, que los y las jóvenes viven en forma pasiva, con aburrimiento, y en la que sus intereses, preocupaciones y problemas no tienen cabida. No existe por parte del personal docente, la capacidad o motivación para incentivar a una participación crítica, creadora, comprometida" (p.6).

Para Kaplún (2004), es necesario realizar una reinvención de esta comunicación, donde, como un elemento fundamental, exista un o una docente con capacidad de escucha, que se interese por conocer la realidad y comprender a sus estudiantes. Así mismo, este autor, expone que: "los educadores necesitan construir alternativas pedagógicas capaces de dialogar con las culturas juveniles y la pedagogía critica debe ser capaz de ofrecer respuestas en este sentido" (2004, p.3).

2.2 Almacén de Datos (o "Data Warehouse")

2.2.1 Almacén de Datos según Inmon (2001)

Un Almacén de Datos, según Inmon (2001), es una arquitectura estructurada de datos, la cual apoya el manejo de estos con las siguientes características:

a. Orientación por tema: Implica que la información se clasifica entorno a los aspectos o temas que son de interés para la organización. Así organizado, los conceptos de negocio relacionados están asociados entre sí, dentro del Almacén de Datos. Estas clasificaciones podrían ser clientes, productos, estudiantes, proveedores, transacciones, órdenes de compra, entre otras. Su organización no es por funciones o

aplicaciones, lo cual permite a los datos cambiar sin afectar a la misma. Lo antes mencionado es de gran importancia, ya que con el pasar del tiempo un Almacén de Datos manipulará gran cantidad de información.

b. Integración: Se refiere a la transformación de los datos, desde la visión de las aplicaciones hacia una percepción compuesta dentro del Almacén de Datos. Esta transformación, se da a partir de una exploración de las estructuras clave de datos: su definición, su disposición, sus relaciones y convenciones. Esta unificación exige, necesariamente, alcanzar el consenso en la definición de los conceptos, de manera que, en el ámbito de empresa, los términos de negocio tengan un único significado y éste sea conocido por todos.
c. Incorpora tiempo a los datos: Esta característica tiene como objetivo, que los cambios producidos en los sistemas operacionales, que afectan a la información contenida en el Almacén de Datos, puedan ser registrados, de manera que sea posible reconstruir su evolución en todo momento. Cada registro, del Almacén de Datos, tiene el tiempo indisolublemente "grabado" en él. Una vez creado el registro de manera fiable, no debe ser modificado (aspecto no volátil). Por lo tanto, un Almacén de Datos se compone de un conjunto de fotos de datos en un momento dado. La práctica recomendada es tener de 5 a 10 años de histórico. Una forma común, es asociar a los datos la “fecha desde” y la “fecha hasta”, para representar un periodo.
d. No volátil (no se pierde información): Las actualizaciones ocurren por excepciones. Cuando se tiene la necesidad de cambiar los datos en un Almacén de Datos, es recomendable realizarlo por medio de una foto de datos en un tiempo dado, lo cual agregará una foto más de datos que se almacena con el cambio realizado.
e. Se compone de agregaciones y datos detallados: Los datos detallados reflejan el nivel atómico de las transacciones. Esto incluye el uso de productos, actividad de cuentas, movimientos de inventario, ventas, etc. Se pueden dar dos tipos de agregaciones: la pública y la de perfil. La agregación por perfil podría ser a nivel de cliente o estudiante. Por el contrario, la agregación pública es de un nivel más alto, por ejemplo departamento, cantón, semanal, entre otros.
f. Enfoque de arriba hacia abajo: Trata de representar todos los datos importantes de una empresa, en el Almacén de Datos. Es un enfoque organizacional, lo cual lo hace difícil de implementar, y riesgoso de fracasar; y esto dificulta la venta de la idea a la gerencia de una empresa.

El Almacén de Datos existe, con el fin de apoyar la toma de decisiones del negocio, y su proceso de planeamiento estratégico. Los datos fluyen de las diferentes fuentes datos, para convertirse en información útil para la empresa; utilizando herramientas de análisis de datos, según se muestra la figura 2.

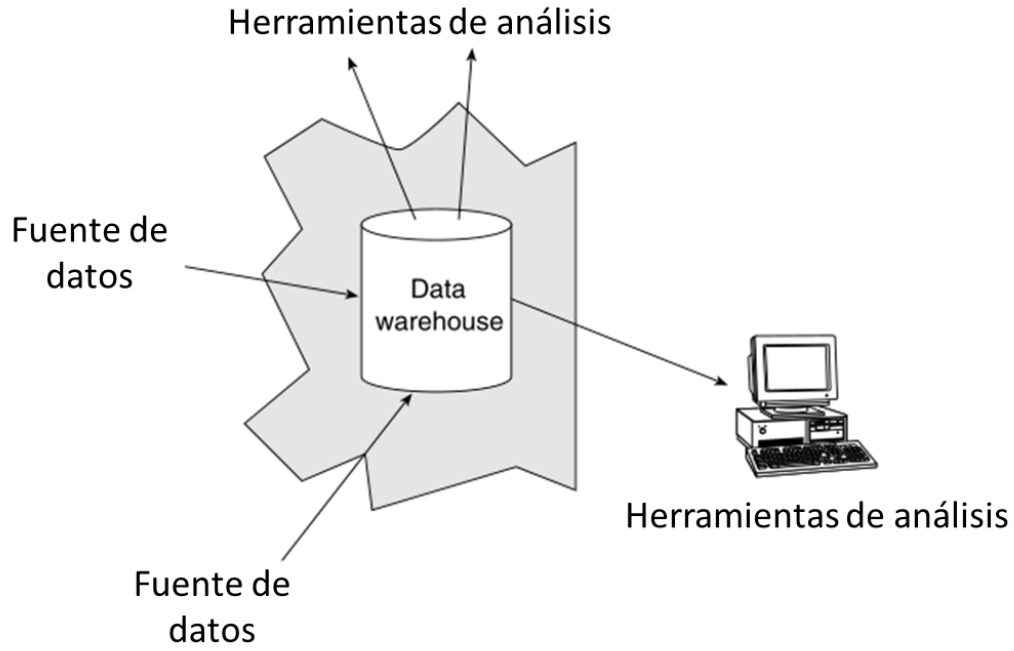


Figura 2: Elementos básicos de un Almacén de Datos.

Mercados de datos

Una vez que se tiene el Almacén de Datos conformado, Inmon (2002) menciona los mercados de datos. Estos se pueden entender, como una colección de datos adaptados, para las necesidades del proceso de toma de decisiones de una organización. Es un subconjunto de datos que ha sido personalizado, para satisfacer necesidades específicas de un área del negocio. Además, este se puede convertir en un recurso compartido para varias áreas. Usualmente, existen varios mercados de datos en un Almacén de Datos. Hay mercados de datos comunes tales como: mercadeo, finanzas, contabilidad, ingeniería entre otros. Estos mercados son un subconjunto del Almacén de Datos, que contienen una gran cantidad de agregaciones, información histórica limitada, y son alimentados únicamente por el mismo.

Los mercados de datos son útiles por:

a. Control: de la información, y de quién accede a ella.

b. Costo: El costo de proceso y almacenamiento es menor, por ser un subconjunto de datos
c. Personalización: Los datos son personalizados, para satisfacer necesidades específicas del negocio.
d. Características comunes: subconjunto de datos, des normalizados, agregados, información histórica limitada, personalizada para reportes y análisis, procesos dedicados para construirlos.

Inmon (2002), plantea tres tipos de mercados de datos; dos de estos se basan en el procesamiento analítico en línea (OLAP). A continuación mencionamos las categorías planteadas de tipos de mercados de datos:

a. Un subconjunto del Almacén de Datos: es decir un resumen simple de este.
b. Procesamiento analítico multidimensional en línea (MOLAP, por sus siglas en inglés Multidimensional Online Analytical Processing): En un mercado multidimensional, los datos son cargados al mercado de datos en forma estructurada, y se crean dimensiones de acuerdo a estos. Estas dimensiones podrían ser de tiempo, localización, entre otras. Una vez creadas las dimensiones, los datos pueden agregarse de acuerdo a estas. Posteriormente, creadas las agregaciones, se pueden navegar los datos, desde las agregaciones hasta el detalle. El objetivo de este mercado es la alta flexibilidad y desempeño.

c. Procesamiento Analítico en Línea Relacional (ROLAP por sus siglas en inglés Relational Online Analytical Processing): Este mercado es de tipo relacional, es decir, que basa la integridad de los datos en el motor de base de datos. Un esquema, el cual tiene afinidad a este tipo de mercados de datos racionales, es el esquema estrella (mostrado en la figura 3). Este organiza los datos de tal forma que, es fácil de navegar y visualizar. Estos esquemas se componen de tablas de hechos (datos de cantidades de unidades o monetarias), y son de gran volumen. Por lo contrario, las tablas de dimensión son pequeñas, y se racionan con las tablas de hecho por medio de alguna clave, como por ejemplo: tiempo o lugar.

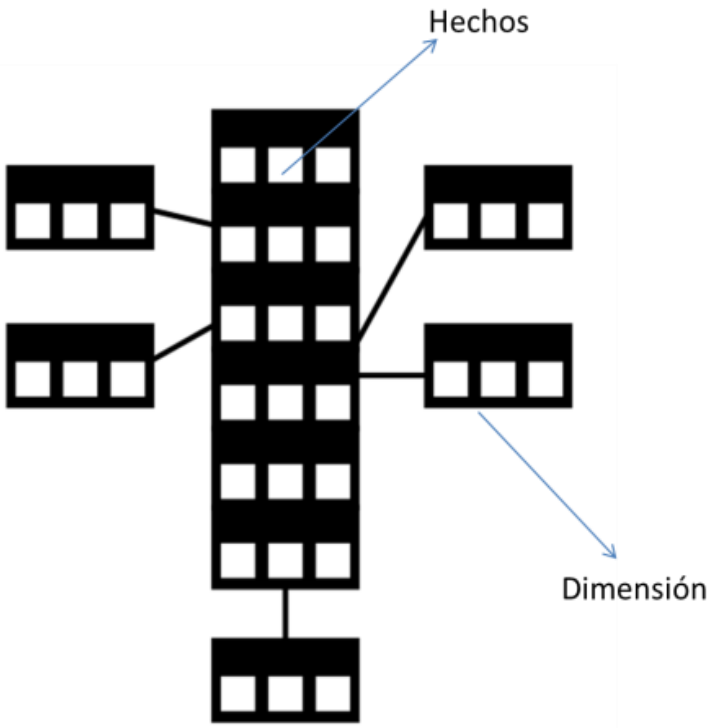


Figura 3: Esquema estrella.

2.2.2 Almacén de Datos según Kimball

De acuerdo a Kimball (2008), los elementos básicos de un Almacén de Datos son: los sistemas operacionales, el área de "stage" de datos, el área de presentación de datos y las herramientas de acceso a datos. Según se muestra en la figura 4.

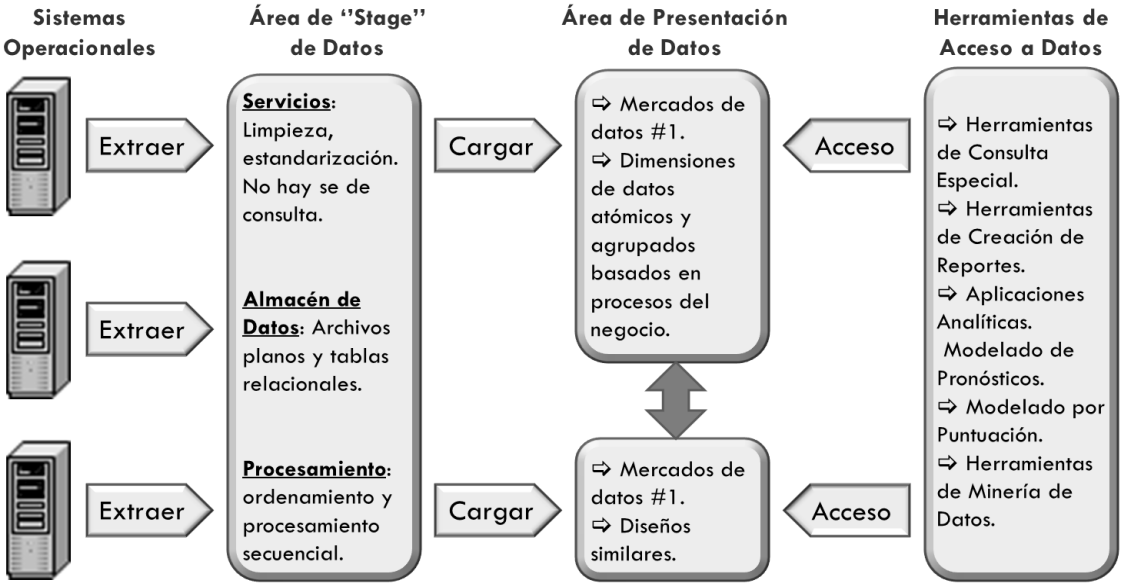


Figura 4: Elementos básicos de un Almacén de Datos según Kimball.

a. Los sistemas operacionales: Se refieren a las principales fuentes de información. Estos son los sistemas transaccionales del negocio; su prioridad es tener alta disponibilidad, minimizando tiempos fuera de servicio. Las consultas a estos sistemas son limitadas, y se asume que contienen pocos datos históricos. Es recomendable, no utilizar sistemas operacionales como fuentes de análisis de datos, ya que este tipo de uso puede arriesgar la disponibilidad del sistema.
b. Área de datos de "Stage": Es un área donde se "arreglan" los datos; es una "limpieza" de datos. Algunas de las tareas que se llevan a cabo en esta área son: combinar, des duplicar y archivar. En esta se separan las fuentes de datos del área de presentación, la cual por tener la responsabilidad de trabajar los datos, tiende a ser grande en términos de requerimientos de hardware. Esta área respeta las reglas del negocio de los datos, pero no utiliza, necesariamente, las bases de datos relacionales; es un área temporal de almacenamiento. El punto clave es no dar ningún servicio de consulta o presentación de datos.
c. Área de Presentación: Es donde los datos son almacenados, para que los usuarios hagan sus consultas, análisis y soporte de toma de decisiones. En esta se localizan los esquemas de estrella o herramientas OLAP (procesamiento analítico en línea).

De acuerdo a lo antes mencionado, Kimball define un Almacén de Datos como un conglomerado de todos los mercados de datos dentro de una empresa; siendo una copia de los datos transaccionales estructurados de una forma especial para el análisis. Su enfoque es de abajo hacia arriba (contrario a Inmon), ya que recomienda construir un mercado de datos a la vez, como primer elemento del sistema de análisis, y posteriormente, ir agregando otros que compartan las dimensiones ya definidas o incluyan otras nuevas. En este sistema, los procesos de carga extraen la información de los sistemas operacionales, y los trabajan en el área "Stage", realizando luego, el llenado de cada uno de los mercados de datos, de una forma individual.

Kimball propone una metodología general, para la construcción de un Almacén de Datos, la cuales es:

a. Selección del proceso de negocio.
b. Definición de la granularidad de la información.
c. Elección de las dimensiones de análisis.
d. Identificación de los hechos o métricas.

2.2.3 ETL (Extract-Transform-Load)

Según Kinball (2004), el acrónimo ETL significa en ingles Extract-Transform-Load o Extraer-Transformar-Cargado (ETC). Este proceso es de suma importancia para un Almacén de Datos. Un diseño apropiado del ETC, cumple con las siguientes características:

a. Extrae datos de los sistemas operacionales.
b. Asegura la calidad de los datos.
c. Asegura consistencias de los datos.
d. Entrega los datos listos, para que el área de presentación pueda utilizarlos.

El ETC hace funcionar el Almacén de Datos, o puede provocar que este falle. Aunque construir el proceso de ETC es una actividad no visible para los usuarios, este consume un setenta por ciento de los recursos de implementación y mantenimiento de un Almacén de Datos típico. El ETC agrega valor a los datos, es decir que es más que extraer los datos de los sistemas operativos de la empresa. Principalmente este se encarga de:

a. Eliminar errores de los datos, corregir datos faltantes.
b. Documenta medidas de confianza de los datos.
c. Captura el flujo transaccional de los datos.
d. Ajusta los datos de múltiples fuentes para ser utilizados conjuntamente.
e. Estructura los datos, para ser utilizados en las herramientas de presentación para usuarios.

Kimball recomienda llevar dos procesos en paralelo a la hora de desarrollar un ETC:
Proceso 01: Planeamiento y diseño: En el proceso de planeamiento, mostrado en la figura 5, se trata de satisfacer todas necesidades del negocio, establecer perfiles de datos, validar requerimientos, establecer requerimientos de seguridad y definir la integridad de los datos, y su latencia como aspectos clave. En cuanto a la arquitectura, es importante definir la forma en que se cargará los datos, las herramientas se van a utilizar, el manejo de la calidad de los datos y los procedimientos de recuperación después de fallas. En cuanto a la implementación, hay que evaluar el hardware, el software, la documentación y las prácticas de codificación. Finalmente, se revisan y definen los procesos de desarrollo, las pruebas, la implementación, el mantenimiento y la afinación de procesos.

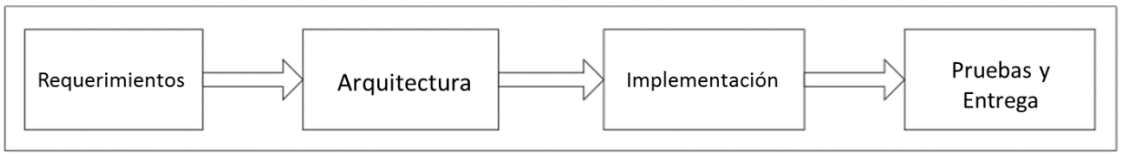


Figura 5: Etapas del proceso de planeamiento.

Proceso 02: Flujo de datos: El proceso de flujo de datos, según se puede observar en la figura 6, se compone de la extracción, limpieza, validación y entrega de los datos. Extraer incluye la lectura de los modelos de fuentes de datos, acceso a estos y calendarización de proceso. En cuanto a la limpieza, esta comprende: velar por el cumplimiento de que los datos tengan las propiedades correctas, estructura correcta, valores correctos; que se cumplan las reglas del negocio complejas; construir la meta data correspondiente, entre otras cosas. En cuanto a la validación cabe mencionar: validación de métricas, validación de desempeño e internacionalización de datos. Y por último, la entrega abarca: cargado y generación de dimensiones, generación de llaves, cargado de tablas de hechos y actualización de agregaciones.

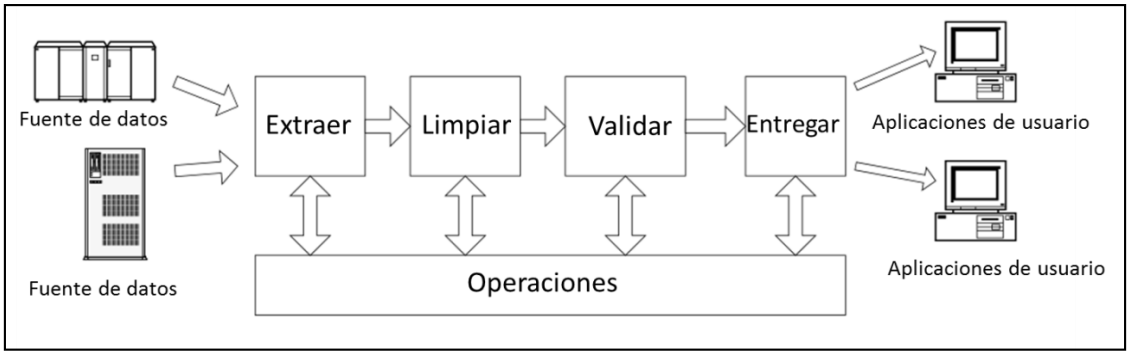


Figura 6: Etapas del proceso de flujo de datos.

El proceso de flujo de datos es apoyado por una serie de operaciones, las cuales se deben considerar a detalle, estas operaciones incluyen:

a. Calendarización.
b. Ejecución de procesos.
c. Manejo de excepciones.
d. Recuperación y reinicio de procesos.
e. Revisiones de calidad.
f. Entrega.
g. Soporte.

2.2.4 Calidad de la información

Para tener un Almacén de Datos exitoso, se debe tener calidad de la información, para lo cual se deben tomar en cuenta los siguientes conceptos:

a. Dato: es algún número u otra información, representada de forma tal que una computadora pueda procesarlo, esto desde el punto de vista de tecnología de la información. Pero desde el punto de vista de negocio, un dato es una representación de hechos. Representan cosas en el mundo real. El dato es un símbolo, o una representación de un hecho de alguna cosa. Por ejemplo, el nombre de una persona, donde el dato es el nombre, y la cosa en sí es la persona.
b. Información: Se puede decir que los datos son la materia prima de la información. La información son los datos en contexto, o datos que se pueden utilizar. Es el significado de los datos, que ayuda a entender los hechos ocurridos. Calidad en la información requiere una clara definición de los datos, valores correctos, y una presentación de datos entendibles, es decir, en el formato correcto. La información es representada por datos correctos, más una correcta definición, y una buena presentación de estos.
c. Conocimiento: El conocimiento es información en contexto. Implica entender el significado de la información. Es el valor agregado de la información por la gente. Con la tecnología existente, las empresas pueden capturar conocimiento de manera electrónicamente, en grandes cantidades. El conocimiento tiene valor cuando las personas tienen el poder de decisión de actuar, con base en el conocimiento.
d. Sabiduría: La sabiduría es conocimiento aplicado. Tiene como objetivo, el generar aprendizaje inteligente de la organización; incrementando la capacidad de crear en el futuro, por medio del aprendizaje compartido.

Lo antes mencionado, ayuda a definir la calidad de la información, y puede ser analizada desde dos puntos de vista:

a. La calidad de la información, desde un punto de vista de precisión, sería el grado de precisión en que la información representa el mundo real, ya que todos los datos son una representación de éste.
b. La calidad de la información como el grado de utilidad, y valor que tienen los datos para el apoyo al funcionamiento de una empresa y el logro de sus metas. Los datos, en un Almacén de Datos, no tienen valor real, es un valor potencial. Adquieren valor cuando estos son utilizados para hacer algo útil y de beneficio para la organización.

La calidad de la información, radica en la habilidad de satisfacer a quienes la usan. Evita malas decisiones, re trabajo, riesgos, cálculos erróneos, pérdida de clientes y oportunidades, daños y mala comunicación. Para tener calidad de la información, se requiere identificar los usuarios de los datos, ya que éstos apoyarán el trabajo. En conclusión, la calidad de información afecta la eficiencia y efectividad de los procesos del negocio, y la satisfacción de los clientes.

2.2.5 Modelos multidimensionales

Con el fin de tener una idea clara de los modelos multidimensionales, se debe comprender el modelo de dos dimensiones. Cualquier reporte de registro de horas por empleado, costos por departamento, o quejas por tienda, es un ejemplo de un modelo multidimensional, de al menos dos dimensiones.

En la tabla 1, se muestra un ejemplo de una dimensión de tiempo (meses), relacionada a diferentes variables o hechos.

Tabla 1: Ejemplo simple de un modelo de una dimensión (tiempo) y cinco variables.

Mes	Ventas	Costos Directos	Costos Indirectos	Total de Costos	Margen
1	¢790	¢480	¢110	¢590	¢200
2	¢850	¢520	¢130	¢650	¢200
3	¢900	¢530	¢150	¢680	¢220
4	¢910	¢600	¢140	¢740	¢170
5	¢880	¢490	¢90	¢580	¢300
6	¢900	¢500	¢120	¢620	¢280
7	¢790	¢620	¢150	¢770	¢20
8	¢820	¢300	¢110	¢410	¢410
9	¢840	¢540	¢100	¢640	¢200
10	¢810	¢570	¢80	¢650	¢160
11	¢840	¢600	¢75	¢675	¢165
12	¢820	¢520	¢95	¢615	¢205

En la anterior tabla, se observan cinco variables descritas por la dimensión tiempo. Estas variables o hechos, por sí solas, no dicen mucho, pero en conjunto con una dimensión adquieren valor y significado. Las dimensiones ayudan a describir cómo los datos están organizados. Si se agrega una dimensión más, por ejemplo producto, se puede visualizar como lo muestra la figura 7.

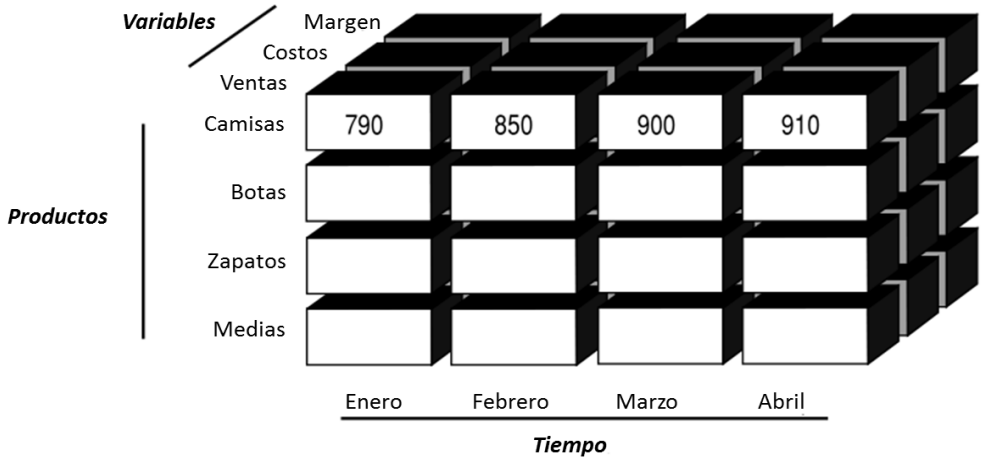
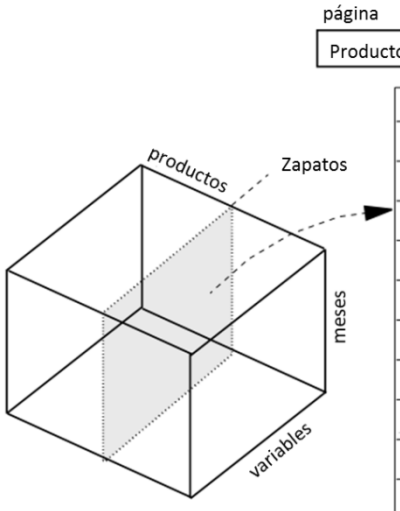


Figura 7: Ejemplo de modelo multidimensional de dos dimensiones y tres variables.

El ejemplo, antes mencionado, describe cómo se pueden representar dos dimensiones en el papel o en una pantalla de computadora; pero más de tres dimensiones es difícil de ejemplificar o verlo en un papel, ya que este está limitado a dos dimensiones. La figura 7, es una metáfora visual de cómo representar datos con varias dimensiones, en este caso tres. Esta limitación incentiva, a grandes pensadores, a considerar otras formas de representación de datos, por medio de la computadora.

Una forma de mostrar varios conjuntos de datos de un modelo multidimensional en una pantalla, es por medio de visualizar porciones de éste, según se muestra en la figura 8.



	Ventas	C. Directos	C. Indirectos	C. Totales	Margen
Enero	520	320	110	430	90
Febrero	400	250	130	380	20
Marzo	430	300	120	420	10
Abril	490	320	150	470	20
Mayo	520	310	180	490	30
Junio	390	230	150	380	10
Julio	470	290	160	450	20
Agosto	500	360	150	510	-10
Setiembre	450	290	140	430	20
Octubre	480	290	140	430	50
Noviembre	510	310	150	460	50
Diciembre	550	330	160	490	60

Figura 8: Representación gráfica de datos multidimensionales.

Sin importar el volumen de la cantidad de datos, en el modelo multidimensional de tres dimensiones, siempre es posible visualizar los datos por medio de porciones. En la

metáfora visual, mostrada en la figura 8, se observa la porción de datos para el producto zapato, para todos los meses de un año, y para las variables de costos, ventas y márgenes.

2.2.6 Representación de tres o más dimensiones

Para representar modelos multidimensionales de tres o más dimensiones, existen las "Estructuras de Tipo Multidimensional (ETMs)", conocidas en inglés como "Multidimensional Type Structures (MTSs)". Esta metáfora visual ayuda a describir los modelos de más de cuatro dimensiones.

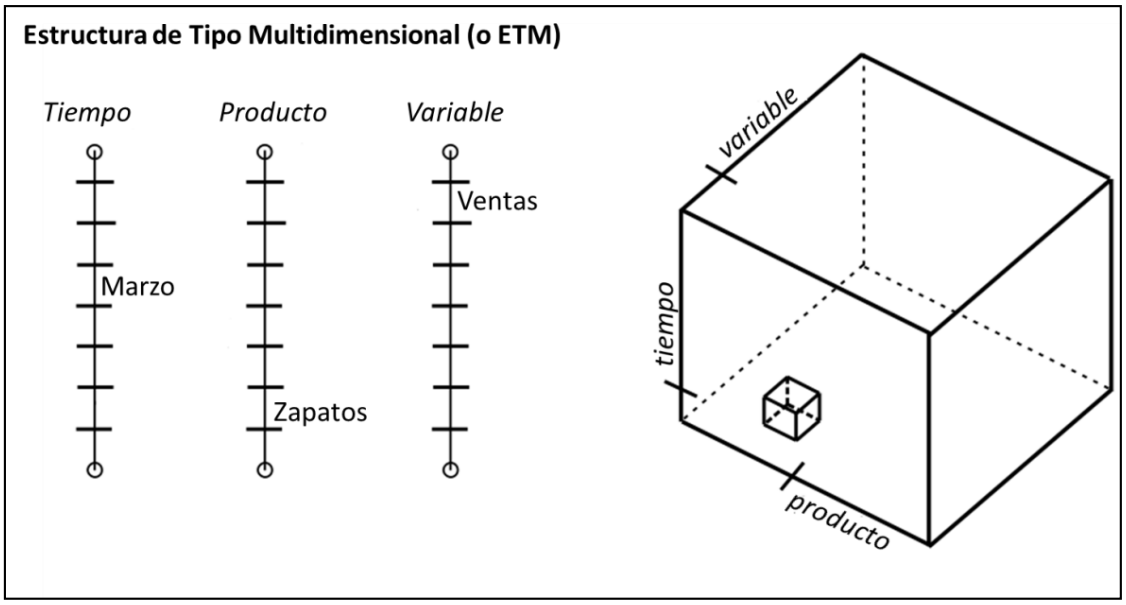


Figura 9: Multidimensional Type Structures (MTSs) o Estructuras de Tipo Multidimensional (ETMs).

En la figura 9, se puede observar dos metáforas visuales, representando el mismo modelo multidimensional de datos, en ambos casos, se está representando las ventas para zapatos en el mes de marzo.

Para representar tres o más dimensiones con ETMs, es tan fácil como agregar otra dimensión, según se muestra en la figura 10.

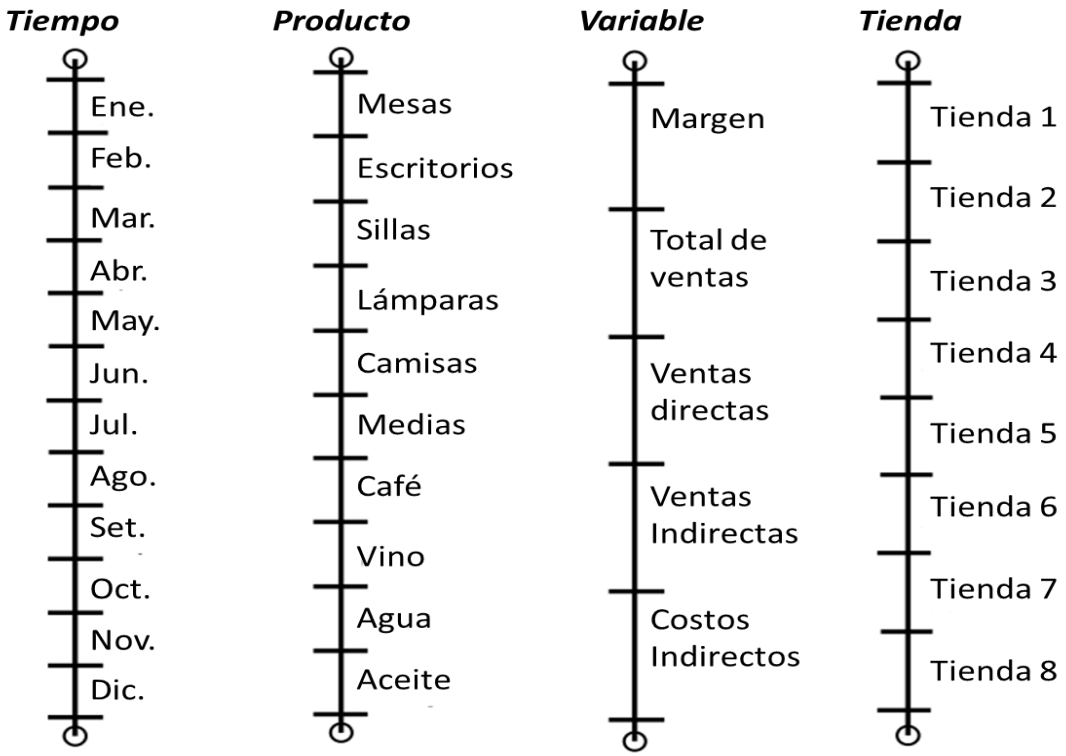


Figura 10: Ejemplo de ETMs con tres dimensiones.

Con los ETMs se pueden representar todas las combinaciones necesarias, entre las variables y las dimensiones; y por este proceso de revisión, validar si realmente la solución que se propone, responde las necesidades de los clientes.

Los ETM's son una forma sencilla de representar modelos y validarlos, ante los grupos de toma de decisión, para determinar que se están comprendiendo los requerimientos de información, antes de continuar con el desarrollo de la herramienta para análisis y apoyo a la toma de decisiones.

2.3 Metodología CRISP para la implementación de Almacén de Datos

2.3.1 Historia

El CRISP-DM es una metodología iniciada, a finales de 1996, por un grupo de empresas europeas: DaimlerChrysler (Thomas Reinartz y Rüdiger Wirth), SPSS-Inglaterra (Julian Clinton, Thomas Khabaza y Colin Shearer), NCR-Dinamarca (Peter Chapman y Randy Kerber), OHRA-Holanda y AG-Alemania; "veteranos" del joven e inmaduro mercado de minería de datos, que deseaban proponer un modelo de proceso

estándar, sin propietarios y libremente disponible; que sirviera de guía a todos los profesionales en el tema.

En 1997, con el apoyo financiero de la comisión Europea, se formó el consorcio conocido con el acrónimo CRISP-DM, por sus siglas en inglés (CRoss-Industry Standard Process for Data Mining); con el objetivo de lograr una herramienta para la industria lo más generalizada posible, a fin de adaptarse a la mayor diversidad de industrias.

En 1999, se presenta el primer modelo de proceso estándar, para atender la comunidad de minería de datos, CRISP-DM 1.0

El CRISP-DM tiene éxito, porque está profundamente basado en la experiencia práctica, la experiencia del mundo real, de cómo la gente conduce proyectos de minería de datos.

2.3.2 Objetivos de la metodología CRISP-DM

a. Surge de la necesidad de aprender nuevas técnicas para aplicar, y comprender de mejor manera a la Minería de Datos y sus resultados, basándose en un proceso jerárquico.
b. Se identifica por perseguir el cumplimiento de objetivos, desde el punto de vista empresarial, dando preferencia a la comprensión del negocio.
c. Desarrollar proyectos de minería de datos mediante un proceso estandarizado.
d. Minimizar los costos, que implica un proyecto de minería de datos en las empresas.

2.3.3 El Modelo CRISP-DM

La metodología se describe en términos de un proceso jerárquico, consistente en un grupo de tareas descritas en cuatro niveles de abstracción (de general a específico). Según se muestra en la figura 11:

a. Fase
b. Tareas Generales
c. Tareas Específicas
d. Instancias del proceso

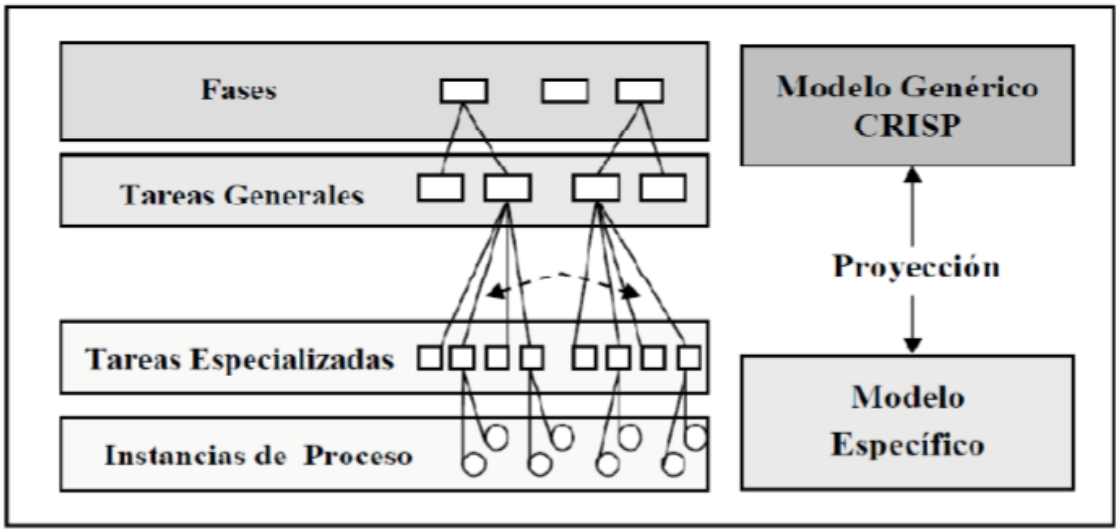


Figura 11: Modelo CRISP-DM.

2.3.4 Ciclo de vida de un proyecto

La metodología CRISP-DM, provee una representación completa del ciclo de vida de un proyecto de Minería de Datos, según se muestra en la figura 12, que se divide en: seis fases, sus tareas y relaciones entre ellas, representadas en la figura 13.

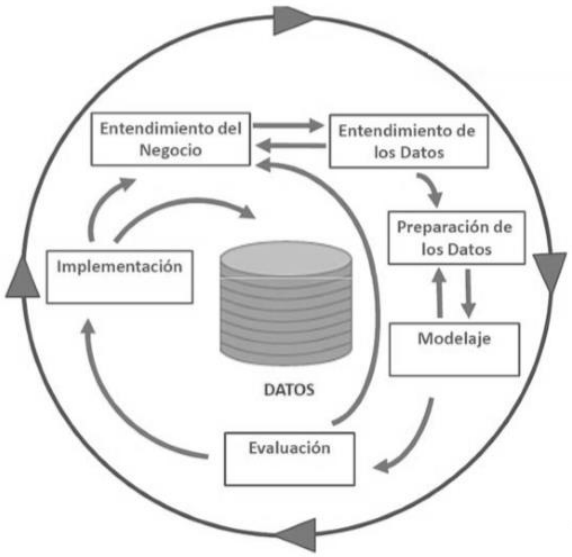


Figura 12: Ciclo de vida de un proyecto CRISP-DM.

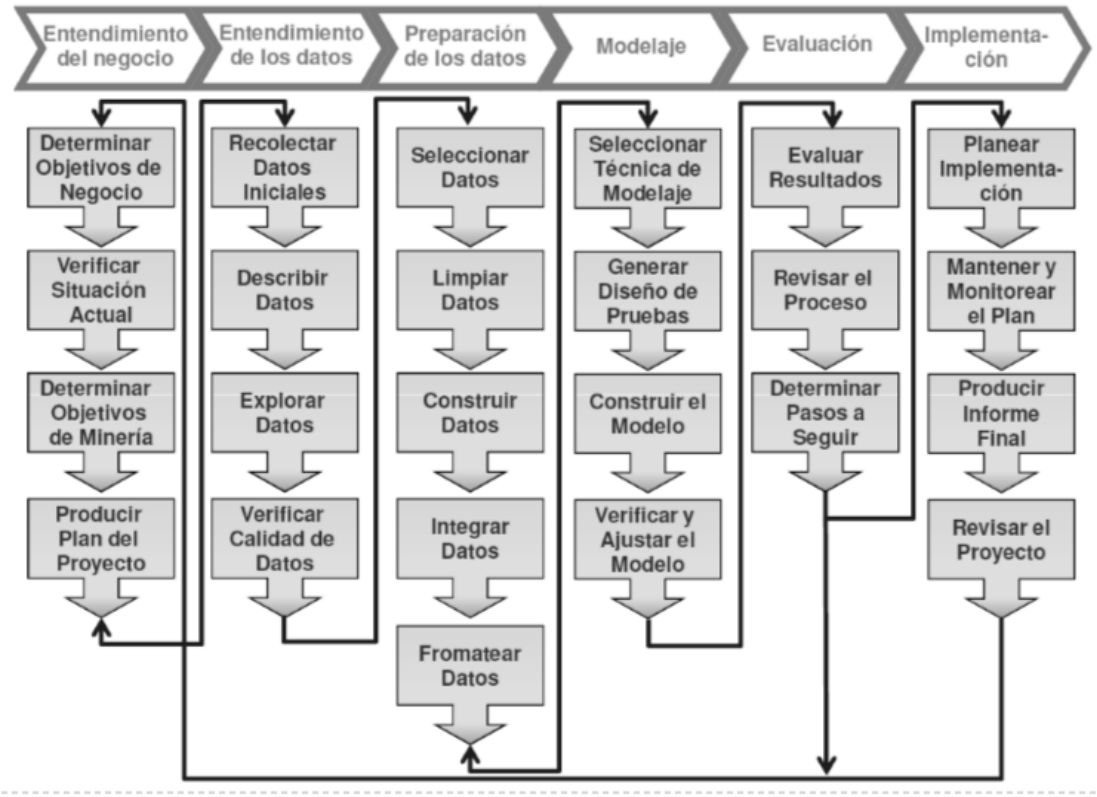


Figura 13: Proceso CRISP-DM.

Fase 1. Entendimiento del negocio

Es la fase más importante de cualquier proyecto de minería de datos. Los diseñadores deben tener un completo conocimiento del negocio, para el que están encontrando una solución. Entender los objetivos y requerimientos del proyecto desde una perspectiva de negocio.

Subfases:

- Establecimiento de los objetivos de negocio (contexto inicial, objetivos y criterios de éxito).
- Evaluación de la situación (inventario de recursos, requerimientos, suposiciones y restricciones, riesgos y contingencias, terminología, costes y beneficios)
- Establecimiento de los objetivos de minería de datos (objetivos de minería de datos, y criterios de éxito)
- Generación del plan del proyecto (plan del proyecto y evaluación inicial de herramientas y técnicas).

Fase 2. Entendimiento de los datos

El analista se familiariza con los datos. Es en esta fase, dónde se descubren ideas iniciales en los datos, o se detectan subconjuntos para formar hipótesis sobre información escondida. (Según los objetivos del negocio de la fase anterior).

Fase 3. Preparación de los datos

Cubre todas las actividades, para construir el conjunto final de los datos que serán utilizados en las herramientas de modelado.

Aquí se incluye: la integración, selección, limpieza y transformación de los datos.

Fase 4. Modelaje

Varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos.

Fase 5. Evaluación

Se evalúa el modelo, y se revisa la construcción.

Se comprueba que se cumplan los objetivos del negocio.

Al final de esta fase, el líder del proyecto debe decidir exactamente cómo utilizar los resultados del proceso.

Fase 6. Implementación o despliegue

Se trata de explotar la potencialidad de los modelos, e integrarlos en los procesos de toma de decisión de la organización, y difundir informes sobre el conocimiento extraído.

El conocimiento obtenido debe ser organizado, y presentado de manera que el usuario lo pueda utilizar.

Revisar el proyecto a fin de identificar fracasos y éxitos, además de potenciales áreas de mejora para el uso en futuros proyectos.

3 Marco Metodológico

3.1 Tipo y Diseño de la Investigación

3.1.1 Enfoque de la Investigación

Esta investigación usa el enfoque cualitativo. Para conceptualizar este enfoque, se ha consultado a varios referentes de la investigación científica. Es así como Hernández, Fernández y Baptista (2006), al referirse a la investigación cualitativa, expresan que ésta se fundamenta en un proceso inductivo (explorar y describir, y luego generar perspectivas teóricas). La recolección de los datos, según los autores, consiste en obtener las perspectivas y puntos de vista de los participantes (emociones, experiencias, significados y otros aspectos subjetivos). Es donde el investigador se introduce en las experiencias individuales de los participantes, y construye el conocimiento, siempre consciente de que es parte del fenómeno estudiado (p. 8).

Hernández et al. (2006) destacan que, en la investigación cualitativa, el marco teórico no prefigura con exactitud el planteamiento del problema. La literatura referenciada es sólo la necesaria, para facilitar que los datos o información emerjan de los participantes, cuya función es la de justificar y documentar la necesidad del proyecto, y su utilidad al finalizar es apoyar o modificar los resultados existentes documentados (p.531).

Una investigación cualitativa es abierta y expansiva, según Hernández et al. (2006), ya que la misma permite un enfoque mayor en los conceptos relevantes, conforme se avanza en el estudio. Inicialmente no están direccionados, se fundamentan en la experiencia o intuición. Está orientada a aprender de experiencias y puntos de vista de los individuos, valores y procesos; generando teorías que se fundamentan en las perspectivas de los participantes.

La investigación planteada, conlleva a que los investigadores involucrados aprendan, y se sumerjan en el rol de negocio de la educación pública, para poder plasmar las necesidades en requerimientos, que permitan la implementación de una solución para inteligencia de negocios que provea de información, a los encargados de generar políticas y propuestas de mejoras, en el ámbito educacional.

El rol de observador participante, en investigaciones cualitativas, es explicado por Valles (2007); aclarando el doble propósito del observador al implicarse en la situación

en estudio y observarla a fondo. Esto demanda un estado de mayor alerta, y prestar atención a los aspectos culturales tácitos. Se caracteriza esta experiencia, como una doble condición de miembro y extraño. Se explotará la introspección natural como instrumento de investigación. Todo esto conducirá, naturalmente, a una exigencia de registrar de manera sistemática las actividades, observaciones e introspecciones (p. 150).

Valles (2007) indica, que las entrevistas en profundidad permiten la obtención de una gran riqueza informativa, en las palabras y enfoques de los entrevistados. Además, proporcionan al investigador la oportunidad de clarificar en una interacción más flexible, que la entrevista estructurada o la encuesta. Provee orientaciones útiles, para acoplar un proyecto a las demandas del cliente, y evitar errores que pueden resultar costosos (p. 196).

En esta investigación se establece un enfoque cualitativo, porque los datos que se recolectarán serán en forma de texto, los cuales se analizarán por medio de categorías de análisis.

3.1.2 Diseño de la Investigación

El diseño de esta investigación es evaluativo. Correa, Puerta y Restrepo (2002) consideran que, la investigación evaluativa se ha convertido en una importante fuente de conocimientos y directrices, en las muchas actividades e instituciones, porque evidencia el grado de eficiencia o deficiencia de los programas, y señala el camino para reformular y valorar el éxito alcanzado con los esfuerzos aplicados (p. 11).

Correa, Puerta y Restrepo (2002) señalan que el método concreto de la evaluación, es la investigación evaluativa, donde las herramientas de la investigación social se ponen al servicio del ideal consistente, en hacer más preciso y objetivo el proceso de juzgar. Establece criterios claros y específicos, para garantizar el éxito del proceso; reúne sistemáticamente información, pruebas y testimonios de una muestra representativa de las audiencias que forman parte del programa u objeto a evaluar; traduce dicha información a expresiones valorativas, y las somete a comparación con los criterios establecidos al inicio; y finalmente saca conclusiones (p. 27).

> Suchman (1967) asevera que: "La investigación evaluativa es un tipo especial de investigación aplicada cuya meta, a diferencia de la investigación básica, no es el descubrimiento del conocimiento. Poniendo principalmente el énfasis en la utilidad, la investigación evaluativa debe proporcionar información para la planificación del programa, su realización y su desarrollo. La investigación evaluativa asume también las particulares características de la investigación aplicada, que permite que las predicciones se conviertan en un resultado de la investigación. Las recomendaciones que se hacen en los informes evaluativos son, por otra parte, ejemplos de predicción". (p.119).

Esta investigación es evaluativa, porque se evaluará la eficiencia y eficacia de la aplicación de la metodología CRISP, en la construcción del Almacén de Datos.

3.2 Población y muestreo:

3.2.1 Población

Arias (2006) define población como: "un conjunto finito o infinito de elementos con características comunes para las cuales serán extensivas las conclusiones de la investigación".

Pérez (ob.cit) lo define de la siguiente manera: "conjunto finito o infinito de elementos que se someten a estudio; pertenecen a la investigación y son la base fundamental para obtener la información".

Basados en las anteriores definiciones, la población, para esta investigación, son los Colegios Académicos Públicos que utilizan el sistema PIAD en Línea, descritos en la tabla 2.

Tabla 2 Población de la investigación

Centros Educativos que usan PIAD en Línea 9/9/2012		
Centro Educativo	**Estudiantes**	**Docentes**
COLEGIO ACADEMICO DE ORIENTACION TECNOLOGICA OMAR SALAZAR OBANDO	424	30
COLEGIO DE GRAVILIAS	800	55
COLEGIO DE TABARCIA	377	33
COLEGIO NOCTURNO GUAYCARA	700	30

LICEO ALEJANDRO QUESADA R	800	38
LICEO ANASTASIO ALFARO	1000	45
LICEO BRAULIO CARRILLO COLINA	1200	80
LICEO DE ASERRI	1600	70
LICEO DE ASERRÍ (Educación Especial)	246	5
LICEO DE CALLE FALLAS	1300	45
LICEO DE COT	840	42
LICEO DE CURRIDABAT	916	50
LICEO DE PURISCAL	1247	68
LICEO DE SABANILLAS	250	18
LICEO DE SAN ANTONIO	1500	50
LICEO DE SAN FRANCISCO	1300	55
LICEO DE SANTA GERTRUDIS	1000	60
LICEO DE SANTO DOMINGO	930	60
LICEO DR VICENTE LACNER	1950	80
LICEO ENRIQUE GUIER SAENZ	516	45
LICEO EXP.BIL DE TURRIALBA	285	28
LICEO EXPERIMENTAL BILINGUE SANTA CRUZ	368	60
LICEO GREGORIO JOSE RAMIREZ	1300	80
LICEO HERNAN ZAMORA ELIZONDO	800	42
LICEO LABORATORIO U.C.R.	487	32
LICEO LAGUNA	214	24
LICEO LOS LAGOS	746	42
LICEO LUIS DOBLES SEGREDA	1750	75
LICEO MARIO VINDAS SALAZAR	1080	60
LICEO MAURO FERNANDEZ ACUÑA	1180	70
LICEO NOCTURNO MARCO TULIO SALAZAR	246	5
LICEO OCCIDENTAL DE CARTAGO	167	17
LICEO PLAYAS DEL COCO	298	23
LICEO SAN JOSÉ RÍO SUCIO	246	5
MARTHA MIRAMBELL(LIC.ATEN)	1700	76
T.V. LA GATA	70	5
T.V. LAS CEIBAS	125	6
T.V. LOS ARBOLITOS	29	5
T.V. UNION DEL TORO	52	5
TV LA CUREÑA	246	5
Totales	**30285**	**1624**

3.2.2 Muestra

Según RENA(2010), una muestra es un conjunto de unidades, una porción del total, que representa la conducta del universo en su conjunto.

La muestra, para esta investigación, está compuesta por diez colegios académicos públicos que utilizan el sistema PIAD en Línea. La selección no es una muestra estadística, sino que se escogió basada en la cantidad de estudiantes y docentes que

componen cada centro, y la completitud de los datos que presenta, para contar con suficientes datos para el análisis. La muestra representa un 25% de los centros educativos, un 40% de la población estudiantil y un 36% de los docentes, como se muestra en la tabla 3.

Tabla 3 Muestra seleccionada para la investigación

Centros Educativos que usan PIAD en Línea 09/09/2012		
Centro Educativo	**Estudiantes**	**Docentes**
LICEO ANASTASIO ALFARO	1000	45
LICEO DE SANTO DOMINGO	930	60
LICEO LUIS DOBLES SEGREDA	1750	75
LICEO DE CALLE FALLAS	1300	45
LICEO ENRIQUE GUIER SAENZ	516	45
LICEO LOS LAGOS	746	42
MARTHA MIRAMBELL(LIC.ATEN)	1700	76
LICEO DE ASERRI	1600	70
LICEO DE SAN ANTONIO	1500	50
LICEO BRAULIO CARRILLO COLINA	1200	80
Total	**12242**	**588**
Total porcentual a la población	**40%**	**36%**

3.3 Métodos de recolección de datos

Según Scheaffer (1987), el método más común de recolección de datos es la entrevista personal o impersonal por algún medio: teléfono, internet u otro.

Para la recolección de datos del presente proyecto, se hará uso de los siguientes métodos:

3.3.1 Entrevista personal

Los datos son frecuentemente obtenidos con entrevistas personales. El procedimiento requiere que el entrevistador realice preguntas preparadas, y registre las respuestas. La gran ventaja de este método de recolección es, que se confronta a las personas con las preguntas directamente, y a la vez se puede notar reacciones físicas, eliminar malos entendidos, y asegurar una comunicación efectiva. Su mayor desventaja es el entrevistador en sí, ya que si este no está debidamente capacitado para realizar la encuesta, conocer y entender las preguntas, puede sesgar los resultados.

3.3.2 Entrevista por teléfono

Cuando se ponen restricciones de acceso a los entrevistadores, se puede hacer uso de la entrevista por teléfono. Sin embargo, existe la dificultad de establecer un espacio que ayude al entrevistado a expresar sus ideas, pues aunque esté escuchando, puede que realice al mismo tiempo otras actividades, lo cual conlleva a no lograr los resultados esperados. Además, al ser estas más cansadas, es aconsejable realizarlas en periodos de tiempo cortos, en comparación con las entrevistas personales.

3.3.3 Entrevista-electrónica/ virtual

Se define como la comunicación entre el sujeto y el investigador, a través de la mediación de la computadora, haciendo uso de recursos tales como: El chat, el correo electrónico y/o foros virtuales, por los cuales se puede propiciar una interacción secuencial, planeada y organizada en un periodo concreto, y con un propósito específico, determinado por las necesidades de información de la investigación. Para ello se recomienda que se diseñe, y se realice desde la visión de la entrevista en profundidad y semiestructurada, con una guía que facilite la identificación de temas centrales, para encontrar las relaciones que los sujetos le dan a las diferentes variables, que afectan el fenómeno en estudio.

En el desarrollo, se deben seguir los mismos protocolos de una entrevista cara a cara: La presentación para establecer conexión con el entrevistado, comunicar el objetivo y las condiciones en las que se va a llevar a cabo, tiempo y forma para la interacción virtual. En el proceso, el investigador debe hacer uso de sus habilidades como entrevistador, para plantear las preguntas de manera clara y precisa; mantener la secuencia y sobre todo lograr el nivel de empatía tan necesario en cualquier entrevista. Esto último, facilitará el diálogo entre ambas partes, para expresar libremente opiniones, creencias, experiencias y vivencias relacionadas con el tema. Por último, dar el cierre y los agradecimientos, así como la conclusión.

En cuanto al protocolo, se recomienda diseñarlo con temas generales y subtemas, mismos que se desarrollan dependiendo del proceso de diálogo entre el entrevistador y el entrevistado. De acuerdo con Tomlinson (1999), el establecer una estructura y jerarquías de los temas y subtemas en la guía, facilita la profundización de los mismos, así como proporciona una excelente herramienta, para que no se pierda el foco durante todo el proceso de la indagación.

El tomar en cuenta los puntos anteriores, permitirá al entrevistador identificar categorías de análisis, así como la construcción de mapas conceptuales sobre la experiencia de los sujetos, aspectos a los que hace referencia Morton y Booth (1997), y los reportes de los estudios de Brew (2001); Lucas (1998); y Drew & et al, (2001).

3.3.4 Observación directa

La observación directa consiste en obtener datos e información, mediante la percepción intencionada y selectiva, ilustrada e interpretativa de un objeto o de un fenómeno determinado, para ser cuantificada y posteriormente, hacer inferencias de la misma.

Un aspecto importante a considerar, es que la información obtenida no sea afectada por las personas que realizan la recolección. Es decir, que siempre tomen las observaciones en condiciones similares, para que los datos sean válidos y comparables. En esta investigación se ha utilizado la técnica de observación participativa, ya que los investigadores llevan a cabo el proceso; involucrándose directamente en los resultados a observar.

Se ha empleado este instrumento, ya que permite que los datos sean tomados en forma directa durante la construcción del Almacén de Datos, utilizando instrumentos y forma de registros diseñados para la recolección de datos, y la adecuación a las condiciones en que se proyecta realizarlas.

3.3.5 Lista de Cotejo

RENA (ob-cit): "Son la relación de aspectos a observar de los que se registran solamente si se presenta o no el aspecto de la conducta, a lo largo de la sesión observada; u objetos en un sitio. También se le conoce como "Check-List", lista de verificación o lista de chequeo".

3.4 Operacionalización de Variables

3.4.1 Variable 1: Inclusión de las variables contenidas en el Reporte de Variables Múltiples, en el desarrollo del Almacén de Datos PIAD.

a. Definición**:** Introducción de las variables, contenidas en el Reporte de Variables Múltiples del PIAD, como dimensiones o hechos del Almacén de Datos PIAD.

b. Naturaleza: Cualitativa

c. Medición: Check List (SI/NO).

Valor	Rango
Inclusión Alta	76 - 100
Inclusión Media	51 - 75
Inclusión Baja	26 - 50
Inclusión Muy Baja	0 - 25

d. Instrumento: Lista de cotejo de las Variables del Reporte de Variables Múltiples del PIAD, incluidas en el Almacén de Datos del PIAD, mostrada en el apéndice 2.

3.4.2 Variable 2: Aplicabilidad de las actividades de CRISP al desarrollo de un Almacén de Datos.

a. Definición: Posibilidad de aplicar las actividades que componen las fases de la metodología CRISP, a la realización de un Almacén de Datos PIAD.

b. Naturaleza: Cuantitativa

c. Medición: Check List (SI/NO)

Valor	Rango
Altamente aplicable	76 – 100
Medianamente aplicable	51 – 75
Bajamente aplicable	26 – 50
Muy poco aplicable	0 – 25

Instrumento: Listas de Cotejo de la Metodología CRISP por las actividades aplicadas en el proyecto, mostrada en el apéndice 3.

Como proyecto de investigación, este trabajo arrojará datos que serán analizados con la finalidad de transmitir el conocimiento obtenido de la experiencia. Es clave organizar los datos, de tal manera que apoyen su análisis, para lo cual se detalla: que vamos a

definir, que vamos a realizar, especificando minuciosamente nuestro proceso de análisis.

Los investigadores serán actores en el proceso, que desea cuantificar las variables que conforman el Reporte de Variables Múltiples del Sistema PIAD, que pueden incluirse en el Almacén de Datos y determinar la cantidad de actividades propuestas en las distintas fases de la metodología CRISP; identificando cuáles son aplicables al desarrollo de un Almacén de Datos. Se presentarán los resultados en un reporte de resultados obtenidos, por medio de tablas y gráficos, los cuales ayuden a demostrar que, partiendo de datos transaccionales contenidos en un reporte, se puede brindar una herramienta de análisis de datos que permita desplegar información de los indicadores de la calidad de la educación costarricense, y que la metodología CRISP) aunque fue creada para proyectos de minería de datos) puede ser aplicada al desarrollo de un Almacén de Datos.

3.5.1 Análisis cualitativo

Según Sandín (2003), el desarrollo del conocimiento supone, entre otros aspectos, descubrir aquello que es significativo en un escenario. Es importante, que las personas desarrollen una comprensión de las maneras en las que se representa el mundo. Para ello, es necesario encontrar modelos de narrativa o interpretación, que permitan apreciar las diversas formas de representación. También, es oportuno fomentar destrezas en el uso y la generalización de la teoría. Dadas sus importantes funciones, una de las tareas básicas en la preparación de investigadores cualitativos, es la práctica en la utilización de la teoría para justificar y dar sentido a lo que se descubre y describe. Para efectos de este proyecto, se pretende determinar cuántas variables, del Reporte de Variables Múltiples, pueden ser incluidas en el Almacén de Datos como dimensiones o en los hechos.

Además, como innovación, se estará haciendo el recuento de las actividades propuestas en cada fase de la metodología CRISP, que puedan ser aplicadas al desarrollo del Almacén de Datos.

3.5.2 Presentación de los datos cualitativos

Según LeCompte (2000), recolectar datos cualitativos, para que estos sean útiles, es una tarea difícil, ya que estos datos hay que convertirlos en resultados. Esta transformación de datos a resultados, es lo que se conoce como análisis. Estos resultados van ayudar a

describir, explicar y predecir el problema que el investigador cualitativo está tratando de resolver.

Análisis de la variable 1

De las veintisiete variables que conforman el Reporte de Variables Múltiples del Sistema PIAD, veintiséis se incluyeron en el desarrollo del Almacén de Datos, la única variable que, por problema de inconsistencia de datos, no se incluyó fue el grupo al que pertenece el estudiante. Del gráfico mostrado en la figura 14, se deduce que en criterios de población, se incluyó solo el 80% de las variables del criterio de Población, para una inclusión total de un 96%, según muestra la tabla 4.

De la investigación realizada, se desprende que el desarrollo del Almacén de Datos tuvo una alta inclusión de las variables contenidas en el Reporte de Variables Múltiples del sistema PIAD.

Tabla 4 Variables incluidas en el Almacén de Datos PIAD.

Variables del Reporte de Variables Múltiples del PIAD incluidas en el Almacén de Datos PIAD			
Criterio	**Cantidad Variables**	**Variables Aplicadas**	**Porcentaje**
Curso Lectivo	2	2	100%
Población	5	4	80%
Rendimiento Académico y Conducta	2	2	100%
Información familiar del estudiante	7	7	100%
Información individual del estudiante	6	6	100%
Criterios de riesgo	5	5	100%
Total	**27**	**26**	**96%**

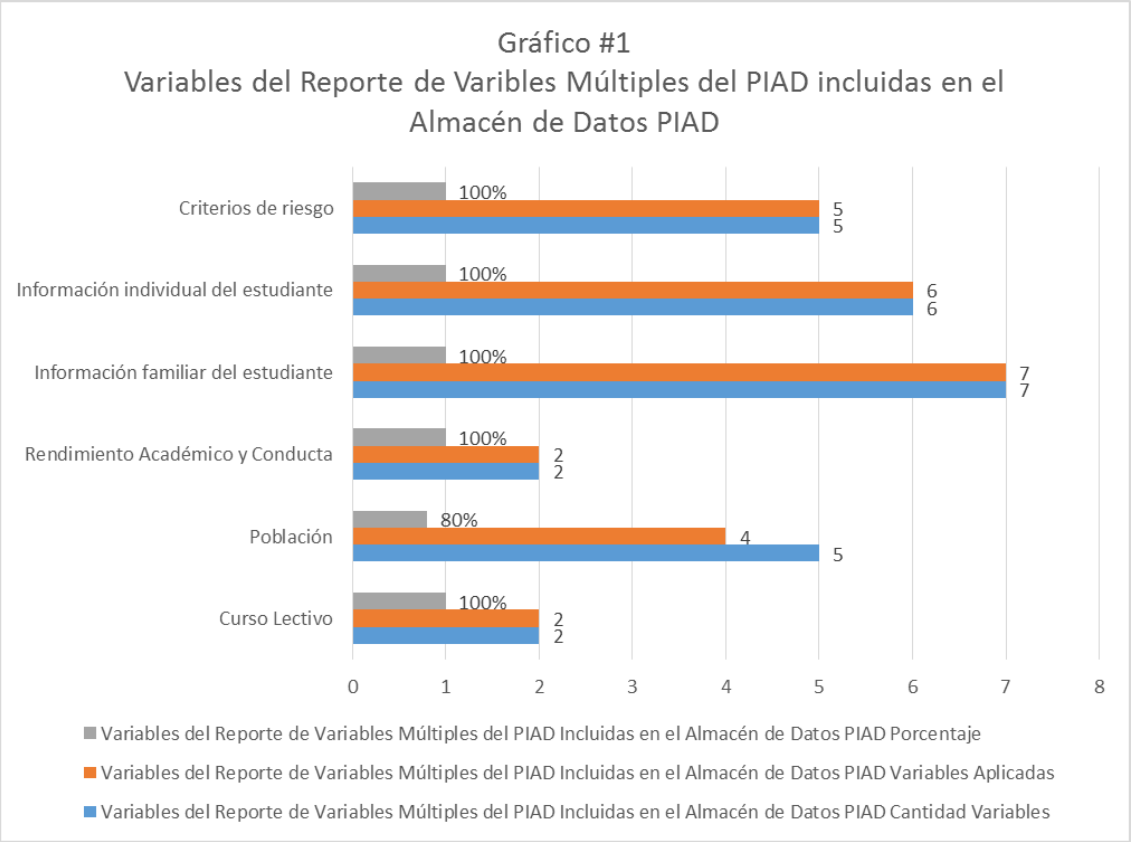


Figura 14 Gráfico de las Variables incluidas en el Almacén de Datos PIAD

Análisis de la Variable 2

De las 229 actividades propuestas en la metodología CRISP, se realizaron 190 en la creación del almacén de datos para determinar los indicadores de la educación pública. Según se muestra en las tablas y gráficos obtenidos mediante las listas de cotejo utilizadas, la aplicabilidad de la metodología dio en un 83%.

Los datos arrojados por la investigación, demuestran que la metodología CRISP, aunque se creó para ser aplicada a proyectos de minería de datos, es altamente aplicable a los proyectos de desarrollo de almacenes de datos.

Tabla 5 Actividades Aplicadas Fase 1 Conocimiento del Negocio.

Fase 1 - Conocimiento del Negocio			
Tarea	**Total Actividades**	**Actividades Realizadas**	**Porcentaje**
1. Determinación de objetivos de negocio	6	6	100%
2. Evaluación de la situación	42	41	98%
3. Determinar objetivos del proyecto	5	4	80%
4. Producción del plan del proyecto	11	11	100%

Total	64	62	97%

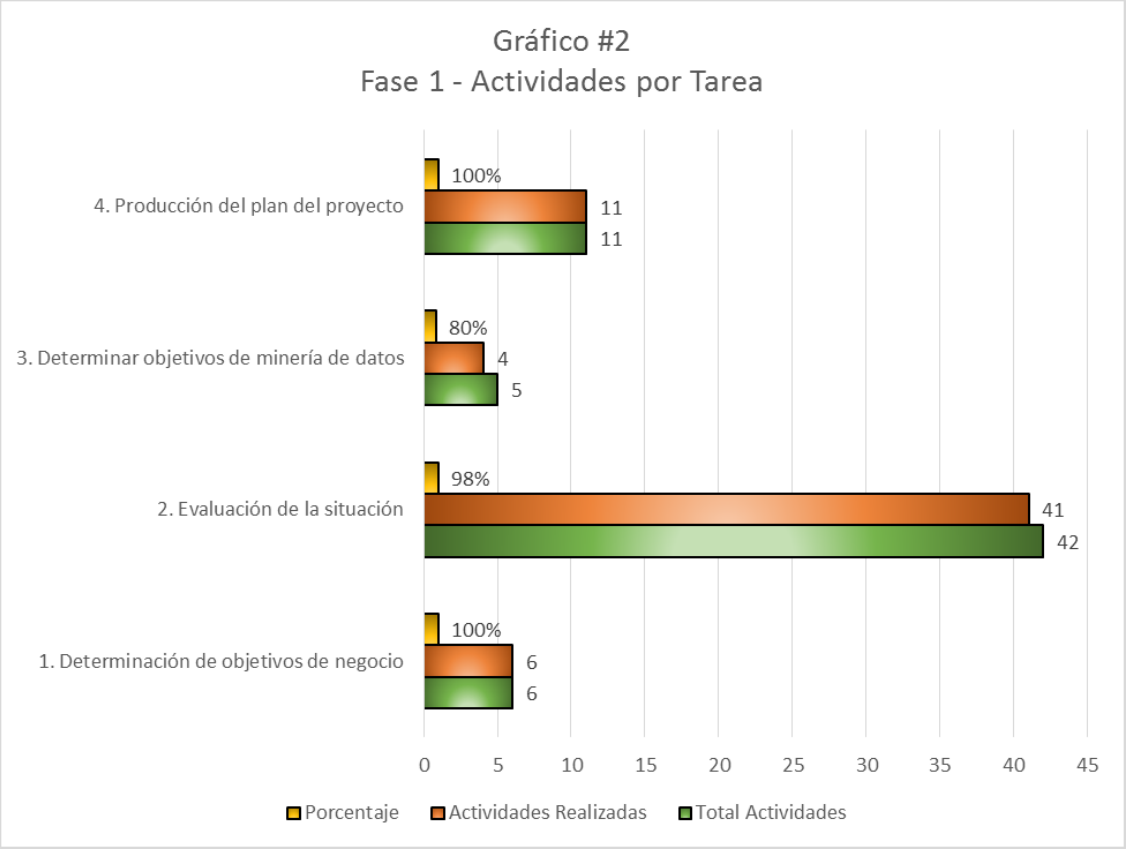


Figura 15 Gráfico de la Aplicación de las actividades de la Fase 1 "Conocimiento del Negocio".

En la Fase 1 de la Metodología CRISP, "Comprensión del Negocio", se proponen sesenta y cuatro actividades de las cuales, para el proyecto de la construcción del Almacén de Datos, se aplicaron un total de sesenta y dos. Las actividades específicas que no se aplicaron fueron:

- Comprobar si la planificación del mantenimiento de hardware se opone a la disponibilidad del hardware, para el proyecto de minería de datos.
- Especificar datos tipo de problema de minería de datos (por ejemplo, la clasificación, la descripción, la predicción, y clustering).

La primera, por considerarse irrelevante para el proyecto, aunque es aplicable a un proyecto de Almacén de Datos; y la segunda por estar orientada específicamente a proyectos de Minería de Datos.

En los resultados arrojados por el gráfico #1, mostrado por la figura 15, se puede observar que la fase de Comprensión de los Datos, fue aplicada en un 97% al proyecto

del Almacén de Datos, ya que trata de conocer el rol del negocio al que se aplicará el proyecto, y de especificar los requerimientos.

Tabla 6 Actividades Aplicadas de la Fase 2 "Comprensión de los Datos".

FASE 2 - Comprensión de los Datos			
Tarea	**Total Actividades**	**Actividades Realizadas**	**Porcentaje**
1. *Recolección de datos iniciales*	9	9	100%
2. *Descripción de datos*	20	18	90%
3. *Exploración de datos*	7	7	100%
4. *Verificación de la calidad de datos*	15	15	100%
Totales	51	49	96%

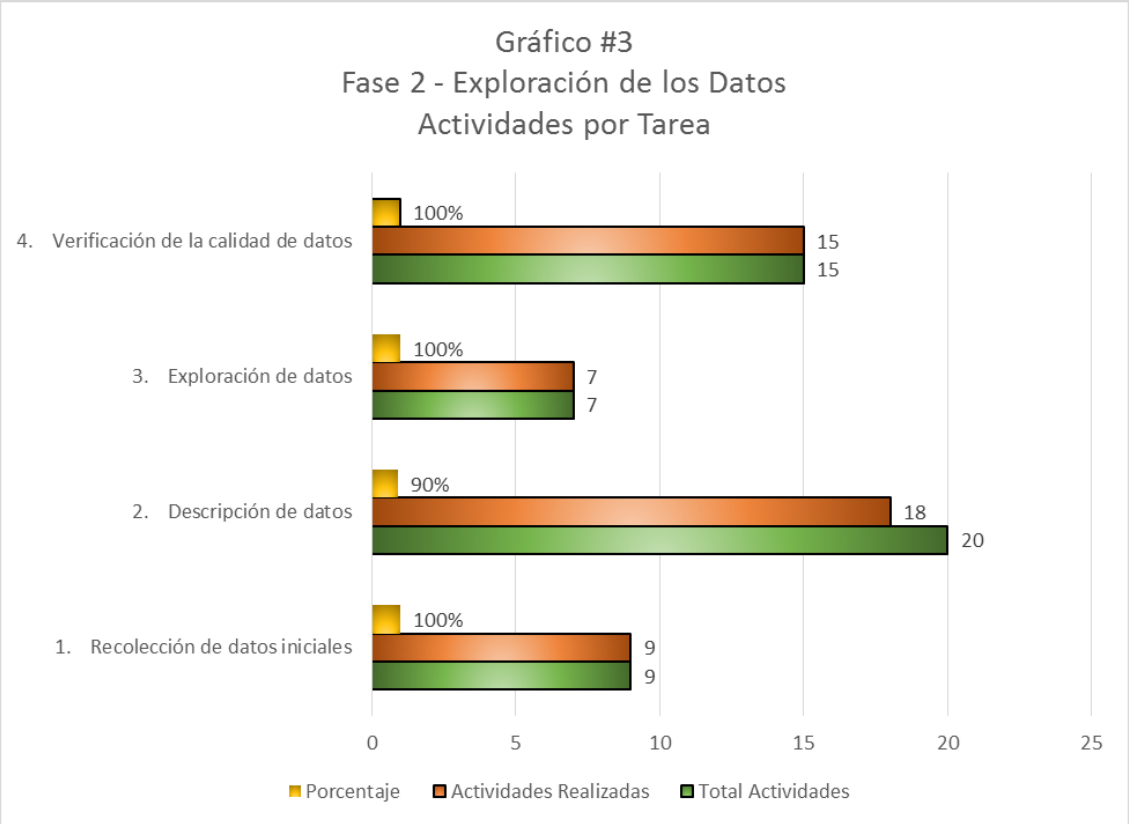


Figura 16 Gráfico de las Actividades aplicadas en la Fase 2 "Comprensión de los Datos".

En la fase 2, "Comprensión de los Datos", compuesta por un total de 51 actividades, se aplicaron 49. Las actividades específicas que no se aplicaron fueron:

- Para cada atributo, calcular la estadística básica (por ejemplo, calcular la distribución, el promedio, el máximo, el mínimo, la desviación estándar, la varianza, la moda, la inclinación, etc.).

- Analizar la estadística básica, y relacionar los resultados con su significado en términos de negocio.

En ambos casos, se consideraron actividades no relevantes para el proyecto del Almacén de Datos.

Según lo visualizado en la figura 16, se deduce que la fase de Comprensión de los Datos fue aplicada en un 96% al proyecto del Almacén de Datos, ya que trata de conocer los datos que van a alimentar al mismo.

Tabla 7 Actividades Aplicadas de la Fase 3 "Preparación de los datos".

Fase 3 - Preparación de los Datos			
Tarea	**Total Actividades**	**Actividades Realizadas**	**Porcentaje**
1. *Selección de Datos*	8	8	100%
2. *Limpieza de datos*	4	4	100%
3. *Construcción de datos*	10	10	100%
4. *Integración de datos*	3	3	100%
5. *Formateo de datos*	4	4	100%
Total	**29**	**29**	**100%**

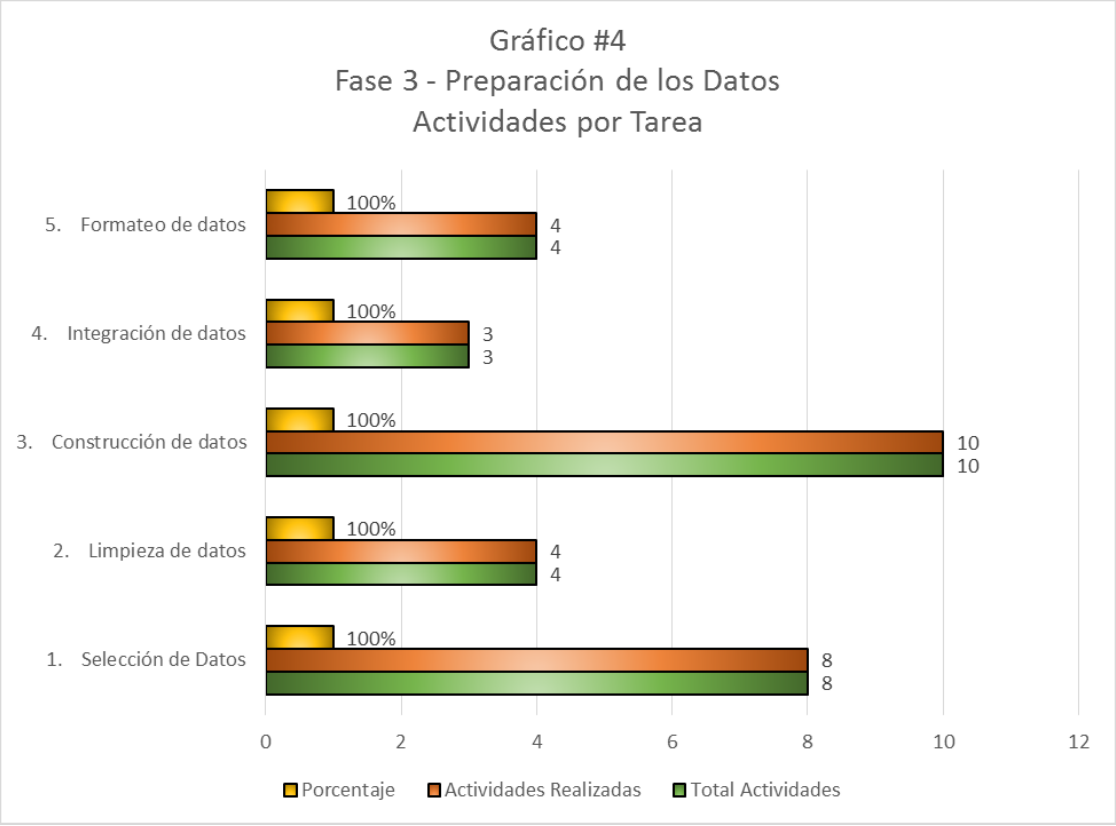


Figura 17 Gráfico de las Actividades aplicadas en la Fase 3 "Preparación de los Datos".

En la fase 3, “Preparación de los Datos”, compuesta por un total de 29 actividades, se aplicaron 29.

Conforme a lo visualizado en la figura 17, se determina la importancia que posee la preparación de los datos; tanto para un proyecto de minería de datos, como para el proyecto actual, ya que la fase fue aplicada en un 100% al proyecto del Almacén de Datos.

Tabla 8 Actividades Aplicadas de la Fase 4 “Modelado”.

Fase 4 – Modelado			
Tarea	**Total Actividades**	**Actividades Realizadas**	**Porcentaje**
1. ***Seleccionar técnicas de modelado***	3	3	100%
2. ***Generar el diseño de prueba***	3	2	67%
3. ***Construcción del modelo***	11	0	0%
4. ***Evaluación del modelo***	16	6	38%
Total	**33**	**11**	**33%**

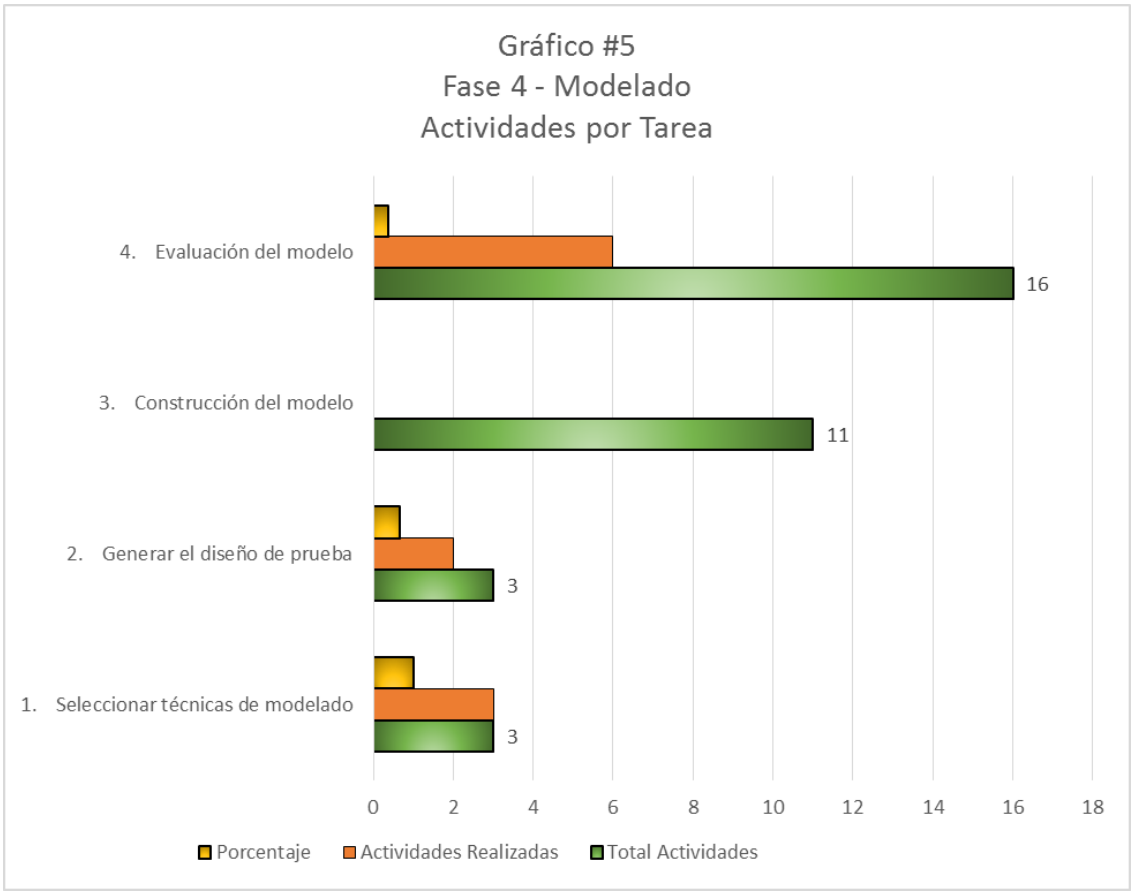


Figura 18 Gráfico de las Actividades aplicadas en la Fase 4 “Modelado”.

En la fase 4, “Modelado”, compuesta por un total de 33 actividades, se aplicaron únicamente 11. Las actividades de esta fase son muy específicas para proyectos de minería de datos:

- Comprobar que existe diseños de prueba separadamente para cada objetivo de minería de datos.
- Decidir los pasos necesarios (el número de iteraciones, el número de desviaciones o curvas, etc.).
- Determinar los parámetros iniciales.
- Documentar las razones para elegir aquellos valores.
- Ejecutar la técnica seleccionada sobre el conjunto de datos de entrada para producir el modelo.
- Post-procesar los resultados de minería de datos (por ejemplo, editar reglas, mostrar árboles).
- Describir cualquier característica del modelo actual que puede ser útil para el futuro.
- Ajustar parámetro de entorno (de registro) usado para producir el modelo.
- Dar una descripción detallada del modelo y cualquier rasgo especial.
- Para modelos basados por regla, listar las reglas producidas, más cualquier evaluación de cada-regla o la exactitud y alcance total del modelo.
- Para modelos no transparentes, listar cualquier información técnica sobre el modelo (como la topología de las redes neuronales) y cualquier descripción de comportamiento producido por el proceso de modelado (como la exactitud o la sensibilidad).
- Describir el comportamiento del modelo y la interpretación.
- Expresar conclusiones respecto a los patrones en los datos (si hay alguno); a veces el modelo revela hechos importantes sobre los datos sin un proceso de evaluación separado (por ejemplo, que la salida o la conclusión son duplicadas en una de las entradas).
- Comparar los resultados de la evaluación y la interpretación.

- Seleccionar los mejores modelos.
- Interpretar los resultados en términos de negocio (tanto como sea posible en esta etapa).
- Conseguir comentarios de los modelos por expertos en datos o en el dominio.
- Comprobar los efectos sobre los objetivos de minería de datos.
- Comprobar los modelos contra una base de conocimiento determinada para ver si la información descubierta es nueva y útil.
- Analizar el potencial para el desarrollo de cada resultado.
- Si hay una descripción verbal del modelo generado (por ejemplo, en forma de reglas), evaluar las reglas: ¿Ellos son lógicos, o ellos son factibles, hay demasiadas reglas o hay demasiado poco, ellos violan el sentido común?
- Conseguir ideas específicas de cada técnica de modelado y ciertos parámetros de ajustes que conduzcan a resultados buenos/malos.
- Ajustar parámetros para producir mejores modelos.

En proyectos de minería de datos se seleccionan algoritmos a aplicar, para obtener los patrones de comportamiento de datos. En el caso del Almacén de Datos, se debe construir un Diseño de Bases de Datos, uno del Servicio de Análisisy uno de despliegue de datos. Las actividades que se aplican a ambos proyectos, es cuando se seleccionan las técnicas para el modelado, y se realizan las suposiciones de la calidad, el formato o las distribuciones de los datos; y en la tarea de evaluar el modelo.

Según lo visualizado en la figura 18, se deduce que, la fase de Modelado fue aplicada en un 33% al proyecto del Almacén de Datos, siendo la de menos aplicación de toda la metodología.

Tabla 9 Actividades Aplicadas de la Fase 5 "Evaluación".

Fase 5 - Evaluación			
Tarea	**Total Actividades**	**Actividades Realizadas**	**Porcentaje**
1. *Evaluación de los resultados*	11	11	100%
2. *Proceso de revisión*	6	6	100%
3. *Determinación de los próximos pasos*	8	7	88%
Total	**25**	**24**	**96%**

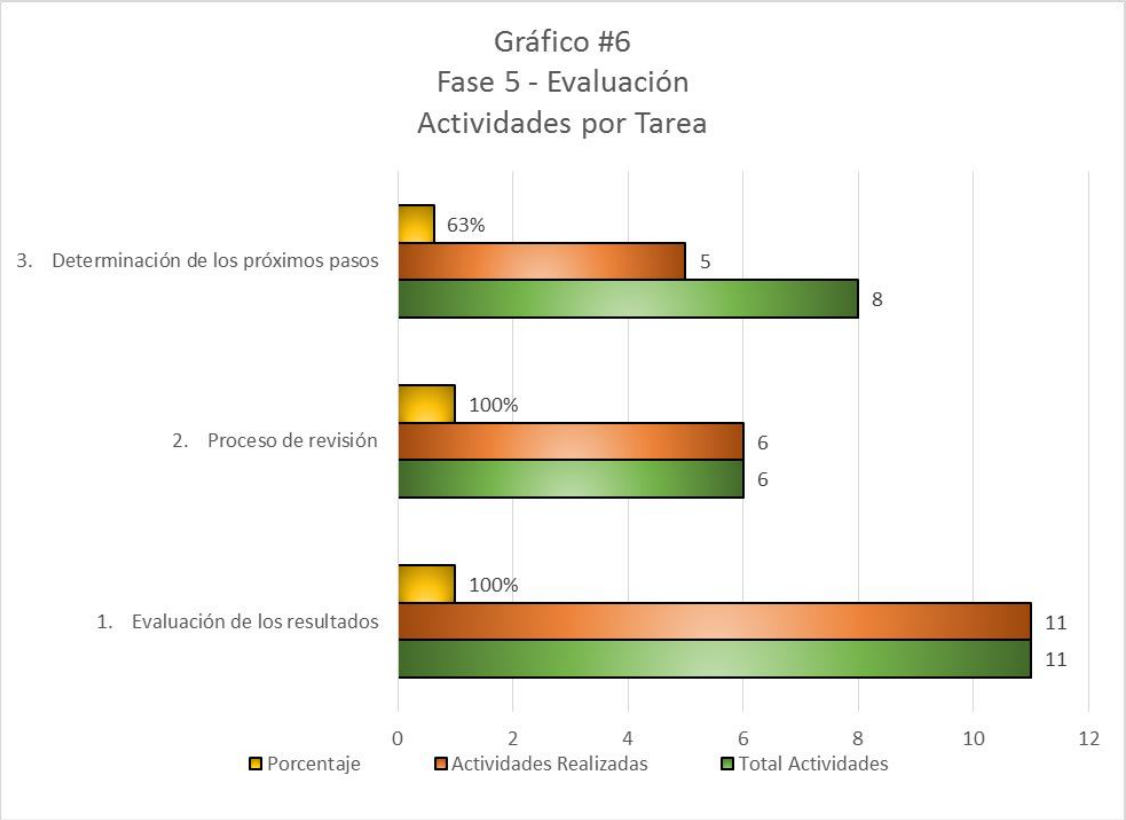


Figura 19 Gráfico de las Actividades aplicadas en la Fase 5 "Evaluación".

En la fase 5, "Evaluación", compuesta por un total de 25 actividades, se aplicaron 24. No se aplicó la siguiente actividad:

- Comprobar los recursos restantes, para determinar si ellos permiten iteraciones de proceso adicionales (o si recursos adicionales pueden estar siendo disponibles).

Como muestra la figura 19, se determina la compatibilidad que existe en la evaluación de un modelo de minería de datos, y la de un proyecto de Almacén de Datos. La fase fue aplicada en un 96% al proyecto del Almacén de Datos.

Tabla 10 Actividades Aplicadas de la Fase 6 "Implementación".

Fase 6 - Implementación			
Tarea	**Total Actividades**	**Actividades Realizadas**	**Porcentaje**
1. *Plan de implementación*	9	4	44%
2. *Supervisión y mantenimiento del plan*	5	5	100%
3. *Producción de Informe final*	7	4	57%
4. *Revisión del proyecto*	6	4	67%
Total	**27**	**17**	**63%**

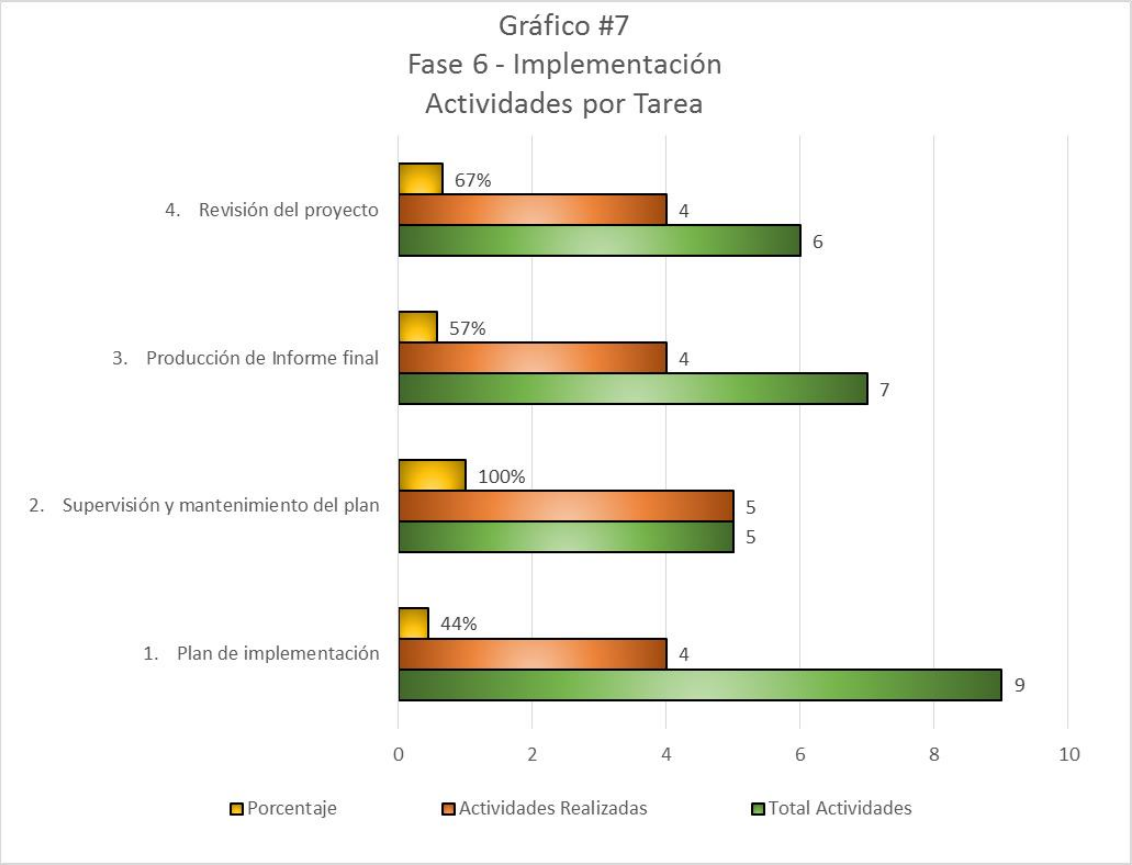


Figura 20 Gráfico de las Actividades aplicadas en la Fase 6 "Implementación".

En la fase 6, "Implementación", compuesta por un total de 27 actividades, se aplicaron 17. Las actividades que no fueron aplicadas, fue por considerarse que están orientadas a los resultados de un proyecto de minería de datos:

- Decidir para cada resultado de conocimiento o información distinto.
- Decidir cómo será supervisado el uso del resultado y medido sus beneficios (donde sea aplicable).
- Decidir por cada resultado de modelo desarrollado o de software.
- Establecer como el modelo o el resultado de software serán desplegados dentro de los sistemas de la organización.
- Determinar cómo su empleo será supervisado y medido sus beneficios (donde sea aplicable).
- Identificar cuáles informes son necesarios (presentación de diapositiva, conclusiones de administración, detalles encontrados, explicación de los modelos, etc.).

- Analizar que tan bien se han encontrado los objetivos de minería de datos iniciales.
- Identificar grupos de objetivos para el informe.
- Entrevistar a toda la gente significativa involucrada en el proyecto y preguntarles sobre su experiencia durante el proyecto.
- Si los usuarios finales trabajan en el negocio con los resultados de minería de datos, entrevistarlos: ¿Están satisfechos? ¿Cómo podría haber sido mejor realizado? ¿Necesitan de apoyo adicional?

Según lo visualizado en la figura 20, se observa que, la aplicabilidad de la fase se dio en un 63% al proyecto del Almacén de Datos, y la tarea donde menos se aplicaron actividades fue en la del plan de implementación, que está muy orientado a los resultados propios de la minería de datos, y no la proporción de información a la que conlleva el Almacén de Datos.

Tabla 11. Aplicabilidad de la Metodología CRISP a un proyecto de Almacén de Datos.

Aplicabilidad de la Metodología CRISP a un proyecto de Almacén de Datos

Etapa	Total Actividades	Actividades Realizadas	Porcentaje
1. ***Conocimiento del Negocio***	64	62	97%
2. ***Conocimiento de datos***	51	49	96%
3. ***Preparación de datos***	29	29	100%
4. ***Modelado***	33	11	33%
5. ***Evaluación***	25	22	88%
5. Implementación	27	17	63%
Total	**229**	**190**	**83%**

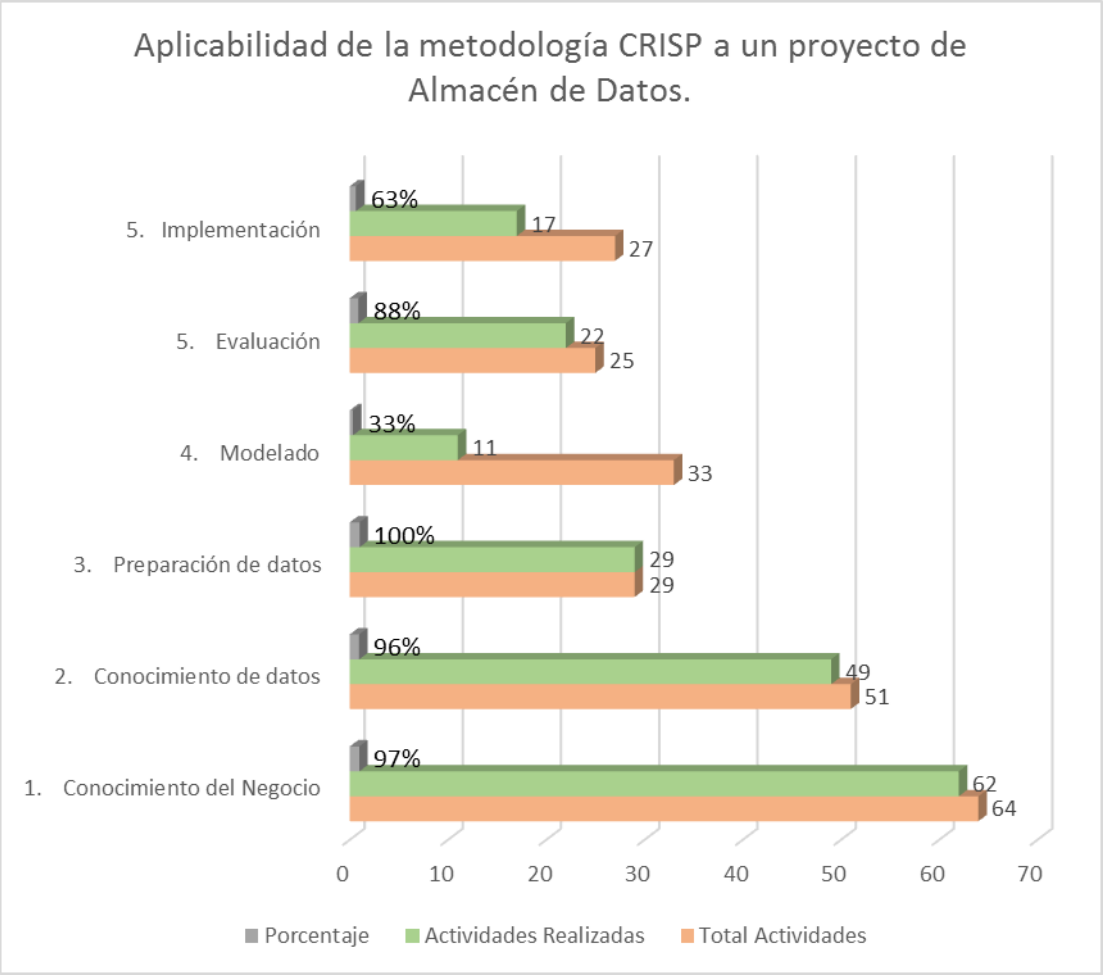


Figura 21. Aplicabilidad de la metodología CRISP a un proyecto de Almacén de Datos.

En resumen, según se muestra en la tabla 11, la metodología CRISP para minería de datos está compuesta por 6 fases, que contemplan en total 229 actividades. En la creación del Almacén de Datos para el PIAD, se aplicaron 190 actividades.

Segú la figura 21, la aplicabilidad de la metodología se dio en un 83%, donde se destaca la fase de preparación de datos, con una aplicabilidad de 100%, y la fase de Modelado con un 33% de aplicabilidad.

Los resultados mostrados hacen evidente que, la metodología CRISP para minería de datos es altamente aplicable a la construcción de un Almacén de Datos, ya que sólo el 17% de las actividades no fueron aplicables a este tipo de proyecto.

4 Propuesta de solución

El presente proyecto pretende desarrollar una herramienta tecnológica de análisis de datos para ASIS.

Se debe integrar en un Almacén de Datos (Data Warehouse) la información que provee el reporte de multivariables del Sistema PIAD, permitiendo integrar la información de distintos centros educativos (10 colegios académicos que utilicen el sistema PIAD en Línea), con el fin de generar indicadores de la calidad de la educación tales como: deserción, ausentismo y rendimiento académico; tomando en cuenta características personales y socioeconómicas de los estudiantes.

Se usará la metodología CRISP para el desarrollo del proyecto, evaluando la eficiencia y eficacia de la misma en el desarrollo de almacenes de datos, por lo cual el presente apartado se desglosa en las seis faces que componen esta metodología.

Para el desarrollo del Almacén de Datos se utilizará la herramienta MS SQL Server 2008 R2, y Excel como herramienta de despliegue de datos por excelencia.

Los orígenes de datos a utilizar, se encuentran en un motor de base de datos MS SQL Server 2008 R2. Las bases de datos se encuentran en forma separada, para cada centro educativo.

4.1.1 Objetivos de negocio y criterios de éxito

Objetivos del negocio

Este apartado detalla los objetivos del negocio que se pretenden alcanzar, con la implementación de un Almacén de Datos, que genere los indicadores de la calidad de la educación pública del país. Los mismos, permitirán medir el alcance y validarlo contra los requerimientos que se detallen.

a. Centralizar la información para toma de decisiones: se requiere centralizar la información de rendimiento académico y asistencia, de los estudiantes de los centros educativos públicos del país, para la presentación de informes que apoyen la toma de decisiones.

b. Permitir la mezcla de variables para generar información: De la manera en que ha sido estructurado el reporte multivariables, realizando combinación de parámetros,

que rigen la información a desplegarse, haciendo posible la mezcla de indicadores. (Apéndice 2)

Criterios de éxito

Al implementar el Almacén de Datos, para la información contenida por los sistemas PIAD, se espera:

CRE - 001 - Tiempos de respuesta al usuario: Para medir el éxito en los tiempos de respuesta, brindados por el Almacén de Datos, se tendrá como indicador de referencia: los tiempos de respuesta que tienen los reportes actuales de los Sistemas PIAD, cuya medición debe ser menor.

CRE – 002 - Capacidad de mantener información histórica: Las bases de datos que almacenan actualmente la información son transaccionales, por lo que mucha información que se modifica no guarda sus valores anteriores. Se espera que el Almacén de Datos permita el guardar la información del día a día, y poder contar con la data histórica de cada carga de información.

CRE -003 - Generación de reportes que permitan el uso de parámetros, para su generación por parte del usuario: Supone que el Almacén de Datos permita al usuario seleccionar entre diferentes parámetros, para generar los reportes utilizando criterios como:

- Periodos.
- Población (Dirección Regional, Circuito, Centro Educativo, Nivel, Grupo).
- Rendimiento académico y conducta.
- Información individual del estudiante.
- Información familiar del estudiante.
- Criterios de riesgo (Adecuación, llegadas tardías, ausentismo, entre otros).

Se espera que la selección de los parámetros, no tenga dependencia de los desarrolladores del Almacén de Datos o los encargados de mantenimiento del mismo.

CRE -004 - Diseño del Almacén de Datos: El diseño del Almacén de Datos debe cubrir las necesidades de información de los indicadores de calidad, de la educación pública. Para ello, debe tomarse en cuenta el reporte variables múltiples del Sistema PIAD en línea.

CRE-005 - Seguridad de la información: Se espera que el Almacén de Datos permita la definición de perfiles y permisos de autorización, sobre la información a desplegar; contemplando la necesidad a cubrir en cada población (Centro Educativo, Circuito, Regional u oficinas centrales).

4.1.2 Inventario de recursos

La sección de Inventario de Recursos identifica: el personal, fuentes de datos, instalaciones técnicas, y otros recursos que requiere la realización del proyecto.

a. Recursos Humanos

RHU – 001 - Líder de Negocio (1): Persona conocedora del Rol de negocio, y las necesidades de información que debe cubrir el Almacén de Datos.

RHU – 002 - Líder Técnico del Sistema PIAD (1): Experto en la aplicación que almacena la información, que será cargada en el Almacén de Datos.

RHU -003 - Desarrollador de aplicaciones de Almacén de Datos (2): **E**ncargado de analizar los requerimientos, diseñar, implementar, y probar la solución de Almacén de Datos.

RHU -004 - Inspector de Calidad (1): Delegado de evaluar la calidad de los entregables del proyecto.

RHU-005 - Experto en Almacenes de Datos (1): Persona experta en el desarrollo e implementación de almacenes de datos, que evaluará la solución propuesta.

RHU-006 - Líder Gerencial (2): Persona de mando gerencial del negocio que evalúa el cumplimiento de los objetivos.

b. Fuentes de datos

Actualmente, el Piad trabaja en cinco versiones de fuentes de datos:

FDT – 001 - Base de Datos de Escuela Primaria: Utiliza un motor de base de datos MySQL en servidor local. Las escuelas primarias cuentan con una única computadora, en la que se instala el sistema y la base de datos, o por medio de una arquitectura cliente servidor en una red interna.

FDT – 002 - Base de Datos de Colegio Académico Local: Emplea un motor de base de datos MySQL en servidor local. Los colegios académicos, que trabajan en forma local, cuentan con una única computadora en la que se instala el sistema y la base de datos, o por medio de una arquitectura cliente servidor en una red interna.

FDT – 003 - Base de Datos de PIAD en Línea: Utiliza un motor de base de datos MS SQL Server 2008 R2, en servidor en línea, que se encuentra en la DGEC. Contiene la información requerida, para aplicar seguridad a los usuarios del PIAD en Línea de los colegios Académicos, que utilizan este sistema.

FDT – 004 – Base de Datos de Colegio Académico en Línea: Utiliza un motor de base de datos MS SQL Server 2008 R2, en servidor en línea, que se encuentra en la DGEC. Contiene la información del Sistema PIAD en Línea de los colegios Académicos.

FDT – 005 - Base de Datos de Colegio Técnico Profesional (CTP): Utiliza un motor de base de datos MySQL en servidor local. Los Colegios Técnicos Profesionales trabajan en forma local, cuentan con una única computadora en la que se instala el sistema y la base de datos, o por medio de una arquitectura cliente servidor en una red interna.

c. Instalaciones técnicas

Las instalaciones técnicas, con las que se cuenta, son las oficinas del PIAD que se localizan en la Escuela Elías Jiménez, en San Rafael debajo de Desamparados en San José.

d. Otros recursos

Servidor de datos: el servidor de datos en que se realizarán las pruebas del Almacén de Datos, será el provisto por ASIS.

Equipos de desarrollo: los equipos de desarrollo serán las computadoras de los desarrolladores del proyecto.

Equipos y materiales de impresión: los equipos y materiales de impresión serán cubiertos por los desarrolladores.

4.1.3 Requerimientos, suposiciones, y restricciones

Esta sección lista los requerimientos generales para la ejecución del proyecto: tipo de resultados de proyecto, presunciones hechas sobre la naturaleza del problema, y de los datos que están siendo usados y restricciones impuestas al proyecto.

a. Requerimientos:

REQ-001 - Recurso Humano: Para el éxito del proyecto, se requiere del tiempo y disposición del siguiente recurso humano:

- Director del PIAD: Es quién dicta los requerimientos de información que se deben cubrir con el Almacén de Datos, y da visto bueno a los informes y avances del proyecto.
- Representante de la Dirección de Gestión y Evaluación de la Calidad del MEP (DGEC): Persona que se encargará de revisar el cumplimiento de los objetivos del proyecto.
- Encargado de desarrollo del PIAD: Encargado de proveer la infraestructura, para implementar el Almacén de Datos; brindar las bases de datos requeridas, para cargar la información; y asesorar en cuanto a la especificación del diseño y los datos de las bases de datos, utilizadas actualmente.
- Desarrolladores del Almacén de Datos: Se contará con dos desarrolladores a medio tiempo, para desarrollar el proyecto.
- Evaluadores del Almacén de Datos: Con el fin de obtener un producto de alta calidad y rendimiento, se va a requerir de al menos dos profesionales especialistas en Almacén de Datos, que evalúen las propuestas de diseño e implementación, y presenten las recomendaciones de mejora.

La disposición de las personas del PIAD y el MEP, es responsabilidad del patrocinador del proyecto.

La labor de los desarrolladores será supervisada por un tutor académico del proyecto.

La participación de los especialistas evaluadores del Almacén de Datos, será coordinada por los desarrolladores.

REQ-002 – Infraestructura: Se requiere contar con la infraestructura que albergará el Almacén de Datos. Inicialmente, el PIAD en Línea estuvo albergado en una arquitectura provista por el Data Center de RACSA. Por razones de negociación, entre esta entidad y el MEP, las bases de datos se instalaron en un servidor provisto por la DGEC.

REQ-003 – Software: Se requiere contar con las licencias de MS SQL Server 2008 R2, en una versión completa, que incluya las herramientas de Inteligencia de Negocios (BI)

y Reporting Services, para la elaboración del Almacén de Datos. El proveer las licencias de software requeridas, es responsabilidad de ASIS.

b. Suposiciones

En el siguiente apartado, se detallan todos los supuestos bajo los que se elaborará el diseño del Almacén de Datos; de darse algún cambio en ellos podría ocasionar que se cambie el diseño del mismo.

SUP-001 – Uso de una selección de datos limitados: Actualmente, existen 35 colegios académicos trabajando la versión del Piad en línea. Para lograr un despliegue de información representativa, se hará una selección de 10, tomando en cuenta la cantidad de datos que poseen.

SUP-002 – Sistema de administración de bases de datos: El sistema de administración de bases de datos a utilizar es el Microsoft SQL Server 2008 R2, ya que es la herramienta de la que se tienen licencias, y bajo la que está funcionando el proyecto piloto PIAD en línea.

SUP-003 – Infraestructura: El Almacén de Datos se implementará sobre la infraestructura, que actualmente alberga el PIAD en línea; de existir algún problema se utilizará el servidor del PIAD que se encuentra en las oficinas de ASIS.

SUP-004 – Sistema operativo: El sistema operativo a utilizar será Windows Server 2008 R2 Enterprise.

SUP-005 – Independencia de los procesos operativos y los de toma de decisiones: Se debe mantener la independencia de las bases de datos del Sistema Operacional PIAD, de las bases de datos del Almacén de Datos de toma de decisiones.

c. Restricciones

Las restricciones del manejo de la información a cargar en el Data Warehouse, se pueden resumir en:

RST-001 – Orígenes de datos: El Almacén de Datos se implementará, usando las herramientas que provee MS SQL Server 2008 R2. Los orígenes de datos que se van a utilizar, serán las bases de datos de aquellos centros educativos que están trabajando en

PIAD en Línea. No se utilizaran las bases de datos en MySQL, ya que están prontas a migrarse a PIAD en línea.

RST-002 – Gestión de usuarios: La gestión de los usuarios debe llevarse a cabo, por parte del administrador que se designe al Almacén de Datos, por lo cual se debe proveer los roles con los permisos requeridos, para realizar la gestión.

RST-003 – Conexiones a Internet: La capacidad de conexión que van a requerir los usuarios del Almacén de Datos, va a estar determinada por el volumen de datos que requieran extraer.

RST-004 – Importación y exportación de información: Actualmente, los registros de los estudiantes son trabajados por los docentes de primaria en Excel; y por los docentes de secundaria en ACCESS. Estos registros, no serán considerados en la información a cargar en el Almacén de Datos, únicamente aquella información que se considere relevante, para determinar las variables que inciden en el rendimiento académico, el ausentismo y deserción en los centros educativos públicos.

RST-005 – Hardware: El sistema se debe implementar sobre la infraestructura (equipos cliente, equipo servidor y sus conexiones) que provea el ASIS para este fin. ASIS es una asociación sin fines de lucro, por lo que sus recursos son limitados. La adquisición de una arquitectura adecuada en software y hardware, es clave para el éxito de este proyecto, y depende del MEP.

RST-006 – Software: El sistema debe desarrollarse con MS SQL Server 2008 R2, utilizando un sistema operativo de Windows Server 2008 R2 Enterprise. El equipo de desarrollo cuenta con herramientas de uso académico.

RST-007 - Seguridad sobre datos sensibles: El sistema debe permitir el manejo de la seguridad sobre el acceso a la información, o acciones propias sobre el Almacén de Datos; permitiendo integrar las políticas que se establezcan para administrar y controlar el acceso, y las acciones que se lleven a cabo con la información procesada. Se debe autenticar y autorizar el ingreso de usuarios al sistema, regular el acceso a una acción sobre los datos mediante el perfil del usuario, entre otros.

4.1.4 Análisis de Riesgos

Un proyecto ambicioso como es el desarrollo de un Almacén de Datos, siempre se enfrentará a situaciones de diferentes orígenes, que comprometen el alcance de los objetivos planteados inicialmente. Tener una correcta administración del riesgo, permitirá controlar y mitigar el impacto que estos tendrán de llegar a materializarse.

A continuación, se presentan los riesgos del proyecto con la propuesta de contingencia agrupados, según su origen.

a. Riesgos de negocio

Riesgo	**Contingencia**
Falta de disponibilidad de los Datos	Selección de un subgrupo de datos representativo.
Falta de disponibilidad de los conocedores del negocio.	Establecer un cronograma de trabajo previo al proyecto.
Apoyo del patrocinador del proyecto.	Establecer carta de compromiso.

b. Riesgos de la organización

Riesgo	**Contingencia**
La prioridad de proyectos. Puede determinarse que otros proyectos son prioritarios para el personal de la organización, sobre el desarrollo del Almacén de Datos.	Utilizar las tecnologías disponibles, para realizar reuniones virtuales y consultas por correo electrónico.
Dependencia del MEP en toma de decisiones, para contar con la infraestructura y responsables del Almacén de Datos.	Establecer una infraestructura mínima por parte de ASIS, para realizar las pruebas. Y como último recurso, usar las computadoras de los desarrolladores como contingencia.

c. Riesgos financieros

Riesgo	**Contingencia**

Presupuesto limitado para el desarrollo de proyectos de la organización.	El proyecto será desarrollado como un proyecto de graduación de la Maestría en Bases de Datos, de la Universidad CENFOTEC.
Alto costo del desarrollo de un Almacén de Datos.	Al tratarse de un proyecto con un fin social, se cuenta con varios profesionales, dispuestos a donar sus horas de servicio, para alcanzar el éxito del mismo.

d. Riesgos técnicos

Riesgo	**Contingencia**
Que la capacidad existente de la infraestructura con la que se cuenta actualmente, no soporte lo requerido por el Almacén de Datos.	De no tener una infraestructura adecuada para las pruebas de la herramienta, estas se llevarán a cabo en las máquinas de los estudiantes.
El impacto de migrar a nuevas versiones de la herramienta utilizada, para el desarrollo del Almacén de Datos (MS SQL Server 2008 R2).	Antes de llevar a cabo una migración, valorar las nuevas características, y cuáles tienen un gran impacto en el desarrollo del Almacén de Datos.

e. Riesgos que dependen de datos y de las fuentes de datos

Riesgo	**Contingencia**
Contar con los respaldos de las bases de datos que se trabajan de forma local.	Se harán las pruebas con las bases de datos que se encuentran en línea.
Los datos incompletos, ya que algunos centros educativos no usan todos los módulos del sistema.	Se realizará un análisis de los datos antes de desarrollar la herramienta, para evitar o dar un tratamiento especial a aquellos que se encuentren incompletos.
Datos sin normalización en las bases de datos.	Trabajar en la carga de los datos por medio de ETL's, antes de cargarlos al Almacén

	de Datos.

4.1.5 Costos y Beneficios

El desarrollo del proyecto se calculó con un total de 185 días de trabajo, según se muestra en el cronograma del apéndice 1, se requirió de recurso humano especializado en seis áreas de negocio, cuyos costos están reflejados en la tabla 12; con las respectivas cargas de trabajo asignadas y un costo por hora significativo. Además, se incluyen los gastos administrativos propios de un proyecto de desarrollo de software.

El beneficio que aportará la herramienta a la educación pública costarricense, se verá reflejado en información verídica y eficaz que se pondrá a disposición de los diferentes mandos, que podrán utilizarla para la toma de decisiones asertivas que lleven a mejora la calidad de la educación y reducir los niveles de pobreza de la población.

a. Costos

Tabla 12: Costos del proyecto Almacén de Datos PIAD.

Proyecto Almacén de Datos PIAD

Costos

Precio Dólar ₡504,00

Descripción	U. Medida	Costo Unitario	Cantidad	Total $	Total ₡
Líder de Negocio.	Hora	60	148	$8.880,00	₡4.475.520,00
Líder Técnico del Sistema PIAD	Hora	35	222	$7.770,00	₡3.916.080,00
Desarrollador de aplicaciones de Almacén de Datos	Hora	35	1480	$51.800,00	₡26.107.200,00
Inspector de Calidad.	Hora	60	148	$8.880,00	₡4.475.520,00
Experto en Almacenes de Datos.	Hora	80	74	$5.920,00	₡2.983.680,00
Líder Gerencial.	Hora	60	444	$26.640,00	₡13.426.560,00
Gastos materiales de Oficina.	Porcentaje	2	2	$219,78	₡110.769,12
Otros gastos	Porcentaje	1	1	$101,23	₡51.019,81
Costo Total del Proyecto				**$110.211,01**	**₡55.546.348,93**

Nota: Todas las personas involucradas en el proyecto han prestado sus servicios en calidad de donación al proyecto PIAD, y los demás costos serán asumidos por los desarrolladores (estudiantes).

b. Beneficios

BNF-001 – Mejora en los procesos de toma de decisiones: Los procesos de toma de decisiones pueden ser mejorados, mediante la disponibilidad de información. Las decisiones empresariales se hacen más rápidas, por gente más informada.

BNF-002 – Procesos empresariales optimizados: Los procesos empresariales pueden ser optimizados. El tiempo perdido esperando por información, que finalmente es incorrecta o no encontrada, es eliminado.

BNF-003 – Datos con significado empresarial: Procesos y datos de los sistemas operacionales, así como los datos en el Almacén de Datos; son usados y examinados. Cuando los datos son organizados y estructurados, para tener significado empresarial, la gente aprende mucho de los sistemas de información. Pueden quedar expuestos posibles defectos en aplicaciones actuales, siendo posible mejorar la calidad de nuevas aplicaciones.

BNF-004 – Mejora en la confianza de la toma de decisiones: Las personas tendrán mayor confianza en las decisiones empresariales que se toman. Tanto, quienes toman las decisiones, como los afectados, conocerán que están basadas en buena información.

BNF-005 – Lenguaje común del negocio: La información compartida conduce a un lenguaje común, conocimiento común, y mejoramiento de la comunicación en la empresa. Se mejora la confianza y cooperación entre distintos sectores de la empresa, viéndose reducida la sectorización de funciones.

BNF-006 – Visibilidad, Accesibilidad y Conocimiento de los Datos: La **v**isibilidad, accesibilidad y conocimiento de los datos, producen mayor confianza en los sistemas operacionales.

4.1.6 Objetivos del Almacén de Datos

Un Almacén de Datos es importante, ya que sirve como un sistema de apoyo a las decisiones. Además, sirve para organizar la utilización de los datos, para llegar a los hechos, las tendencias o las relaciones, que pueden ayudarles a tomar decisiones efectivas o crear estrategias eficaces, para lograr los objetivos de la organización.

El desarrollo del Almacén de Datos, para el sistema PIAD, plantea los siguientes objetivos como herramienta para la toma de decisiones:

OAD-001 – Hacer que la información de la organización sea accesible: Los contenidos del Almacén de Datos son entendibles y navegables, y el acceso a ellos es caracterizado por el rápido desempeño. Estos requerimientos no tienen fronteras y tampoco limites fijos. Cuando hablamos de entendible significa, que los niveles de la información sean

correctos y obvios. Y navegables significa, reconocer el destino en la pantalla y llegar a donde queramos con solo un clic. Rápido desempeño significa, cero tiempo de espera.

OAD-002 – Hacer que la información de la organización sea consistente: La información de una parte de la organización, puede hacerse coincidir con la información de la otra parte. Si dos medidas de la organización tienen el mismo nombre, entonces deben significar la misma cosa; y a la inversa, si dos medidas no significan la misma cosa, entonces son etiquetados diferentes. Información consistente significa, información de alta calidad. Es decir, que toda la información es contabilizada y completada.

OAD-003 – Permitir que la información sea adaptable y elástica: El Almacén de Datos debe ser diseñado para cambios continuos. Cuando se hagan nuevas preguntas al Almacén de Datos, los datos existentes y las tecnologías no cambian ni se corrompen. Cuando se agreguen datos nuevos al Almacén de Datos, los datos existentes y las tecnologías tampoco deben cambiar. El diseño de los Mercados de datos (Data Marts) debe ser distribuido e incrementado.

OAD-004 – Ser un soporte a la toma de decisiones inteligentes: El Almacén de Datos tiene los datos correctos, para soportar la toma de decisiones. La información que presenta, es un estado del día a día de la organización, ya que los datos son alimentados de los sistemas transaccionales y los repositorios, utilizados en el trabajo diario.

OAD-005 – Proteger los valores de la información: El Almacén de Datos, no solamente controla el acceso efectivo a los datos, sino que, brinda a los dueños de la información gran visibilidad en el uso y abusos de los datos, aún después de haber dejado el Almacén de Datos. Todo lo demás es un compromiso, y por consiguiente algo que queremos mejorar.

4.1.7 Criterios de éxitos del Almacén de Datos

El establecimiento de los criterios de éxito permite llevar a cabo las mediciones correspondientes, para verificar que la herramienta desarrollada cumple los objetivos planteados.

CEA- 001 –Ágil, flexible y amigable al usuario: El Almacén de Datos debe ser ágil en la presentación de la información, flexible a los datos que debe contener en diferentes formatos, y fácil de usar para el usuario.

CEA -002 – Facilidad de uso: Debe presentar al cliente los datos, de forma fácil y simple. El despliegue de los datos en herramientas, como Excel, conocidas en los distintos gremios de trabajo, debe dar esa facilidad para combinar, examinar y analizar la información.

CEA-003 – Seguridad: Mostrar la información a aquellos que cuentan con la debida autorización, haciendo uso de la autenticación y autorización de la herramienta utilizada.

CEA-004 – Soporte a la agregación de datos y generación de reportes: El Almacén de Datos debe permitir que se agreguen nuevos datos, para su posterior análisis; y que, a partir de estos, se generen los reportes requeridos para la toma de decisiones.

CEA-005 – Integración datos: Posibilitar la integración de los datos de las escuelas primarias, colegios técnicos profesionales y colegios académicos, para lograr generar indicadores que determinen la calidad de la educación pública, en el país.

CEA -006 - El uso de variables temáticas: El Almacén de Datos debe aceptar filtrar la información, según distintas variables temáticas que el usuario pueda seleccionar.

4.2 Fase 2 – Comprensión de los datos

La fase de comprensión de datos implica, estudiar más de cerca los datos que alimentarán el Almacén de Datos. Este paso es esencial para evitar problemas inesperados durante fase de preparación de datos, que suele ser la más larga de un proyecto.

La comprensión de los datos implica acceder a los datos, y explorarlos con la ayuda de tablas y gráficos que se pueden organizar, con el fin de determinar la calidad de los datos.

4.2.1 Recolección de datos inicial

El sistema transaccional utilizado en los diferentes centros educativos, ofrece a los usuarios un gran número de reportes. Estos, en su mayoría, son de carácter operativo, por cuanto sólo competen a la institución, donde se ejecutó el reporte. Estos son reportes, los cuales no agrupan a más de una institución; pero existe un reporte, el cual es utilizado para analizar estratégicamente la calidad de la educación en cada institución, que aunque este se ejecute, por ahora, en forma local, en cada institución

educativa, se podría utilizar a nivel nacional; cuando este pueda ser aplicado y ejecutado en el Almacén de Datos.

Existe un reporte utilizado para combinar variables y generar información, para generar los indicadores de la calidad de la educación pública, llamado "Reporte de Variables Múltiples". Este, realiza conteos de estudiantes, de acuerdo a una serie de criterios, tales como: repitencia, deserción, asignación de beca, adecuación, entre otros. Por lo cual, este reporte será el que defina la información, que debe contener el Almacén de Datos a crear.

Este reporte es de gran importancia para el éxito del proyecto, por el cual se tomará como la base de modelo multidimensional, que resultará en el cubo de análisis de la calidad de la educación en Costa Rica, de las instituciones que sean incorporadas en el Almacén de Datos.

La información utilizada, por el "Reporte de Variables Múltiples", está 100% disponible en la base de datos de cada institución educativa. Sin embargo, existe otro tipo de problema el cual tuvimos que resolver: el sistema transaccional fue desarrollado de tal forma que, genera las consultas en forma dinámica, es decir que dentro del código fuente de la aplicación, contiene partes de las consultas necesarias, para poder ejecutar el reporte en cuestión. Estas consultas se generan con base a los criterios que el usuario elija, a la hora que desea ejecutar el reporte.

Para poder analizar e identificar las fuentes exactas que utiliza el reporte, fue necesario crear un ambiente de pruebas local, el cual se compone del código fuente de la aplicación transaccional, que se encuentra en producción.

Una vez establecido el ambiente de prueba, se definieron una serie de casos basados en las opciones de parámetros, que ofrece el reporte a la hora de ejecutarlo. Además, se utilizó una técnica de desarrollo de aplicaciones, la cual se llama *depurar* que consiste en colocar puntos de revisión dentro de la aplicación, lo que ayudo a extraer las distintas consultas realizadas a la base de datos, para generar el reporte; esto mientras se ejecutaba la aplicación.

El reporte de variables múltiples consta de las siguientes secciones de parámetros:

- Curso lectivo
- Criterios de población
- Criterios de rendimiento académico y conducta
- Información individual del estudiante

- Información familiar del estudiante
- Criterios de riesgo

Cada uno de las secciones, antes mostradas, contiene una serie de opciones las cuales el usuario puede seleccionar, para la generación del reporte. Cada opción que el usuario seleccione afectará la consulta final, que realice a la base de datos. Más adelante, se mostrara con detalle los casos seleccionados, para la generación de consultas.

Criterios de selección

Para poder identificar los datos necesarios, utilizados en el reporte de variables múltiples, se definieron una serie de casos, los cuales ayudarían a identificar las consultas enviadas por la aplicación transaccional a la base de datos. A continuación, se muestra en la tabla 13, el detalle de los casos definidos, para la técnica de depuración antes mencionada:

Tabla 13: Criterios de selección del Reporte de Variables Múltiples del Sistema PIAD

Caso	Tiempo	Criterios de población	Criterios de rendimiento académico y conducta	Información individual del estudiante	Información familiar del estudiante	Criterios de riesgo
1	Curso lectivo: 2009, Periodo : Primer periodo	Nivel o Año: Séptimo	Asignatura básica individual: Matemática, Notas > 100 and Notas <= 100	Ninguna	Ninguna	Ninguna
2	Curso lectivo: 2009, Periodo : Primer periodo	Centro Educativo	Asignatura básica individual: Ciencias, Notas > 100 and Notas <= 100, Conducta	Ninguna	Ninguna	Adecuaciones: No significativa
3	Curso lectivo: 2009, Periodo : Primer periodo	Grupo: 1	Asignaturas básicas, Notas > 100 and Notas <= 100	Ninguna	Ninguna	Llegadas tardías
4	Curso lectivo: 2009, Periodo : Primer periodo	Estudiante= ALTAMIRANO….	Asignatura técnica o artística: Religión, Notas > 100 and Notas <= 100	Ninguna	Ninguna	Caso de ausentismo
5	Curso lectivo: 2009, Periodo : Primer periodo	Grupo: 1	Todas las asignaturas, Notas > 100 and Notas <= 100	Ninguna	Ninguna	Prioridad en el comedor

6	Curso lectivo: 2009, Periodo : Primer periodo	Grupo: 1	Asignatura básica individual: Ciencias, Notas > 100 and Notas <= 100	Ninguna	Ninguna	Beca
7	Curso lectivo: 2009, Periodo : Primer periodo	Nivel o Año: Decimo	Asignatura básica individual: Matemática, Notas > 100 and Notas <= 100	Ninguna	Ninguna	Hogar uniparental
8	Curso lectivo: 2009, Periodo : Primer periodo	Grupo: 4	Asignatura básica individual: Matemática, Notas > 100 and Notas <= 100	Ninguna	Ninguna	Tipo de beca: Comedor Estudiantil

Los casos mostrados, fueron clave para poder identificar las consultas enviadas por la aplicación a la base de datos, basada en los criterios seleccionados por los usuarios. Es necesario aclarar, que para los criterios de información individual del estudiante y la información familiar, se realizó un análisis diferente, ya que el ambiente de prueba generaba problemas a la hora de utilizar dichos parámetros; posteriormente se explicará el procedimiento realizado. En los criterios de rendimiento académico y conducta, se utilizaron rangos de notas fuera de proporción, con el fin de que no se devolvieran datos; ya que el objetivo de los casos es generar la consulta en lenguaje SQL, que referencia la tabla o tablas y los campos que contienen la información, en un menor tiempo de respuesta.

Con respecto a la información individual del estudiante y la información familiar, se realiza un análisis muy diferente. Al tener problemas de ejecución en el ambiente de pruebas, se toma la decisión de utilizar las tablas del sistema del motor de bases de datos de SQL Server 2008 R2. Estas tablas describen el catálogo de objetos, dentro de una base de datos, en este caso tablas, vistas y columnas, entre otras cosas. Para poder identificar la información necesaria para las secciones antes mencionadas, se utiliza la siguiente consulta:

```
SELECT t.name AS table_name,
SCHEMA_NAME(schema_id) AS schema_name,
c.name AS column_name
FROM sys.tables AS t
INNER JOIN sys.columns c
```

```
ON t.OBJECT_ID = c.OBJECT_ID
WHERE upper(c.name) LIKE '%COND%'
ORDER BY schema_name, table_name;
```

El código mostrado, busca en las tablas del sistema nombres de datos importantes, para el análisis. El reporte de Variables Múltiples contiene una serie de parámetros de suma importancia, que se requieren en el análisis. A continuación, se presenta la tabla 14 y la tabla 15, que describen a detalle todos los parámetros utilizados en este reporte:

Tabla 14: Parámetros del reporte de Variables Múltiples (1).

Curso Lectivo	Criterios de población	Criterios de rendimiento académico y conducta	Información individual del estudiante
Curso lectivo Periodo	Centro educativo Docente Estudiante Grupo Nivel o Año	Asignatura básica individual Asignatura técnica o artística Asignaturas básicas Conducta (and) **(PROB)** Notas mayores a (and no opcional) Todas las asignaturas	Adecuación curricular Asistencia Casos de ausentismo Deserción Edad Prioridad en el comedor escolar Recibe beca Recibe bono escolar Recibe transporte Repitencia Sexo Tipos de becas Trabaja el estudiante

Tabla 15: Parámetros del reporte de Variables Múltiples (1).

Información familiar del estudiante	Criterios de riesgo
Acceso a Internet	Adecuaciones
Encargados que viven con el estudiante	Llegadas tardías
Escolaridad de la madre	Caso de ausentismo
Escolaridad del padre	Prioridad en el comedor
Hogar uniparental	Beca
Ingreso mensual per cápita	Ingreso per cápita mensual
Tipo de vivienda	Hogar uniparental

	Tipo de beca

El análisis realizado ayudó a identificar las tablas y los campos necesarios, para poder implementar un modelo de análisis para la empresa. Nuestro propósito, es incorporar en el Almacén de Datos todos los datos existentes, con el fin de poder agruparlos y mostrar las instituciones en forma grupal, y de esta manera tener una visión más global de la calidad de la educación, con base en un grupo de instituciones. En la siguiente sección se identifica, a detalle, los datos encontrados durante la etapa de análisis de consultas.

4.2.2 Descripción de datos

Informe general de descripción de datos

A continuación, la tabla 16 presenta los datos identificados con base en el análisis de consultas de la aplicación, y las consultas realizadas a las tablas del sistema:

Tabla 16: Descripción de datos de la información individual del estudiante.

Información individual del estudiante	Fuente de Datos (Tabla)	Nombre Columna	Tipo dato
Adecuación curricular	TEstudiante	adecuación	int
Asistencia	TAsistenciaDiaria, TAsistenciaLección	sePresento	int
Casos de ausentismo	THojaDatos	casoAusentismo	int
Deserción	TTraslado	Tipo	int
Edad	TEstudiante	fechaNacimiento	datetime
Prioridad en el comedor escolar	THojaDatos	asisteComedor	int
Recibe beca	THojaDatos	poseeBeca	int
Recibe bono escolar	THojaDatos	poseeBono	int
Recibe transporte	THojaDatos	recibeTransporte	int
Repitencia	TMatrícula	esRepitente	int
Sexo	TPersona	Sexo	int
Tipos de becas	TBeca	código	int
Trabaja el estudiante	THojaDatos	tipoTrabajo	int

Tabla 17: Descripción de datos de la información familiar del estudiante.

Información familiar del estudiante	Fuente de Datos (Tabla)	Nombre Columna	Tipo dato
Acceso a Internet	THojaDatos	poseeInternet	int
Encargados que viven con el estudiante	TEncargado	viveConEstudiante	int
Escolaridad de la madre	TEncargado	escolaridad/parentesco	int
Escolaridad del padre	TEncargado	escolaridad/parentesco	int
Hogar uniparental	TEncargado	estadoCivil	int
Ingreso mensual per cápita	TEncargado	ingresoMensual	float
Tipo de vivienda	TNucleoFamiliar	estadoCasa	int

Tabla 18: Descripción de datos de las tablas de casos.

Tablas de Casos	Descripción de datos
TAsignatura	Información relacionada con las materias.
TCursoLectivo	Muestras los años lectivos iniciando en el 2000.
TEncargado	Información del encargado, identificación, trabajo, estado civil, etc.
TEstudiante	Información del estudiante identificación, edad, adecuación, etc.
TExpediente	Registro de la identificación y su imagen.
TGrupo	Información general de los grupos, tal como cupo, estado, etc.
TGrupoEstudiante	Registro del estudiante a los grupos.
THojaDatos	Información detallada de los estudiantes.
TPeriodo	Describe todos los periodos de lo centros educativos, por ejemplo primero, segundo, etc.
TPersona	Describe en general las personas de los sistemas, incluye estudiantes y profesores.
TRendimiento	Registro de notas.

Los datos, antes expuestos, son registrados por la aplicación transaccional. Para poder utilizar estos datos, se emplean los “Integration Services” de SQL Server 2008 R2. Existen tablas, como por ejemplo: TRendimiento, las cuales contienen alrededor de doscientos mil registros, para varios años lectivos; lo que indica que la carga de información al Almacén de Datos se va a tornar pesada, por lo que se debe tener en cuenta la aplicación de comandos, que provea la herramienta utilizada, y permitan optimizar las inserciones.

En la exploración de los datos se encontraron muchos valores nulos, o que carecen del formato esperado, por ejemplo las cédulas de identidad. Para estos casos, se han definido transformaciones que conviertan estos datos en algo significativo, para el negocio. La ventaja que existe en utilizar los datos del reporte de variables múltiples, es que estos ya han sido validados por los expertos del negocio. Por lo tanto, el reto está en recrear dicho reporte, en el Almacén de Datos. Otra ventaja presente, es que la fuente de datos es claramente definida, es decir que provienen de los centros educativos que utilizan el sistema transaccional.

4.2.3 Exploración de datos

Se lleva a cabo la exploración de datos, con la finalidad de determinar cantidades y requerimientos de limpieza, transformación y creación de datos.

Haciendo revisión del reporte de variables múltiples, se ha determinado la necesidad de contar con los siguientes datos:

a. Datos del centro educativo

Para obtener los datos del centro educativo, se debe revisar la tabla *tparametro,* cuando el campo nombre sea *codigoCentro* y unirlo al campo de *codigoPresupuesto* de la tabla *tcentroEducativo*, que se encuentra Base de Datos BdPiadLinea.

```
SELECT BdPiadLinea.dbo.tcentroeducativo.codigoPresupuestario,
       BdPiadLinea.dbo.tcentroeducativo.nombre   AS CentroEducativo,
       BdPiadLinea.dbo.tcircuito.numero          AS Circuito,
       BdPiadLinea.dbo.tdireccionregional.nombre AS DireccionRegional,
       BdPiadLinea.dbo.tprovincia.nombre         AS Provincia,
       BdPiadLinea.dbo.tcanton.nombre            AS Cantón,
       BdPiadLinea.dbo.tdistrito.nombre          AS Distrito
FROM   BdPiadLinea.dbo.tcentroeducativo
       INNER JOIN BdPiadLinea.dbo.tcircuito
               ON BdPiadLinea.dbo.tcentroeducativo.codCircuito =
                  BdPiadLinea.dbo.tcircuito.codigo
       INNER JOIN BdPiadLinea.dbo.tdireccionregional
               ON BdPiadLinea.dbo.tcircuito.codDireccion =
                  BdPiadLinea.dbo.tdireccionregional.codigo
       INNER JOIN BdPiadLinea.dbo.tdistrito
               ON BdPiadLinea.dbo.tcentroeducativo.codDistrito =
                  BdPiadLinea.dbo.tdistrito.codigo
       INNER JOIN BdPiadLinea.dbo.tcanton
               ON BdPiadLinea.dbo.tdistrito.codCanton =
                  BdPiadLinea.dbo.tcanton.codigo
       INNER JOIN BdPiadLinea.dbo.tprovincia
               ON BdPiadLinea.dbo.tcanton.codProvincia =
                  BdPiadLinea.dbo.tprovincia.codigo
WHERE  codigoPresupuestario = (SELECT valor
                               FROM   dbo.TParametro
                               WHERE  nombre = 'codigoCentro')
```

Tabla 19: Exploración de datos de centros educativos.

Código	Código Presupuestario	Nombre	Circuito	Dirección Regional	Provincia	Cantón	Distrito
1	3988	LICEO DE SAN ANTONIO	1	Desamparados	San José	Desamparados	San Antonio

b. Ausentismo

En la exploración de datos se ha encontrado que:

- Se deben estipular los valores del campo *tipo*, de la tabla *tasistencialeccion*.
- Formato de la cédula no es consistente.
- La cantidad de lecciones a la que se ausentó aparece en cero (0).

```
SELECT dbo.tasistenciadiaria.fecha,
       dbo.testudiante.cedula,
       dbo.tasignatura.nombre          AS Asignatura,
       dbo.tperiodo.Nombre             AS Periodo,
       dbo.tasistencialeccion.tipo     TipoAsistencia,
       dbo.tasistencialeccion.cantidad AS LeccionesAusente
FROM   dbo.testudiante
       INNER JOIN dbo.tasistenciadiaria
               ON dbo.testudiante.cedula = dbo.tasistenciadiaria.cedula
       INNER JOIN dbo.tasistencialeccion
               ON dbo.tasistenciadiaria.codigo =
                  dbo.tasistencialeccion.codAsistenciaDiaria
       INNER JOIN dbo.tasignatura
               ON dbo.tasistencialeccion.codAsignatura = dbo.tasignatura.codigo
       INNER JOIN dbo.tperiodo
               ON dbo.tasistenciadiaria.codPeriodo = dbo.tperiodo.codigo
WHERE  Datepart(yyyy, dbo.tasistenciadiaria.fecha) = 2012
```

Tabla 20: Exploración de datos de ausentismo

Fecha	Cédula	Asignatura	Periodo	Tipo Asistencia	Lecciones Ausente
26/03/2012		Inglés	Primer período	2	2
26/03/2012		Estudios Sociales	Primer período	2	2
26/03/2012		Inglés	Primer período	0	1
27/03/2012		Conducta	Primer período	0	1
27/03/2012		Conducta	Primer período	3	1
16/07/2012		Física	Segundo período	4	0
16/07/2012		Psicología	Segundo período	4	0
16/07/2012		Educación Cívica	Segundo período	4	0
16/07/2012		Física	Segundo período	4	0
27/03/2012		Francés	Primer período	3	2
27/03/2012	155811032415	Informática Educativa	Primer período	2	2
27/03/2012	1-1672-0078	Conducta	Primer período	2	1

c. Becas por Estudiante

La cédula del estudiante presenta formato incorrecto.

```
SELECT dbo.tcursolectivo.anno,
       dbo.tbeca.nombre,
       dbo.testudiantebeca.cedEstudiante Estudiate
FROM   dbo.tbeca
       INNER JOIN dbo.testudiantebeca
               ON dbo.tbeca.codigo = dbo.testudiantebeca.codBeca
       INNER JOIN dbo.testudiante
               ON dbo.testudiantebeca.cedEstudiante = dbo.testudiante.cedula
       INNER JOIN dbo.tcursolectivo
               ON dbo.testudiantebeca.codCursoLectivo = dbo.tcursolectivo.codigo
                  AND dbo.tcursolectivo.codigo = 13
GROUP  BY dbo.tcursolectivo.anno,
          dbo.tbeca.nombre
```

Tabla 21: Exploración de datos de becas por estudiante.

Curso Lectivo	Beca
2012	Avancemos
2012	Avancemos
2012	Avancemos
2012	Avancemos

d. Grupos de Estudiantes

- Se debe determinar el valor del campo estado, de la tabla *testudiante*.
- Cédulas inconsistentes.

```
SELECT DISTINCT dbo.tcursolectivo.anno AS CursoLectivo,
                dbo.tnivel.nombre      AS Nivel,
                dbo.tgrupo.numero      AS Grupo
                dbo.testudiante.cedula,
                dbo.tpersona.primerApellido,
                dbo.tpersona.segundoApellido,
                dbo.tpersona.nombre,
                dbo.testudiante.estado,
                dbo.testudiante.adecuacion
FROM   dbo.tgrupo
       INNER JOIN dbo.tgrupoestudiante
               ON dbo.tgrupo.codigo = dbo.tgrupoestudiante.codigo
       INNER JOIN dbo.testudiante
               ON dbo.tgrupoestudiante.cedula = dbo.testudiante.cedula
       INNER JOIN dbo.tpersona
               ON dbo.testudiante.cedula = dbo.tpersona.cedula
       INNER JOIN dbo.tcursolectivo
               ON dbo.tgrupo.codCursoLectivo = dbo.tcursolectivo.codigo
       INNER JOIN dbo.tnivel
               ON dbo.tgrupo.numNivel = dbo.tnivel.numero
       INNER JOIN dbo.tasignatura
               ON dbo.tgrupoestudiante.codAsignatura = dbo.tasignatura.codigo
WHERE  dbo.tcursolectivo.anno = 2012
ORDER  BY dbo.tcursolectivo.anno,
          dbo.tnivel.nombre,
          dbo.tgrupo.numero
```

Tabla 22: Exploración de datos de grupos de estudiantes.

Curso Lectivo	Nivel	Grupo	Estado	Adecuación
2012	Décimo	1	3	0
2012	Décimo	1	3	0
2012	Décimo	1	3	0
2012	Décimo	1	3	0
2012	Décimo	2	3	1
2012	Décimo	3	3	1
2012	Décimo	3	3	0
2012	Décimo	3	5	0
2012	Décimo	3	3	1
2012	Décimo	3	3	0

e. Traslado Estudiante

- Campos que no contienen información como: CodigoCentroDestino y NombreCentroDestino.
- El tipo de traslado presenta la descripción del valor, sólo en la aplicación.

```
SELECT dbo.ttraslado.fecha,
       dbo.tpersona.cedula,
       dbo.tpersona.primerApellido,
       dbo.tpersona.segundoApellido,
       dbo.tpersona.nombre,
       dbo.testudiante.estado      AS EstadoEstudiante,
       dbo.tcentroeducativo.nombre AS CentroEducativoOrigen,
       dbo.ttraslado.codigoEscuelaDestino,
       dbo.ttraslado.escuelaDestino,
       CASE
         WHEN dbo.ttraslado.tipo = 0 THEN 'Inclusión'
         WHEN dbo.ttraslado.tipo = 1 THEN 'Exclusión'
         WHEN dbo.ttraslado.tipo = 2 THEN 'Deserción'
         WHEN dbo.ttraslado.tipo = 3 THEN 'Prematricula'
       END                         AS TipoTraslado
FROM   dbo.ttraslado
       INNER JOIN dbo.texpedienteest
               ON dbo.ttraslado.codigo = dbo.texpedienteest.codigo
       INNER JOIN dbo.testudiante
               ON dbo.ttraslado.cedula = dbo.testudiante.cedula
       INNER JOIN dbo.tpersona
               ON dbo.testudiante.cedula = dbo.tpersona.cedula
       INNER JOIN dbo.tcentroeducativo
               ON dbo.ttraslado.codigo = dbo.tcentroeducativo.codigo
```

Tabla 23: Exploración de datos de traslado de estudiantes.

Estado Estudiante	CentroEducativoOrigen	CodigoCentroDestino	NombreCentroDestino	TipoTraslado
2	Escuela Elías Jiménez Castro		LSA	Inclusión
3	Escuela Francisco Gamboa			Deserción
2	Escuela García Monge			Deserción
2	Escuelas Privadas			Deserción
2	Escuela Francisco Schimith			Deserción
2	Escuela Gravilias			Deserción
2	Escuela Fátima			Deserción
2	Otras Fuera de Circuito 01			Deserción
2	Otras del circuito 01			Deserción

f. Deserciones

- Para saber si un estudiante ha desertado del centro educativo, se debe buscar en la tabla dbo.ttraslado el campo tipo cuando el valor sea 2.
- Se debe validar que la deserción puede ser para un curso lectivo en específico.

```
SELECT dbo.ttraslado.fecha,
       dbo.tpersona.cedula,
       dbo.tpersona.primerApellido,
       dbo.tpersona.segundoApellido,
       dbo.tpersona.nombre,
       dbo.testudiante.estado AS EstadoEstudiante
FROM   dbo.ttraslado
       INNER JOIN dbo.texpedienteest
               ON dbo.ttraslado.codigo = dbo.texpedienteest.codigo
       INNER JOIN dbo.testudiante
               ON dbo.ttraslado.cedula = dbo.testudiante.cedula
       INNER JOIN dbo.tpersona
               ON dbo.testudiante.cedula = dbo.tpersona.cedula
       INNER JOIN dbo.tcentroeducativo
               ON dbo.ttraslado.codigo = dbo.tcentroeducativo.codigo
WHERE  dbo.ttraslado.tipo = 2
```

Tabla 24: Exploración de datos de deserciones.

Fecha	Cédula	PrimerApellido	SegundoApellido	Nombre	Estado Estudiante
22/08/2008	1-1531-0326	BRENES	ARCE	LUIS DIEGO	3
22/08/2008	1-1548-0978	CHAVARRIA	JIMENEZ	ANGIE	2
22/08/2008	1-15460-0592	GARCIA	QUIROS	JOSE GUILERMO	2
22/08/2008	1-1513-0954	RODRIGUEZ	FERNANDEZ	BRYAN ANDREY	2
22/08/2008	1-1541-0778	SEGURA	TELLERIA	RAQUEL	2
22/08/2008	1-1470-0804	UREÑA	ROJAS	OSCAR	2
22/08/2008	1-1521-0632	FALLAS	ARIAS	ALEXANDER	2
25/08/2008	3-0464-0936	BRENES	CHAVEZ	DANIELA	2

g. Rendimiento Académico

- Se visualizaron promedios en 0, se debe averiguar cuál es la nota mínima.
- Cédulas inconsistentes, y que no cumplen el formato.

```
SELECT dbo.tcursolectivo.anno  AS CursoLectivo,
       dbo.tperiodo.nombre     AS Periodo,
       dbo.tasignatura.nombre  AS Asignatura,
       dbo.trendimiento.cedula AS Cedula,
       dbo.trendimiento.nota   AS Nota
FROM   dbo.trendimiento
       INNER JOIN dbo.tasignatura
               ON dbo.trendimiento.codAsignatura = dbo.tasignatura.codigo
       INNER JOIN dbo.tperiodo
               ON dbo.trendimiento.codPeriodo = dbo.tperiodo.codigo
       INNER JOIN dbo.tcursolectivo
               ON dbo.tperiodo.codCursoLectivo = dbo.tcursolectivo.codigo
WHERE  dbo.tcursolectivo.anno = 2012
```

Tabla 25: Exploración de datos de rendimiento académico.

CursoLectivo	Periodo	Asignatura	Cédula	Promedio
2012	Primer período	Educación Física	01-1626_-0006____	0.00
2012	Primer período	Educación Física	115830406	62.00
2012	Primer período	Educación Física	_-____-____	0.00

h. Beneficios Económicos

- Los datos no están siendo actualizados. Se hallaron estudiantes en la tabla testudianteBeca, que en la tabla thojadatos el campo poseeBeca es igual a 0; asumimos que 0 es igual a No y 1 es igual a Sí.
- Se encuentran cédulas que no cumplen el formato, y son inconsistentes.

```
SELECT dbo.tcursolectivo.anno AS CursoLectivo,
       dbo.testudiante.cedula,
       CASE WHEN dbo.thojadatos.poseeBeca        = 1 THEN 'SI'
            ELSE 'NO' END     AS PoseeBeca,
       CASE WHEN dbo.thojadatos.poseeBono        = 1 THEN 'SI'
            ELSE 'NO' END     AS poseeBono,
       CASE WHEN dbo.thojadatos.asisteComedor    = 1 THEN 'SI'
            ELSE 'NO' END     AS asisteComedor,
       CASE WHEN dbo.thojadatos.recibeTransporte = 1 THEN 'SI'
            ELSE 'NO' END     AS recibeTransporte
FROM   dbo.thojadatos
       INNER JOIN dbo.tcursolectivo
               ON dbo.thojadatos.codCursoLectivo = dbo.tcursolectivo.codigo
       INNER JOIN dbo.texpedienteest
               ON dbo.thojadatos.codExpedienteEst = dbo.texpedienteest.codigo
       INNER JOIN dbo.texpediente
               ON dbo.texpedienteest.codigo = dbo.texpediente.codigo
       INNER JOIN dbo.testudiante
               ON dbo.texpediente.cedula = dbo.testudiante.cedula
WHERE  dbo.tcursolectivo.anno = 2012
```

Tabla 26: Exploración de datos de beneficios económicos.

CursoLectivo	Cedula	RecibeBeca	RecibeBono	AsisteComedor	RecibeTransporte
2012	10400470104	NO	NO	NO	NO
2012	0-1096-583_	NO	NO	NO	NO
2012	01-1626_-0006____	NO	NO	NO	NO
2012	40520120275	NO	NO	NO	NO
2012	1002153245	NO	NO	NO	NO

i. Encargado del Estudiante

- Se debe asignar los valores de parentesco, ya que se encuentran fijos en la aplicación.
- El formato de la cédula es incoherente; tanto para el estudiante, como para el encargado.

```
SELECT dbo.testudiante.cedula AS CedulaEstudiante,
       dbo.tencargado.cedula  AS CedulaEncargado,
       CASE
         WHEN dbo.tencargado.viveConEstudiante = 1 THEN 'SI'
         ELSE 'NO' END       AS viveConEstudiante,
       dbo.tencargado.parentesco
FROM   dbo.testudiante
       INNER JOIN dbo.testudianteencargado
               ON dbo.testudiante.cedula =
                  dbo.testudianteencargado.cedEstudiante
       INNER JOIN dbo.tencargado
               ON dbo.testudianteencargado.cedEncargado = dbo.tencargado.cedula
```

Tabla 27: Exploración de datos del encargado del estudiante.

Cédula	Parentesco
0	1
135-RE-017233	0
_-____-____	0
_C1780726-____-____	0
6780	1
6780	7
6780	0
8-3006-4___	0
8888060088	7
c0830064	0
10400470104	0
160400094936	1

j. Datos personales del estudiante

La exploración de datos muestra fechas de nacimiento con error, y los valores del Estado Civil se encuentran en el código de la aplicación.

- Fecha de nacimiento presenta error, por lo que al calcular la edad se obtienen valores o muy bajos, o muy altos.
- Se deben buscar los valores del Estado Civil, porque se encuentran fijos en la aplicación.
- Cédulas que no cumplen el formato, y son inconsistentes.

```
SELECT dbo.tpersona.cedula,
       dbo.tpersona.primerApellido,
       dbo.tpersona.segundoApellido,
       dbo.tpersona.nombre,
       dbo.tnacionalidad.nombre                                    AS Nacionalidad,
       CASE WHEN dbo.tpersona.sexo = 0 THEN 'FEMENINO'
         ELSE 'MASCULINO'  END                                     AS Sexo,
       dbo.testudiante.fechaNacimiento,
       Datediff(yy, dbo.testudiante.fechaNacimiento, Getdate()) AS Edad,
       dbo.testudiante.estadoCivil
FROM   dbo.tpersona
       INNER JOIN dbo.testudiante
               ON dbo.tpersona.cedula = dbo.testudiante.cedula
       INNER JOIN dbo.tnacionalidad
               ON dbo.tpersona.codNacionalidad = dbo.tnacionalidad.codigo
```

Tabla 28: Exploración de datos personales del estudiante.

Valor Cédula	FechaNacimiento	Edad	Estado Civil
115020087	27/02/2012	1	0
_-____-____	15/02/2013	0	0
_C1780726-____-____	0001-01-01:00	2012	0
6780	0855-11-16:00	1158	0
7100460	1897-03-04:00	116	0
0-0710-0460	18/12/8989	-6976	0
8888060088			

4.3 Fase 3 - Preparación de los datos

La preparación de los datos es una de las fases determinantes, en el éxito de la creación del Almacén de Datos. Se identifica el conjunto de datos que se usará para el trabajo de análisis principal, y que estarán basados en el conjunto de datos desplegados por el reporte de variables múltiples.

4.3.1 Datos seleccionados

La selección de los datos se basó en una exploración de todas las tablas utilizadas, en la elaboración del reporte de Variables Múltiples, del sistema PIAD en Línea, como muestra el diagrama de la figura 22. Esta labor se llevó a cabo, recorriendo las porciones de código ejecutadas por la aplicación, y anotando cada sentencia de

Transact-SQL utilizadas para consultar y valorar los datos. El reporte de variables múltiples no utiliza un procedimiento almacenado de la base de datos, sino que se construye dinámicamente desde la aplicación.

Figura 22 Entidades que alimentan el Almacén de Datos.

4.3.2 Limpieza de datos

Se descartan los estudiantes, cuyo formato de cédula no cumpla con las siguientes especificaciones:

- Tener una longitud de 11 caracteres.
- El carácter 2 y 7 deben corresponder a un guion (-).
- El primer carácter debe ser un número del uno (1) al nueve (9).
- El último carácter debe ser un número del cero (0) al nueve (9).

Se excluyen los registros repetidos en varias de la tablas, por utilizar llaves primarias subrogadas, se anotan registros con el resto de la información repetida, por lo cual a la hora de utilizar la información, se decidió agruparla por el resto de los campos para eliminar los repetidos.

4.3.3 Construcción de datos

Edad del Estudiante: La edad del estudiante se construye tomando el campo FechaNacimiento de Testudiante, y se obtiene la diferencia con la fecha actual de la base de datos (Getdate) en años.

HogarUniparental: Por cada estudiante, se seleccionan aquellos encargados donde el parentesco sea padre o madre, y que viven con él.

4.3.4 Integración de datos

EscolaridadPadre: se selecciona la escolaridad que posee el encargado, cuyo parentesco es padre.

EscolaridadMadre: se selecciona la escolaridad que posee el encargado, cuyo parentesco es madre.

4.3.5 Formateo de datos

Transformar formateando se refiere, principalmente, a modificaciones sintácticas hechas a los datos que no cambian su significado, pero podría ser requerido por la herramienta de modelado.

Como la dimensión de sexo empieza el Id con 1, y la tabla de persona guarda 0 para el sexo Femenino, y 1 para el sexo Masculino; se toma el campo Sexo y se le suma uno. Este proceso se aplica de igual manera a la adecuación.

4.4 Fase 4 - Modelado

El modelado del Almacén de Datos, se ha llevado a cabo bajo la filosofía de Kimball, haciendo uso de los elementos básicos: los sistemas operacionales (son las distintas bases de datos que contienen la información de los colegios académicos, que utilizan el sistema Piad en Línea); el área de transformación (esquema llamado *TransformacionDatos*), donde se realizan las operaciones necesarias, para transformar los datos; el área de extracción de datos (esquema llamado *ExtraccionDatos*) donde se copiarán los datos sin cambios de las fuentes de información originales; y el área de presentación (esquema que se ha llamado *SoporteDecisiones*, que contendrá las tablas del esquema estrella).

La construcción está basada en la selección del proceso de negocio, definido por la información provista en el reporte de variables múltiples: la granularidad llevada hasta la cantidad de estudiantes; la elección de las dimensiones, tomada de los distintos filtros de información del reporte mencionado; la identificación de hechos, basada en los indicadores de cantidad de estudiantes (el ingreso per cápita familiar promedio, el rendimiento determinado por las notas de las asignaturas de los estudiantes, y el ausentismo basado en las lecciones a las que se ausenta, con o sin justificación)

Además, de que los proceso recomendados por Kimball, se encuentran integrados en la metodología CRISP, permitiendo complementar la parte innovadora de este proyecto.

4.4.1 Generar el diseño de prueba

Antes de construir un modelo, es necesario definir un procedimiento para probar la calidad del modelo y la validez. Por ejemplo, en tareas de minería de datos supervisadas, como la clasificación, es común usar tasas de error como medidas de calidad. Por lo tanto, el diseño de prueba específica, que el conjunto de datos debería ser separado en el entrenamiento, y en el conjunto de prueba. El modelo está construido sobre el conjunto de entrenamiento, y su calidad estimada sobre el conjunto de prueba.

4.4.2 Construcción del modelo

Arquitectura del ETC.

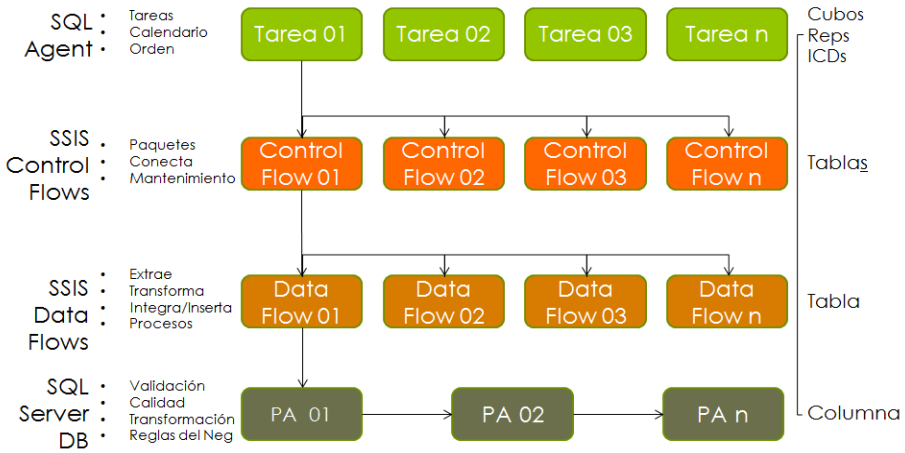


Figura 23 Arquitectura del Almacén de Datos Piad para el proceso de ETC (Extracción, transformación y cargado)

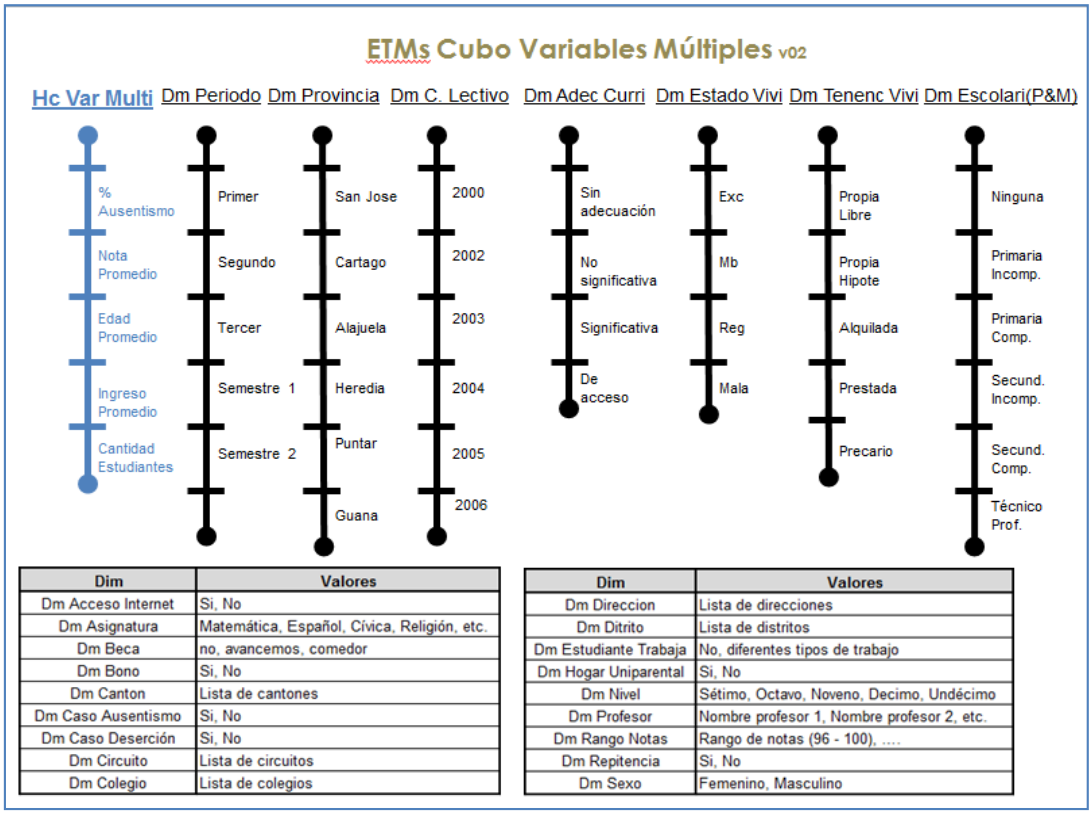
La arquitectura mostrada en la figura 23, permite modelar un Almacén de Datos escalable y robusto. Esta arquitectura servirá de guía, para que nuevas etapas del Almacén de Datos se lleven a cabo, en forma estandarizada y ordenada.

Estructuras de Tipo Multidimensional (ETMs)

Se utilizaron las estructuras de tipo multidimensional, para mostrar al cliente la información que se podría desplegar desde el modelo multidimensional.

Como muestra la figura 24, todos los ETMs (como por ejemplo DM Periodo, Dm Provincia, Dm Curso Lectivo, entre otras) de color negro son las dimensiones de la

estrella (incluyendo las tablas de dimensiones mostradas), y serán los datos por los que se pueda agrupar y filtrar la información, que se desprende del ETM Hc Variables Múltiples (color azul, el cual representa los hechos). No se despliegan todas las barras de dimensiones, para poder tener una imagen más clara. Se opta por documentar la totalidad del modelo, por medio de ETMs y tablas descriptivas de nombre de dimensiones, y sus correspondientes valores.



Dim	Valores
Dm Acceso Internet	Si, No
Dm Asignatura	Matemática, Español, Cívica, Religión, etc.
Dm Beca	no, avancemos, comedor
Dm Bono	Si, No
Dm Canton	Lista de cantones
Dm Caso Ausentismo	Si, No
Dm Caso Deserción	Si, No
Dm Circuito	Lista de circuitos
Dm Colegio	Lista de colegios

Dim	Valores
Dm Direccion	Lista de direcciones
Dm Ditrito	Lista de distritos
Dm Estudiante Trabaja	No, diferentes tipos de trabajo
Dm Hogar Uniparental	Si, No
Dm Nivel	Sétimo, Octavo, Noveno, Decimo, Undécimo
Dm Profesor	Nombre profesor 1, Nombre profesor 2, etc.
Dm Rango Notas	Rango de notas (96 - 100), ...
Dm Repitencia	Si, No
Dm Sexo	Femenino, Masculino

Figura 24 ETMs del cubo de variables múltiples

Diseño de la estrella.

Según muestra la figura 25, se elige un modelo en estrella para el Almacén de Datos, cuya simpleza, al no mantener relaciones de las dimensiones con otras tablas, le hace más eficiente en las consultas.

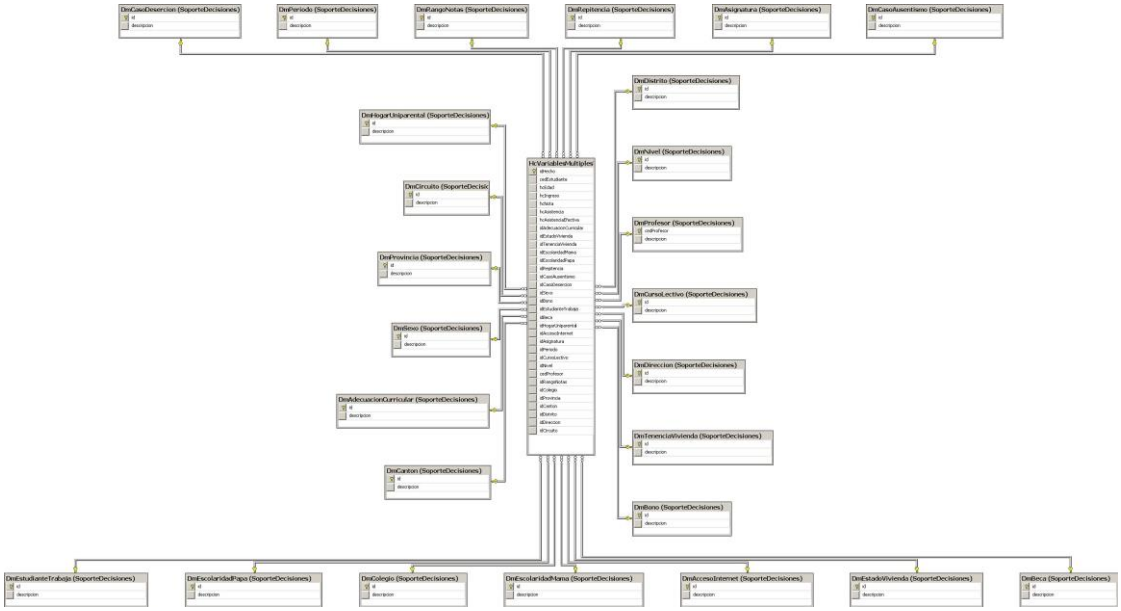


Figura 25 Diagrama estrella del cubo variables múltiples

Despliegue de información

Para el despliegue de información se puede utilizar una herramienta de generación de reportes, o una herramienta de despliegue de información. Para efectos de este proyecto, se hará uso de la herramienta Microsoft Excel.

Implementación de Indicadores clave de desempeño (o ICD)

El término en inglés se conoce como KPI o "Key Performance Indicators". Los servicios de análisis del SQL Server 2008, brinda la posibilidad de agregar ICDs a las bases de datos, con el fin de evaluar el desempeño de los negocios, estos indicadores se reflejarían en los datos del modelo multidimensional. Un ICD está relacionado a los hechos de un modelo multidimensional, y a los cálculos que se puedan llevar a cabo con estos hechos. Estos cálculos se realizan por medio de un lenguaje llamado: *declaraciones de expresiones multidimensionales* (o MDX, en inglés sería "Multidimensional Expressions (MDX) statements").

Para que un ICD pueda evaluar el desempeño de un negocio, se debe de tomar en cuenta las siguientes cuatro características:

- El valor en sí del ICD, por ejemplo, nota promedio, el cual se calcula como el total de notas entre la cantidad de estudiantes, por un conjunto dado de características.
- El valor meta del ICD, este se define de acuerdo a las necesidades del negocio.

- El estado del ICD. El negocio debe de definir cuando la meta es alcanzada satisfactoriamente, cuando se está alejando los resultados de esta, y finalmente cuando el valor obtenido está totalmente desviado de la meta definida.
- La tendencia del ICD con respecto al tiempo. El negocio define, con respecto al tiempo, como se debe de comportar los valores obtenidos.

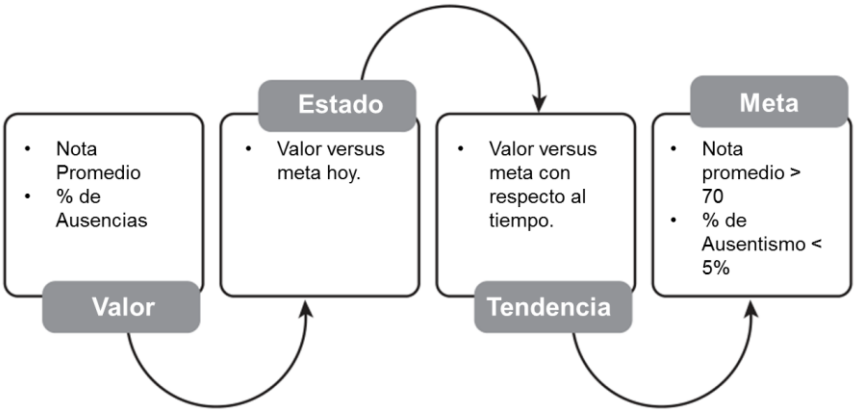


Figura 26 Definición de un ICD para el modelo multidimensional variables múltiples.

El diagrama mostrado en la figura 26, describe las cuatros características que se pueden definir para un ICD, para el modelo multidimensional de variables múltiples del Almacén de Datos PIAD. El diagrama describe la relación entre las partes, y los posibles valores para cada una de las características del ICD (nota promedio y % de ausencias).

Los ICD's de nota promedio y porcentaje de ausencias, son de gran ayuda para que el negocio pueda monitorear el desempeño de la educación en los centros educativos, que utilizan el sistema PIAD en línea. Los ICDs pueden variar su valor y sus estados dependiendo, de las variables que el usuario seleccione.

Colegio	LICEO DE SAN ANTONIO				
	Values				
Row Labels	icd_Ausentismo	icd_Ausentismo Status	icd_Promedio_Notas	icd_Promedio_Notas Status	Hc Conteo Estudiantes
Biología	3.70%		71.64		5,941.00
Ciencias	2.44%		70.14		14,077.00
Educación Cívica	4.38%		72.49		19,538.00
Español	2.50%		68.19		20,077.00
Estudios Sociales	2.71%		69.24		20,196.00
Física	3.73%		73.24		5,862.00
Francés	3.36%		72.20		14,550.00
Inglés	3.26%		74.50		18,849.00
Matemática	2.31%		66.39		20,620.00
Química	3.69%		69.04		6,066.00
Grand Total	**2.94%**		**70.44**		**145,776.00**

Figura 27 Ejemplo de ICDs para el Almacén de Datos piad

La figura 27, describe como los ICDs, ausentismo y promedio de notas, muestran sus valores y estado, con respecto a las variables colegio y asignatura. En el caso del ICD ausentismo, su meta es ser menor de 5%. En el caso del colegio Liceo de San Antonio, esta meta se cumple por materia, por lo que su indicador de estado se encuentra en color verde. En el caso del ICD promedio de notas, para este mismo colegio se define la meta como promedio de notas mayor a 70. En caso de que, el promedio de notas se encuentren entre 60 y 70, se considera una advertencia de que se está alejando de la meta, y su estado será de color amarillo. Con base en lo antes mencionado, por ejemplo para la materia español, el promedio de notas está en estado de advertencia, ya que este tiene un promedio 68.19, el cual es menor a la meta definida.

4.5 Fase 5 Evaluación

La evaluación de los datos es un proceso clave para el éxito de la creación del Almacén de Datos. Esta evaluación ocurre en todo el proceso de ETC (o proceso de extracción, transformación y cargado de datos).

En el proceso de evaluación se realizan actividades como:

- Validación de cantidades.
- Validación de tipos de datos.
- Validación de sus clasificaciones.
- Consolidación de datos.

4.5.1 Evaluación del cumplimiento de los objetivos del negocio y los criterios de éxito establecidos.

Objetivos del negocio

Centralizar la información para toma de decisiones: se ha centralizado la información de diez colegios académicos que utilizan el sistema PIAD en línea, para permitir el análisis de la información en forma conjunta, para la producción de los indicadores de la calidad.

Permitir la mezcla de variables para generar información: Se ha implementado 26 variables, de las 27 contenidas en el Reporte de Variables Múltiples, logrando implementar en un 96% la información requerida. Se espera, en la próxima versión del Almacén de Datos, incluir la dimensión del grupo al que pertenece el estudiante,

mediante la agregación de las tablas: tsubgrupo, tsubcargaacademica y tsubgrupoestudiante. Se incluyeron nuevas variables, tales como rango de notas, Dirección Regional y circuito.

Criterios de éxito

Según los criterios de éxitos, definidos en la primera fase del desarrollo del Almacén de Datos, se ha evaluado el cumplimiento de los mismos para determinar el éxito del proyecto.

CRE - 001 - Tiempos de respuesta al usuario: El proceso de generar el cubo de información, involucra la carga y transformación de datos. Por lo cual, el tiempo requerido para completar el proceso va a variar según la cantidad de datos a procesar y las características de infraestructura tecnológica en la que se ejecute. Para efectos de este proyecto las pruebas realizadas tuvieron una duración de treinta minutos. Una vez cargado el cubo, la respuesta de provisión de información con la herramienta de Excel es inmediata.

CRE – 002 - Capacidad de mantener información histórica: El Almacén de Datos PIAD es capaz de registrar la información histórica de rendimiento, asistencia, deserción entre otros valores para ser analizadas y determinar patrones de comportamiento.

CRE -003 - Generar reportes que permitan el uso de parámetros para su generación, por parte del usuario: Utilizando como visualizador de dato la herramienta Excel, se provee al usuario de una capacidad ilimitada, para generar reportes y gráficos, según sus necesidades. También, se puede utilizar Reporting Service para la generación de reportes que provean información a los usuarios, según sus roles y permisos.

CRE -004 - Diseño del Almacén de Datos: El Almacén de Datos incluye veintiséis, de las veintisiete variables del reporte de variables múltiples del sistema PIAD, queda pendiente incluir la dimensión del grupo al que pertenece el estudiante, mediante la agregación de las tablas: tsubgrupo, tsubcargaacademica y tsubgrupoestudiante, en la próxima versión del Almacén de Datos.

CRE-005 - Seguridad de la información: El Almacén de Datos PIAD cuenta con la seguridad que brinda la herramienta MS SQL Server 2008 R2, para la definición de perfiles y permisos de autorización sobre la información a desplegar. También se puede

hacer uso de la herramienta de Reporting Service, para la generación de reportes específicos, regulando el acceso a los mismos, de acuerdo a las políticas de seguridad que se definan.

El Almacén de Datos PIAD es una herramienta, que viene a proveer al negocio de una forma dinámica de análisis de grandes volúmenes de información, que permiten una toma de decisiones acertada, como la posibilidad de implementar, a futuro proyectos de minería de datos que permita determinar patrones de conducta en los estudiantes, que ayuden a mejorar la calidad de la información, así como las condiciones en que se ofrece, para disminuir el ausentismo y la deserción y mejorar el rendimiento.

4.5.2 Evaluación del proceso

La evaluación de los resultados ocurre en todo el proceso del ETC, ya que de esta manera se asegura el tener la mejor información posible, para la creación de las herramientas de soporte a la toma de decisiones.

Cuando se está en el proceso de decidir, cuáles fuentes de datos hay que cargar al Almacén de Datos, se debe evaluar el tipo de dato que se piensa utilizar. Por ejemplo, el tamaño de los textos, ya que una mala decisión a la hora de definir un tamaño, puede provocar una pérdida de información. Además, no se puede asumir que todas las fuentes de datos son iguales, hay que realizar validaciones; tanto de base de datos, como de los datos en sí, para estar seguros que estamos tomando la mejor decisión.

Una vez definidos los tipos de datos a considerar, estos se deben de evaluar. En el caso de nuestro Almacén de Datos, se definen una serie de variables, las cuales ayudan al análisis de datos, pero estas deben ser evaluadas, con el fin de asegurar que se están utilizando correctamente; y de la misma forma en todas las fuentes de datos contempladas en el Almacén de Datos.

En caso de los datos que formaran parte de hechos, es decir los datos utilizados para el análisis, es sumamente importante evaluar, cuáles serían los mejores valores por defectos, para cuando no se puede definir o encontrar algún dato en especial. Por ejemplo, es importante definir el valor por defecto, cuando no se pueda definir la nota de un estudiante en algún momento dado (en nuestro caso definimos el valor por defecto igual a cero).

La valoración y evaluación de los identificadores es un paso importante, para asegurar resultados confiables para el Almacén de Datos. En este proyecto, en particular, la evaluación de las cédulas de los estudiantes fue clave, para poder contar con datos fiables. El formato fue vital, para asegurar que no se dé la duplicación de datos en el Almacén de Datos; es algo básico que los datos posean una calidad aceptable. Además, se deben identificar y separar los casos especiales, para que sea el negocio quién determina el tratamiento que deba aplicárseles.

Al efectuar la evaluación de resultados, es importante validar si el alcance del prototipo es significativamente representativo, con respecto al total de fuentes que alimentarían el Almacén de Datos. Para el presente trabajo, el total de centros educativos, incluidos en el prototipo, fue de diez colegios, estos se seleccionaron por ser los que tenían número significativo de datos, y mayor completitud de los mismos. Se escogieron diez colegios (aproximadamente un 25%), del total de bases de datos que utilizan el sistema PIAD en línea. Al ser un número significativo, se enfrentaron problemas muy similares, a los que presenta un Almacén de Datos en producción. La mayoría de dificultades encontradas, se relacionan con la calidad de datos, y el uso no estandarizado de la información.

Un paso clave en la evaluación de los datos, es donde se lleva a cabo. Inicialmente, se llevó a cabo una exploración de los datos en los repositorios individuales, de cada colegio. Sin embargo, fue cuando se unificaron en la extracción, cuando los problemas grandes surgieron. El tener la información de los diez colegios en un único repositorio evidenció, inconvenientes del uso de los datos, de calidad y falta de estandarización al utilizar los términos. Por ejemplo, los códigos de los periodos lectivos y de las asignaturas eran diferentes, para cada centro educativo.

Según lo mencionado en el marco teórico, en un modelo multidimensional existen tablas de hechos, y tablas de dimensiones. Al intersecar las dimensiones, se logra identificar hechos que pueden ayudar a analizar situaciones. Un ejemplo, es la identificación de las ventas (en este caso es el hecho), para un producto en un mes específico (estas son las dimensiones). En el presente proyecto, se construyeron veintiséis dimensiones y ocho hechos. Es importante, que antes de generar el cubo se evalué que las dimensiones se están utilizando correctamente, y que las consultas a la tabla de hechos, devuelven resultados confiables y válidos.

La evaluación de los datos se realizó por etapas:

a. Se comprobó que los datos que se están generando son correctos, para lo que se utiliza la ejecución de consultas clave de prueba, dentro del proyecto del cubo, utilizando la herramienta de Visual Studio.

b. Una vez identificado el dato clave, se debe hacer la misma consulta en el modelo multidimensional, generado en el motor de base de datos en SQL Server.

c. Si la anterior comparación es satisfactoria, se valida el resultado con la aplicación cliente, que se conecta a la base de datos multidimensional. Para efectos de este proyecto, las pruebas se realizaron con el cliente de Excel.

En el momento que se estén evaluando los procesos de extracción, transformación y carga de datos, se debe realizar un análisis estratégico, para definir cuánto procesar a la vez. Por ejemplo, evaluar a cada estudiante para saber que nota obtuvo cada por año, periodo, nivel, materia, grupo y colegio, es un proceso pesado, que debe realizarse por separado de la evaluación, de cuántas veces un estudiante ha faltado por las mismas características. Este tipo de separación se define, por medio pruebas con cantidades de datos controladas, y midiendo el impacto en el motor de base de datos por tiempo de procesamiento. Un proceso, con una duración de 5 a 10 minutos, es aceptable para grandes cantidades de datos, pero procesos con una duración de más de 10 minutos, al llevarlos a un ambiente de producción, se convertirán en procesos que tardan horas en llevarse a cabo.

Para evaluar las trasformaciones, es aconsejable tener un lugar temporal, que permita analizar que los resultados fueron los esperados. Además, se aconseja utilizar técnicas de confirmación a base de dato falso (o “commit false”), lo cual ayuda a que se puedan utilizar los datos de prueba varias veces. Para este caso, se definieron tres lugares (o esquemas) en la base de datos: extracción, trasformación y soporte de decisiones. El primero recibe los datos sin cambios de las fuentes; en transformaciones como su nombre lo dice ocurren todos los cambios estratégicos, y finalmente en soporte están los datos listos, para generar el cubo. Únicamente, se aplican los cambios finales a los hechos, cuando la trasformación ha sido validada en el lugar temporal en la sección de trasformaciones.

En la etapa de extracción y consolidación de los datos de las fuentes, se puede dar el caso que se utilicen conceptos en forma irregular. Por ejemplo, representar los periodos en forma diferente en cada centro educativo, es decir que el primer trimestre de un curso

lectivo se represente con identificador diferente, o que un profesor exista en varios colegios con la misma cedula de identidad, pero con nombres distintos. Este tipo de situaciones se presentaron, por lo que hay que diseñar trasformaciones claves, para poder traducir estas irregularidades en un grupo de datos estandarizado. Por medio de la evaluación de los datos consolidados, se puede corregir este tipo de situaciones.

Es importante a la hora de evaluar los datos, entender las reglas del negocio, ya que al entenderlas, se pueden diseñar pruebas las cuales busquen validar estas reglas. Por otro lado, el entender las reglas de negocio, ayuda a realizar un plan efectivo para procesar los datos y evaluarlos. Finalmente, la prueba que siempre hay que realizar, antes de generar un cubo, es probar que todos los datos de las dimensiones correspondan a los hechos, y que no existan valores desconocidos dentro del modelo.

Finalmente, realizadas todas pruebas, antes mencionadas, corresponde al negocio realizar el uso de la herramienta de análisis, y comenzar un proceso de mejora continua con el equipo de desarrollo, ya que los datos son cambiantes, constantemente crecen, y representan el comportamiento los estudiantes. Las necesidades de cambios que se generarán, por parte del negocio, deben de someterse a un proceso de priorización y planeamiento, para que estos se vayan aplicando de acuerdo a un orden basado en las necesidades del negocio.

4.6 Fase 6 Implementación o despliegue

Para efectos de la presente investigación, la herramienta elaborada no será implementada aún, ya que se está en un proceso de unificar las bases de datos, utilizadas por los centros educativos, en un único repositorio, por lo cual el diseño de la base de datos y sus estructuras serán variados.

Con el fin de llevar a cabo la investigación, y completar la Fase 6 de la metodología CRISP, se están implementando los apartados, de esta sección, a manera de recomendación.

4.6.1 Resultados del Almacén de Datos

Este apartado detalla la evaluación de los resultados del desarrollo del Almacén de Datos, y la recomendación de un plan para llevar a cabo la implementación del mismo.

Extracción de Datos

Se ha implementado un Almacén de Datos que inició con una carga de datos, para el análisis, compuesta por 24 tablas, con un total de seis millones quinientos treinta y cuatro mil ochocientos cincuenta y un registros (6.534.851), representados en la tabla 29.

Tabla 29. Procedimientos almacenados de la transformación de datos.

Tabla	Cantidad Registros
tasignatura	353
tasistenciadiaria	1828006
tasistencialeccion	2811293
tbeca	20
TCargaAcademica	19944
tcursolectivo	131
tdocente	920
tencargado	20835
tencargadonucleofamiliar	5452
testudiante	78474
testudiantebeca	4343
testudianteencargado	29349
TExpediente	86020
tgrupo	1503
tgrupoestudiante	122468
thojadatos	16196
tleccionefectiva	3607
tmatricula	47147
tnivel	50
tnucleofamiliar	3441
tperiodo	599
tpersona	51403
trendimiento	1390008
ttraslado	13289
Total	**6534851**

Transformación de Datos: se crearon N Tablas y N procedimientos almacenados, que permitieron la limpieza, transformación y creación de los datos requeridos para el Almacén de Datos, según la información obtenida de las 10 bases de datos, pertenecientes a 10 colegios académicos que utilizan el programa PIAD en línea.

a. Tablas:

- TAsignatura
- TAsignaturaClaseEfectiva
- TCanton
- TCentroEducativo
- TCircuito
- TDireccionRegional
- TDistrito
- TEscolaridadMadre
- TEscolaridadPadre
- TEstudiante
- TEstudianteAusentismo
- TEstudianteBeca
- TEstudianteCasoAusentismo
- TEstudianteDesercion
- TEstudianteIndicador
- TEstudianteNotas
- TEstudianteProfesorGrupo
- TEstudianteRepitente
- TEstudianteUbicacion
- THogarUniparental
- THojaDatosUnica
- TNucleoFamiliar
- TPeriodosAD
- TProvincia
- TSistemaInformacion

b. Procedimientos Almacenados:

Los procedimientos almacenados, programados con la finalidad de aplicar extracción, limpieza, transformación de datos y la carga de los mismos en la estrella, se encuentran representados con su respectiva cantidad de líneas de código en la tabla 12. Se requirió al menos de diez mil líneas de códigos en total, entre consultas y pruebas de código,

para llegar a la construcción del Almacén de Datos PIAD. Sin embargo, las mencionadas en la tabla 30 son los que se han registrado en la base de datos.

Tabla 30 Procedimientos Almacenados del Almacén de Datos PIAD.

N°	Nombre Procedimiento	Cantidad Líneas
1	[actualizarAsistenciaEfectivaAusencias]	40
2	[actualizarAsistenciaEfectivaTodosEstudiantes]	29
3	[actualizarEstudianteCasoAusentismo]	30
4	[actualizarEstudianteCasoRepitencia]	29
5	[actualizarEstudianteIndicadores]	40
6	[actualizarEstudianteRangoNotas]	53
7	[actualizarEstudianteRangoNotas]	31
8	[actualizarTEstudianteBeca]	28
9	[actualizarUbicacionEstudiantes]	31
10	[borradoTablasPiadEnLinea]	44
11	[insertarDimensionesVarMultV1]	255
12	[insertarDmAsignatura]	36
13	[insertarDmColegio]	34
14	[insertarDmCursoLectivo]	34
15	[insertarDmNivel]	32
16	[insertarDmPeriodo]	31
17	[insertarDmProfesor]	44
18	[insertarEscolaridadPadres]	86
19	[insertarHcVariablesMultiplesV1]	44
20	[insertarTAsignaturaClaseEfectiva]	35
21	[insertarTEstudiante]	57
22	[insertarTEstudianteAusentismo]	88
23	[insertarTEstudianteBeca]	94
24	[insertarTEstudianteCasoAusentismo]	101
25	[insertarTEstudianteDesercion]	68
26	[insertarTEstudianteIndicador]	96
27	[insertarTEstudianteNotas]	53
28	[insertarTEstudianteProfesorGrupo]	46
29	[insertarTEstudianteProfesorGrupo]	43
30	[insertarTEstudianteUbicacion]	68
31	[insertarTHogarUniparental]	50
32	[insertarTHojaDatosUnica]	57
33	[insertarTNucleoFamiliar]	56
34	[insertarTPeriodosAD]	41
35	[transformacionTAsignatura]	34
Total		**1938**

Soporte de Decisiones: Se creó un esquema de datos, para albergar las estructuras de la estrella que proveerá la información del cubo multidimensional, del Almacén de Datos. La estrella quedó compuesta por 26 dimensiones y una tabla de hechos.

a. Dimensiones:

- **DmAccesoInternet:** Dimensión que registra SI – NO como indicadores de que el estudiante posee o no acceso a internet.
- **DmAdecuacionCurricular:** Expone los tipos de adecuación curricular, que puede aplicarse a un estudiante.
- **DmAsignatura:** Registra las asignaturas que fueron tomadas en cuenta, para la información del Almacén de Datos.
- **DmBeca:** Muestra las distintas becas que se otorgan a los estudiantes.
- **DmBono:** Dimensión que registra SI- NO como indicador, de que un estudiante recibe bono.
- **DmCanton:** Registra los cantones de Costa Rica.
- **DmCasoAusentismo:** Dimensión que registra SI- NO, como indicador de que un estudiante es un caso de ausentismo.
- **DmCasoDesercion:** Dimensión que registra SI- NO, como indicador de que un estudiante es un caso de deserción, del sistema de educación.
- **DmCircuito:** Registra los circuitos educativos.
- **DmColegio:** Dimensión que registra los centros educativos (Colegios académicos públicos).
- **DmCursoLectivo:** Registra los cursos lectivos.
- **DmDireccion:** Dimensión que registra las direcciones regionales del Ministerio de Educación Pública.
- **DmDistrito:** Dimensión que registra los distritos de Costa Rica.
- **DmEscolaridadMama:** Registra los niveles de escolaridad de la madre.
- **DmEscolaridadPapa:** Registra los niveles de escolaridad del padre.

- **DmEstadoVivienda:** Dimensión que registra los distintos estados que puede tener una vivienda.
- **DmEstudianteTrabaja:** Dimensión que registra SI- NO, como indicador de que un estudiante trabaja.
- **DmHogarUniparental:** Dimensión que registra SI- NO, como indicador de que un estudiante vive sólo con uno de sus padres.
- **DmNivel: R**egistra los niveles que se cursan en la educación pública.
- **DmPeriodo:** Registra los periodos lectivos.
- **DmProfesor:** Dimensión que registra los docentes (profesores).
- **DmProvincia:** Registra las provincias de Costa Rica.
- **DmRangoNotas:** Registra los rangos de notas.
- **DmRepitencia:** Dimensión que registra SI- NO, como indicador de que un estudiante se encuentra repitiendo el curso lectivo.
- **DmSexo:** Dimensión que registra los valores para sexo.
- **DmTenenciaVivienda:** Registra la condición de tenencia de una vivienda.

b. Tabla de hechos:

Para en análisis de información, el Almacén cuenta con una tabla de hechos (HcVariablesMultiplesV1) compuesta por campos calculados y llaves foráneas a las dimensiones. Esta tabla contiene seiscientos noventa y cuatro mil ochocientos ochenta y cinco registros.

Campos de la tabla HcVariablesMultiplesV1

- **cedEstudiante:** Cédula del estudiante
- **cedProfesor:** Cédula del docente (profesor).
- **hcAsistencia:** Cantidad de ausencias del estudiante.
- **hcAsistenciaEfectiva:** Cantidad de lecciones por materia a la debió asistir el estudiante.
- **hcEdad:** Edad del estudiante.
- **hcIngreso:** Ingreso familiar per cápita del estudiante.

- **hcNota:** Nota de rendimiento del estudiante por materia.
- **idAccesoInternet:** Indicador de que el estudiante posee internet.
- **idAdecuacionCurricular:** El tipo de adecuación curricular que recibe el estudiante.
- **idAsignatura:** Identificador de la asignatura cursada por el estudiante.
- **idBeca:** el tipo de beca que recibe el estudiante (en esta se incluye el acceso al comedor y transporte, entre otras).
- **idBono:** Indicador de que el estudiante posee bono escolar.
- **idCanton:** Identificador del cantón donde reside el centro educativo.
- **idCasoAusentismo:** Indicador de que el estudiante es un caso de ausentismo, este se determina si el ausentismo sobrepasa el 15% de la asistencia efectiva.
- **idCasoDesercion:** Indicador de que el estudiante desertó en un curso lectivo, del sistema de educación.
- **idCircuito:** Identificador del circuito al que pertenece el centro educativo.
- **idColegio:** Identificador del centro educativo al que asiste el estudiante.
- **idCursoLectivo:** Identificador del año lectivo que cursa el estudiante.
- **idDireccion:** Identificador de la Dirección Regional, a la que pertenece el centro educativo.
- **idDistrito:** Identificador único del distrito en el que se ubica el centro educativo.
- **idEscolaridadMama:** Identificador del nivel de escolaridad de la madre del estudiante.
- **idEscolaridadPapa:** Identificador del nivel de escolaridad del padre del estudiante.
- **idEstadoVivienda:** Identificador del estado en que se encuentra la vivienda del estudiante.
- **idEstudianteTrabaja:** Indicador de que el estudiante trabaja.
- **idHecho:** Identificador único del registro del hecho.

- **idHogarUniparental:** Indicador de que el estudiante sólo vive con uno de sus padres.
- **idNivel:** Identificador del nivel de escolaridad que cursa el estudiante.
- **idPeriodo:** Periodo lectivo para el curso lectivo.
- **idProvincia:** Identificador de la provincia donde se ubica el centro educativo.
- **idRangoNotas:** Rango de notas al que pertenece la nota obtenida por el estudiante, en una determinada materia.
- **idRepitencia:** Indicador de que el estudiante es repitente.
- **idSexo:** Sexo del estudiante.
- **idTenenciaVivienda:** Identificador de la tenencia de la vivienda de la familia del estudiante.

El Almacén de Datos mantiene un diseño en estrella, permitiendo que, por cada dimensión, puedan realizarse agrupaciones o filtros a los datos contenidos en la tabla de hechos.

Al generarse el cubo, se produce una permutación de todas las dimensiones con la tabla de los hechos, permitiendo agrupar o filtrar los datos de los hechos, por los valores contenidos en las dimensiones.

Para la visualización de la información se utiliza la herramienta de MS Excel, y se han implementado Indicadores Clave de Desempeño que permiten, de una forma gráfica y por colores, identificar el estado de la información analizada.

Como se puede observar, en las dimensiones y la tabla de hechos, se ha incluido todas las variables del Reporte de Variables Múltiples, excepto el grupo al que pertenece el estudiante. Cuando se trabajaba con los datos, se presentaban muchas inconsistencias con el grupo, por lo que se excluyó de las dimensiones. El problema se presentó, por no contemplar que se trabajan subgrupos en algunos casos, y se debían incluir las tablas de tsubgrupo, tsubcargaacademica y tsubgrupoestudiante. Se debe incluir dichos datos en la próxima versión del Almacén de Datos.

4.6.2 Plan de Implementación

El Almacén de Datos desarrollado no es una versión definitiva, por lo que se hace una propuesta de los puntos que deben incluirse en el plan de implementación, una vez que se cuente con la versión final.

a. Preparar de la infraestructura.
    - Hardware requerido.
    - Software requerido.
    - Conectividad.
b. Identificar de los orígenes de datos a incluir.
    - Crear un mapa de los orígenes de datos, identificando la unidad de producción, ubicación y formato.
    - Descripción de los datos en el ámbito del negocio.
c. Actualizar la extracción de los datos.
d. Realizar la transformación de datos.
e. Cargar los datos en la estrella.
f. Generar o actualizar el cubo.
g. Determinar los usuarios de la información, y lo que se puede usar.
h. Definir el despliegue de la información en la organización.

4.6.3 Supervisión y mantenimiento del plan

La supervisión y el mantenimiento son cuestiones importantes, si los resultados del Almacén de Datos se hacen parte del negocio cotidiano y de su ambiente. Una preparación cuidadosa de una estrategia de mantenimiento, ayuda a evitar errores en los datos o la disponibilidad de los mismos. Para supervisar el desarrollo de los resultados del Almacén de Datos, se requiere contar con un plan detallado para su vigilancia y mantenimiento.

La supervisión y mantenimiento del plan de implementación del Almacén de Datos debe contemplar:

- **Definición de nuevos requerimientos de información**: de la misma manera que sucede con cualquier proyecto, las necesidades de hoy, no serán las del mañana; por lo tanto, se deberá mantener un constante análisis de las necesidades de información del negocio.
- **Análisis de los datos:** se deberá entender y preparar los datos, para ser incluidos en el diseño del Almacén de Datos.
- **Diseño y modelado:** Actualizar el diseño de la base de datos de extracción, transformación y carga de los datos que se desean analizar; incluyendo la metada de éstos. Actualizar el diseño del servicio de análisis para disponer la información y las estructuras de despliegue de datos.
- **Implementar los nuevos requerimientos:** la implementación implica, la extracción de datos del sistema operacional y transformación de los mismos, la carga de los datos validados al Almacén de Datos, según la periodicidad definida y las necesidades de actualización del negocio. La explotación de los datos se hace mediante el uso de reportes y consultas, o herramientas de visualización, como Excel. Es necesario, mantener el control de los datos, por medio de los metadatos técnicos y de negocio, y que estos se encuentren accesibles a los usuarios finales y al administrador.
- **Revisión:** determinar las posibles mejoras que se pueden implementar a futuro.
- **Diseño de una estructura de capacitaciones:** las capacitaciones tienen como objetivo, el proporcionar formación estadística necesaria, para aprovechar de la mejor manera la funcionalidad incluida en la aplicación. Se deberán realizar prácticas sobre los nuevos datos implementados, que permitan fijar los conceptos adquiridos, y sirvan de formación a los usuarios.

4.6.4 Informe del proyecto

Como última actividad, se ha realizado un informe sobre el desarrollo del Almacén de Datos, que contempla un resumen de cada una de las fases implementadas de la metodología CRISP.

Contenido:

a. Resumen de la comprensión del negocio: contexto, objetivos, y criterios de éxito

b. Resumen del proceso de Modelado del Almacén de Datos

c. Resumen de los resultados

d. Evaluación de los resultados

e. Resumen de la implementación y de los planes de mantenimiento

f. Análisis Costo/Beneficio

g. Conclusiones para el negocio

h. Conclusiones para trabajos futuros.

5 Conclusiones

Se concluye que al documentar los requerimientos y los datos que aportan valor para el diseño del Almacén de Datos, se debe tomar en cuenta que:

C-01-01: El proceso de conocer las reglas del negocio y la correcta interpretación de los datos, requiere de una gran cantidad de tiempo de los programadores, por no contar éstos con el conocimiento del rol de negocio, ni de los orígenes de datos de la organización.

C-01-02: La cantidad de información disponible para análisis y toma de decisiones, se ve afectada por la calidad de los datos contenidos en los sistemas transaccionales.

Se concluye que para diseñar un Almacén de Datos, mediante el modelado de estructuras, para contener los datos objeto del análisis, se debe contar con:

C-02-01: Una arquitectura de extracción, transformación y carga (ETC) sólida y escalable, debe estar diseñada de tal manera, que utilice una serie de herramientas, las cuales satisfagan las diferentes necesidades de un proceso de ETC, para un Almacén de Datos. Para este proyecto, se utilizó el conjunto de herramientas Microsoft para inteligencia de negocios de la siguiente manera: Nivel 1: El agente SQL se encargará de ejecutar tareas, las cuales darán orden, calendarizarán y ejecutarán actividades para la creación de un cubo, reporte o ICD (indicador clave de desempeño). Nivel 2: El controlador de flujo SSIS (Sql Server Integration Services), se encargará de crear y ejecutar paquetes, los cuales apliquen procesos a nivel de tablas (varias). Nivel 3: El controlador de flujo de datos SSIS, se ocupará de extraer, transformar, integrar, insertar y ejecutar procesos para una tabla. Nivel 4: El motor de base de datos, por medio de procedimientos almacenados, validará la calidad de los datos, los trasformará, y validará reglas de negocio, para una o más columnas de una tabla.

C-02-02: Una arquitectura sólida y escalable para un Almacén de Datos, debe contar con tres áreas principales (o esquemas), como mínimo. Estas áreas deben contemplar: la de extracción de datos (recibe los datos sin transformaciones de las fuentes de información del Almacén de Datos), la transformación de datos (se aplican los cambios necesarios a los datos para limpiar, validar la calidad, y prepararlos para procesos de

carga) y el soporte de decisiones (donde existen y se crean los repositorios base, para los modelos multidimensionales).

Del trabajo realizado para la implementación del Almacén de Datos, las herramientas de análisis de datos y apoyo al proceso de toma de decisiones, se ha llegado a las siguientes conclusiones:

C-03-01: Definir los valores por defecto para los hechos, apoya al mejoramiento de la calidad de datos, ya que con esta práctica ayuda a fijar un valor, cuando un dato no se puede determinar. Por ejemplo, asignar un valor por defecto para el ingreso por mes o la nota de un estudiante, cuando esta no está disponible.

C-03-02: Es importante, definir en forma adecuada los tipos de datos que se utilizarán en el Almacén de Datos, y asegurarse que ésta definición se respete, en todo momento, en el Almacén de Datos. Esta definición se deriva de un adecuado análisis de tipos de datos, de las diferentes fuentes información del Almacén de Datos.

C-03-03: La escogencia de la cantidad de datos (significativa), para el desarrollo del Almacén de Datos, ayuda a identificar problemas similares al ambiente de producción, lo que permite identificar, con mayor claridad, los problemas de calidad de datos.

C-03-04: Evaluar y validar los identificadores, es una actividad clave para asegurar una data confiable. Por ejemplo, que los identificadores de los profesores, estudiantes, asignaturas y periodos sean válidos y únicos; esto fomenta que el Almacén de Datos sea confiable.

C-03-05: La validación de los datos claves del Almacén de Datos es una tarea necesaria durante el diseño. Los datos, considerados clave, se derivan del análisis y la exploración de datos. Se consideraron datos clave: estudiante, periodo, nivel, materia, grupo y colegio.

C-03-06: La evaluación de datos en los orígenes por separado, aunque sean de tamaño significativo, conlleva a que se omitan procesos de transformación necesarios, ya que no permite identificar problemas de calidad de datos, como repeticiones y el uso de valores diferentes, para un mismo significado.

C-03-07: La utilización de estructuras de tipo multidimensional, para mostrar al cliente la información que proveerá el modelo multidimensional, permite confirmar la validez de la información del cubo.

C-03-08: La realización de pruebas, a las distintas capas de la solución de inteligencia de negocios, (proyecto de Análisis Service en Visual Studio, la base de datos multidimensional en el Sistema Administrador de Bases de Datos MS SQQL Server 2008 R2 y el cliente de datos Excel) asegura que la información es confiable para el uso por parte de los usuarios.

En el proceso de evaluación de la aplicabilidad de las actividades, propuestas en la Metodología CRISP, para el desarrollo de un Almacén de Datos, se llegó a las siguientes conclusiones:

C-04-01: La fase de conocimiento del negocio es altamente aplicable a un proyecto de Almacén de Datos, ya que se puede aplicar el 96% de las actividades que la componen. Mediante la definición de objetivos de negocio, objetivos del Almacén de Datos y criterios de éxito; permite establecer, con el cliente, lo que se espera de la herramienta, y contar con parámetros de medición del éxito alcanzado.

C-04-02: Tanto los proyectos de minería de datos, como los proyectos de desarrollo de almacenes de datos, comparten la particularidad de que requieren de datos existentes, provenientes de diferentes orígenes de datos para ser alimentados. Por lo cual, es de vital importancia que, antes de diseñar y desarrollar el proyecto, se analicen y conozcan los datos, y se preparen para el respectivo uso y análisis. Las fases de conocimiento y preparación de los datos son altamente aplicables y necesarias, para ambos tipos de proyectos.

C-04-03: La fase de modelado de un proyecto de minería de datos, difiere a la de un proyecto de desarrollo de almacén de datos. El primero, usa funciones de minería de datos y aplica algoritmos; mientras que el segundo, debe contar con tres modelos: el de negocio, el dimensional y el físico. Esta fase es bajamente aplicable, ya que solo el 33% de las actividades que la componen fueron aplicadas al proyecto.

C-04-04: La fase de evaluación en el desarrollo del Almacén de Datos permite determinar el alcance de los objetivos y los criterios de éxito, tanto para el negocio, como para el proyecto. Se aplicó el 100% de las actividades en la evaluación de los

resultados y el proceso de revisión, en la determinación de los siguientes pasos se aplica el 63% de las actividades, siendo altamente aplicable y necesaria.

C-04-05: En un proyecto de minería de datos se arrojan resultados propios de la exploración de los datos, mientras que en un proyecto de Almacén de Datos se trabaja la información para ponerla a disposición de los usuarios, que serán quienes determinen los resultados para la toma de decisiones, por el uso de la herramienta. Es por esta razón, que la fase de implementación fue medianamente aplicada en este proyecto, de veintisiete actividades se aplicaron únicamente 15, para una aplicabilidad de un 63%.

C-04-06: La metodología CRISP es una herramienta muy valiosa, que puede ser altamente aplicable a un proyecto de Almacén de Datos. Su aporte más valioso se encuentra en una guía de tareas, compuestas por múltiples actividades que llevan a la elaboración de productos documentados del proyecto. Esta metodología, se convierte en un instrumento muy poderoso que lleva de la mano desde un novato, hasta un experto en el desarrollo de soluciones de Inteligencia de Negocios.

6 Recomendaciones

R-01-01: Es recomendable que la organización provea un Experto en la Materia (SME: Subject Matter Expert), que tenga el conocimiento del rol del negocio y los orígenes de datos que serán utilizados para alimentar el Almacén de Datos, esto para que pueda transmitir de forma rápida y eficiente la inducción al conocimiento del negocio a los desarrolladores, y colabore en la ubicación de los datos que van a aportar valor. La dedicación de tiempo del SME debe ir de un 30% a 50%, para asegurar que cuenta con el tiempo adecuado, para sesiones de entrenamiento y sesiones de preguntas por parte de los instruidos.

R-01-02: Es aconsejable que la organización, previo a la decisión de implementar un Almacén de Datos, realice un análisis del estado de la información que desean incluir para análisis y apoyo a la toma de decisiones, con el fin de hacer un plan de mejoramiento de su data. En el caso de las instituciones públicas que proveen información para toma de decisiones de gobierno, se recomienda la creación de una directriz por parte de la Contraloría General de la República u otro entre de control, que establezca estándares de la calidad de los datos.

R-01-03: Con la finalidad de poder incluir, en el Almacén de Datos, la mayor cantidad de información disponible en los orígenes de datos; es aconsejable establecer políticas de gobernanza de la información, que se implementen en las restricciones de los repositorios de datos y las reglas de negocio sobre completitud y calidad de los datos, requeridos para análisis y apoyo al proceso de toma de decisiones.

R-02-01: Se recomienda que el equipo que brindará soporte y mantenimiento al Almacén de Datos, posea un conocimiento adecuado de las siguientes tecnologías: SSIS, SSAS, SSRS, Ms Sql Server 2008, SQL, Transact–Sql, procedimientos almacenados, Visual Studio 2008, afinamiento de consultas, tablas pivotes en Excel (fórmulas de análisis y creación de reportes). La versión de Excel a utilizar debe ser 2007, o más reciente.

R-02-02: Se recomienda que el repositorio de datos, que albergará el Almacén de Datos, cuente al menos con tres esquemas, que permitan separar el trabajo con los datos en extracción, transformación y soporte de decisiones. Con el tiempo, el Almacén de

Datos tenderá al crecimiento, presentando la necesidad de especializar los datos; por lo que se aconseja, la especialización de los esquemas de soporte de decisiones, para que reflejen el dominio del negocio que representan. Por ejemplo, se deberá crear un esquema de soporte de decisiones para los datos de colegio, de la escuela, de preescolar y escuelas de enseñanza especial.

R-03-01: Se recomienda el uso de valores por defecto, de esta manera se evita la inclusión de valores nulos carentes de significado para la toma de decisiones, o discriminar la información por falta de completitud en los mismos.

R-03-02: Utilizar los tipos de datos, que incluye el Sistema Gestor de Bases de datos, que más se ajusten a las necesidades presentes y futuras de la información a contener, en cada repositorio de datos. Es aconsejable utilizar datos estándar de SQL, para asegurar la independencia de los datos del SGBD utilizado.

R-03-03: Es recomendable contar con la opinión de un experto del negocio, asignado por la empresa, que indique la cantidad de orígenes de datos a incorporar en el Almacén de Datos, y colabore en la determinación de posibles problemas de integración de datos.

R-03-04: Una vez realizada la extracción de los datos de los orígenes y almacenarlas en un único repositorio, es recomendable realizar el proceso de exploración, con el fin de poder identificar y solucionar problemas de calidad de datos, antes de iniciar el proceso de transformación y carga al Almacén de Datos.

R-03-05: Incorporación del SME desde el inicio del proceso de diseño, y antes de iniciar el proceso de ETC, para que lleve a cabo la validación de datos a extraer y sus respectivas transformaciones, con el fin de ahorrar tiempo y asegurar un resultado de calidad de datos.

R-03-06: Es recomendable unificar los orígenes de datos, antes de realizar la evaluación e integración de datos, esto para poder determinar, sobre el conjunto de datos completo, los requerimientos de transformación para cargar el Almacén de Datos.

R-03-07: Es recomendable la utilización de ETM's, que permitan asegurar el entendimiento, de los expertos del negocio, de la información contenida en las dimensiones, y el valor que la misma va a aportar al negocio, con el propósito de que ellos puedan dan por aceptado el éxito del proyecto.

R-03-08: Se aconseja la participación de los expertos del negocio y los usuarios de la información, en la validación del proceso de desarrollo del Almacén de Datos, para asegurar el cumplimento de los requisitos de información y la calidad de esta.

R-04-01: Establecer, con los líderes del negocio, los objetivos del negocio para el Almacén de Datos, los objetivos del Data Warehouse y los criterios de éxito, que deberán ser satisfechos con el proyecto, esto permitirá que los alcances sean claros a todos los involucrados.

R-04-02: Para desarrollar un Almacén de Datos, capaz de proveer accesibilidad a información consistente, adaptable y elástica para la toma de decisiones, es necesario comprender los orígenes de datos que lo van a alimentar, y analizar la limpieza y transformación que estos requieren. Por lo cual, se recomienda asignar, en el plan del proyecto, una cantidad de tiempo considerable a estas tareas, y contar con experto en materia del negocio que apoye en la inducción del conocimiento a los desarrolladores.

R-04-03: Para utilizar la metodología CRISP, en el desarrollo de un Almacén de Datos, se recomienda ajustar las tareas de la fase de modelado, por las etapas de diseño que permitan generar el modelo del negocio, el modelo dimensional y el modelo físico con las respectivas actividades, para cada una de las etapas.

R-04-04: Para llevar a cabo una evaluación del desarrollo del Almacén de Datos asertiva, se recomienda integrar un equipo de evaluadores que incluya: un experto del negocio, que analice la información producida; un conocedor de los orígenes de datos que analice las fuentes utilizadas, para la producción de información; usuarios de la información que validen el valor de la misma para el negocio; y los desarrolladores de la herramienta, para validar que ésta funciona de forma óptima.

R-04-05: La implementación de la herramienta de análisis de datos, y apoyo al proceso de toma de decisiones, requiere la formulación de un plan, cuyas actividades determinen su realización en el negocio. Además, en el proceso de evaluación de un proyecto de Almacén de Datos, es muy posible que siempre se descubran nuevos requerimientos de información. Por lo cual, es aconsejable que al elaborar el plan, sea realizado en conjunto con personal de toma de decisiones del negocio, que valide el trabajo realizado, y posea el poder de definir ante la organización las pautas para la implementación.

R-04-06: Se recomienda la metodología CRISP como estándar, para la generación de los productos entregables en el desarrollo de un proyecto de Almacén de Datos.

7 Reflexiones Finales – Trabajos Futuros

7.1 Basados en la incompletitud de la información sobre el ingreso per cápita de la familia del estudiante, consideramos una opción práctica el utilizar la tabla de salarios mínimos del Ministerio de Trabajo, para determinar el ingreso de los encargados de los estudiantes, según la ocupación que manifestaron ejercer, de este modo, poder contar con un ingreso per cápita para análisis.

7.2 Para una óptima administración del Almacén de Datos, se recomienda al definir un equipo de gobernanza de la información, que incluya los siguientes roles:

- Encargado de Gobernanza de Datos, que realice las labores de definición de datos, definición de la interpretación de los datos, creación, y actualización del diccionario de datos y nuevos requerimientos.
- Administrador de Bases de Datos del Almacén de Datos, que se encargue de la administración, mantenimiento y afinamiento del Almacén de Datos.
- Técnico de Soporte de Análisis a los usuarios, que se dedique a la interpretación de los datos, hacer resolución de consultas sobre problemas del Almacén de Datos, y dar capacitación a los usuarios en las herramientas y los datos. Es importante, que el asignado a este rol, mantenga una correcta administración de tiquetes de asistencia a usuarios.
- Desarrollador de Almacén de Datos, responsable del desarrollo de nuevos requerimientos.

7.3 Se recomienda al negocio, que para la inclusión de la data de centros educativos, según su nivel de enseñanza, se separe el esquema de soporte de decisiones en un esquema para cada tipo de centro, ya que las reglas de negocio que aplican para generar los indicadores son distintas. Por ejemplo, la determinación de un caso de ausentismo o el registro de la asistencia regular al comedor, son diferentes para primaria y secundaria.

7.4 Como trabajo futuro, se recomienda implementar al Almacén de Datos la información de preparación y desempeño de los docentes, por considerarse un indicador relevante correlacionado al desempeño y asistencia de los estudiantes.

7.5 Con el fin de mejorar la calidad de los datos requeridos para análisis, y apoyo al proceso de toma de decisiones, se recomienda implementar un plan de limpieza y

completitud de los datos; dando prioridad a los repositorios de los centros educativos que, actualmente, se encuentran utilizando el sistema PIAD en línea y posteriormente, para los centros educativos que se irán incorporando a esta modalidad.

7.6 Definir un proyecto de mejora para la aplicación PIAD, que asegure la obtención de los datos considerados clave para el análisis, y soporte a la toma de decisiones. Por ejemplo, la hoja de datos del estudiante.

7.7 Crear una metodología para definir, planear y priorizar los nuevos proyectos de inteligencia de negocios (ampliación del ETL, Nuevos reportes, nuevos cubos).

7.8 Realizar una prueba de concepto, para analizar la implementación de Sharepoint, para mostrar y compartir datos de Inteligencia de Negocios.

7.9 Evaluar la factibilidad de la migración de las herramientas a la versión 2012.

7.10 Realizar un proyecto de minería de datos, que permita la obtención de patrones de conducta que agravan o mejoran, de alguna manera, el rendimiento y ausentismo de los estudiantes.

7.11 Crear un proceso de inclusión de nuevos centros educativos, enfocado en la calidad de datos.

7.12 Desde un ámbito profesional, se considera un trabajo futuro el ajustar la metodología CRISP-DM, a una metodología CRISP-DW, eliminando las tareas y actividades que no se ajusten a un proyecto de Almacén de Datos, e incluyendo las tareas y actividades que sirvan de guía en la fase de Modelado e Implementación de éste.

7.13 Crear una propuesta de calidad de datos para las entidades públicas, que esté basada en cuatro pilares: Los usuarios, interfaz de usuario de las aplicaciones, integridad de los datos y su normalización (uso de valores predeterminados, llaves únicas, llaves primarias, llaves foráneas, entre otros) y por último los procesos de negocio (cada proyecto debe contener un capítulo orientado a la calidad de datos).

8 Glosario

Actividad: Es parte de una tarea en la Guía de Usuario; describe las acciones para realizar una tarea.

AED: Asociación Empresarial para el Desarrollo.

ANDE: Asociación Nacional de Educadores.

ASIS: Asociación para la Innovación Social.

Caso del proceso: Un proyecto específico descrito en términos del modelo de proceso

Contexto de minería de datos: Un conjunto de restricciones y presunciones, tales como el tipo de problema, las técnicas o herramientas, el dominio de aplicación.

CRISP-DM: Procesos Estándar para la industria de la minería de datos (CRoss-Industry Standard Process for Data Mining).

ICD: Indicadores clave de desempeño, conocidos en inglés como KPI o "Key Performance Indicators".

DGEC: Dirección de Gestión y Evaluación de la Calidad.

ETL: Extracción, Transformación y Carga de datos (Extract-Transform-Load).

ETM: Estructura de Tipo Multidimensional.

Especializado - Una tarea que hace presunciones específicas, en contextos específicos de minería de datos.

Fase - Un término para la parte de alto nivel del modelo de proceso CRISP-DM; consiste en tareas relacionadas.

Genérico - Una tarea que mantiene un cruce con todos los proyectos de minería de datos posibles.

Guía de usuario - Asesoramiento específico sobre cómo realizar proyectos de minería de datos.

MEP: Ministerio de Educación Pública.

Metodología de CRISP-DM - El término general para todos los conceptos desarrollados y definidos en el CRISP-DM.

MDX: declaraciones de expresiones multidimensionales (Multidimensional Expressions statements).

Modelo de proceso - Define la estructura de proyectos de almacenes de datos, y proporciona la guía para su ejecución; consiste en el modelo de referencia y en la guía de usuario

Modelo de referencia - Descomposición de proyectos de almacenes de datos en fases, tareas, y salidas.

MOLAP: Implementación OLAP que almacena los datos en una base de datos multidimensional.

Multivariables: combinación de dos o más variables.

OLAP: Procesamiento Analítico en Línea (OLAP)

PIAD: Proyecto de Informatización de Alto Desempeño.

POA: Plan Operativo Anual.

Repitencia: Término utilizado en el medio educativo, para referirse a la repetición escolar.

ROLAP: Implementación OLAP que almacena los datos en un motor relacional.

Salida - El resultado tangible de la ejecución de una tarea.

SME: Sujeto Experto en la Materia o Experto del Rol de Negocio

Tarea - Una serie de actividades para producir una o más salidas; parte de una fase.

9 Referencias

Abarca, S. (1995). Psicología de la motivación. San José, Costa Rica. EUNED.

Casullo, M. Cayssials, A. (1994). Proyecto de vida y decisión vocacional. Buenos Aires. Argentina. Paídos.

Chapman,P.,Clinton,J.,Keber,R.,Khabaza,T.,Reinartz,T.,Shearer,C.,Wirth,R. (2000).CRISP-DM1. Step by Step BI Guide.Edited by SPSS.

Comisión Económica para América Latina (CEPAL). (2002). Panorama Social de América Latina 2001-2002. Capitulo III Deserción escolar, un obstáculo para el logro de los Objetivos del Desarrollo del Milenio. Recuperado de: http://www.eclac.org/cgibin/getProd.asp?xml=/publicaciones/xml/4/11254/P11254.xml&xl=/dds/tpl/p9f.xsl&base=/tpl/top-bottom.xslt

English, Larry P. (1999). Improving Data Warehouse and Business Information Quality. Canada. Wiley.

Informe XI Estado de la Nación. Recuperado de: http://www.estadonacion.or.cr/biblioteca-virtual/costa-rica/estado-de-la-nacion/informes-anteriores/informes-por-tema/297-informe-xi

Gorbach,I., Berger,A., Melomed,E. (2008). Microsoft SQL Server 2008 Analysis Services UNLEASHED. USA. Sams.

Guadamuz, L. (2008). Análisis de los determinantes del rendimiento académico. Recuperado de http://www.lorenzoguadamuz.net/Docs/AN%C3%81LISIS%20DE%20LOS%20DETERMINANTES%20DEL%20RENDIMIENTO%20ESCOLAR.pdf

Gutiérrez, L. (2007). Educación y trabajo en jóvenes costarricenses. Revista Instituto Nacional de Investigaciones en Educación. Universidad de Costa Rica. Recuperado de: http://www.revista.inie.ucr.ac.cr/articulos/2-2007/archivos/jovenes.pdf

Harinath, S.,Carroll,M., Meenakshisundaram,S. Zare,R.,Guang-Yeu Lee,D. (2009). Professional Microsoft SQL Server Analysis Services 2008 with MDX. Canada. Wiley.

Holland, J. (1987). La elección vocacional: teoría de las carreras. México. Trillas.

Inmon, W.H. (2001). Corporate Information Factory. Canada. Wiley.

Irola, A. (2002). Análisis cualitativo de los factores que provocan la deserción escolar en estudiantes de sétimo año del Colegio Técnico Profesional de Pacayas. Programa de Estudios de Posgrado en Administración Educativa para optar por el grado de Magíster. Universidad de Costa Rica.

Kaplún, G. (2004). Culturas juveniles y educación: pedagogía crítica, estudios culturales e investigación participativa. Publicado en los jóvenes; múltiples miradas, UNC Neuquén. Recuperado de: http://www.liccom.edu.uy

Kimball, Ralph. (2008). The Data Warehouse Lifecycle Toolkit. Canada. Wiley.

Kinball, Ralph. (2004). The Data Warehouse ETL Toolkit. Canada. Wiley.

Krauskopf, D. (2002). Adolescencia y Educación. Costa Rica. EUNED.

Laura L. Reeves, A Manager's Guide to Data Warehousing, WIley, página 4.

Margaret D. LeCompte, Analyzing Qualitative Data, Lawrence Erlbaum Associates (Taylor & Francis Group) 2000.

María P. Sandín, La Enseñanza De La Investigación Cualitativa, Revista de Enseñanza Universitaria 2003.

Richards, C. (2006).Historia de desertores: la vida detrás de las cifras. Recuperado de: http://innovemos.unesco.cl/medios/historiadevidas

Ruiz, A. (2005). Universalización de la educación secundaria y reforma educativa. Ponencia preparada para el Undécimo Informe Estado de la Nación. San José, Programa Estado de la Nación.

Schiefelbein, E. and Simmons, J. (1981). The determinants of school achievement: A review of the research for Developing Countries. Otawa: IDRC.

Sheldon, Robert (02 March 2010). Agregando un ICD a un cubo de análisis de servicio de SQL Server. Recuperado de: https://www.simple-talk.com/sql/reporting-services/adding-a-kpi-to-an-sql-server-analysis-services-cube/

Suchman, E. A. (1967). Evaluative Research: Principles and Practice in Public Service and Social Action Programs. New York : Russell Sage Foundation.

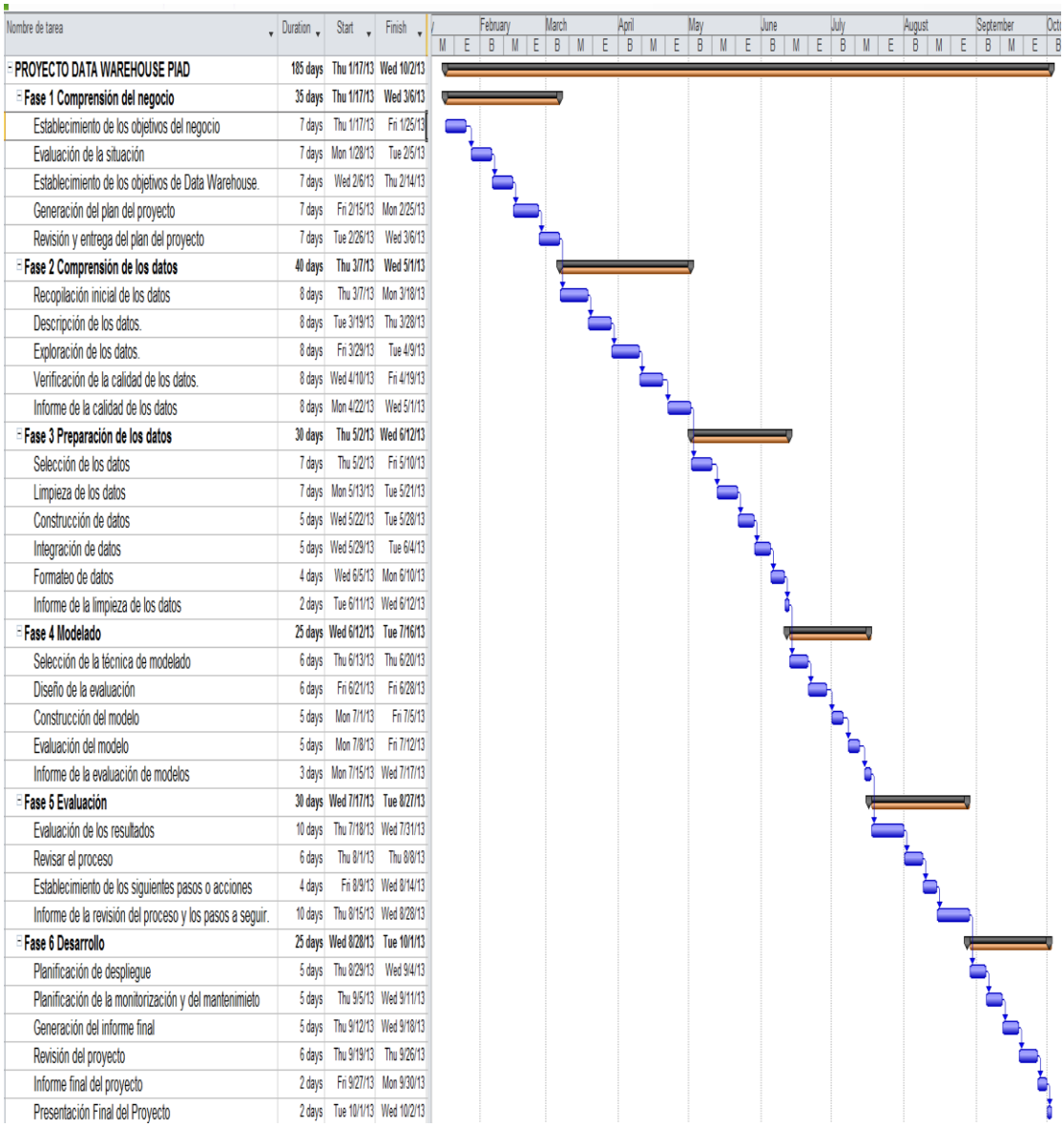
Thomsen, E. (2002). Olap Solutions Building Multidimensional Information Systems. Canada. Wiley.

Uruñuela (2005). ABSENTISMO ESCOLAR. Recuperado de http://weib.caib.es/Documentacio/jornades/jornades_menors/p_urunuela.pdf

Vargas, V. (2006). Educación y Deserción. Editorial de Opinión. Consultado el día. Recuperado de: http://www.nacion.com/ln_ee/2006/diciembre/31/opinion945461.html

10 Anexos

Anexo 1. Cronograma (17 de Enero del 2013 al 02 de Octubre del 2013) Sujeto a disposición de PIAD.



Nombre de tarea	Duration	Start	Finish
PROYECTO DATA WAREHOUSE PIAD	185 days	Thu 1/17/13	Wed 10/2/13
Fase 1 Comprensión del negocio	35 days	Thu 1/17/13	Wed 3/6/13
Establecimiento de los objetivos del negocio	7 days	Thu 1/17/13	Fri 1/25/13
Evaluación de la situación	7 days	Mon 1/28/13	Tue 2/5/13
Establecimiento de los objetivos de Data Warehouse.	7 days	Wed 2/6/13	Thu 2/14/13
Generación del plan del proyecto	7 days	Fri 2/15/13	Mon 2/25/13
Revisión y entrega del plan del proyecto	7 days	Tue 2/26/13	Wed 3/6/13
Fase 2 Comprensión de los datos	40 days	Thu 3/7/13	Wed 5/1/13
Recopilación inicial de los datos	8 days	Thu 3/7/13	Mon 3/18/13
Descripción de los datos.	8 days	Tue 3/19/13	Thu 3/28/13
Exploración de los datos.	8 days	Fri 3/29/13	Tue 4/9/13
Verificación de la calidad de los datos.	8 days	Wed 4/10/13	Fri 4/19/13
Informe de la calidad de los datos	8 days	Mon 4/22/13	Wed 5/1/13
Fase 3 Preparación de los datos	30 days	Thu 5/2/13	Wed 6/12/13
Selección de los datos	7 days	Thu 5/2/13	Fri 5/10/13
Limpieza de los datos	7 days	Mon 5/13/13	Tue 5/21/13
Construcción de datos	5 days	Wed 5/22/13	Tue 5/28/13
Integración de datos	5 days	Wed 5/29/13	Tue 6/4/13
Formateo de datos	4 days	Wed 6/5/13	Mon 6/10/13
Informe de la limpieza de los datos	2 days	Tue 6/11/13	Wed 6/12/13
Fase 4 Modelado	25 days	Wed 6/12/13	Tue 7/16/13
Selección de la técnica de modelado	6 days	Thu 6/13/13	Thu 6/20/13
Diseño de la evaluación	6 days	Fri 6/21/13	Fri 6/28/13
Construcción del modelo	5 days	Mon 7/1/13	Fri 7/5/13
Evaluación del modelo	5 days	Mon 7/8/13	Fri 7/12/13
Informe de la evaluación de modelos	3 days	Mon 7/15/13	Wed 7/17/13
Fase 5 Evaluación	30 days	Wed 7/17/13	Tue 8/27/13
Evaluación de los resultados	10 days	Thu 7/18/13	Wed 7/31/13
Revisar el proceso	6 days	Thu 8/1/13	Thu 8/8/13
Establecimiento de los siguientes pasos o acciones	4 days	Fri 8/9/13	Wed 8/14/13
Informe de la revisión del proceso y los pasos a seguir.	10 days	Thu 8/15/13	Wed 8/28/13
Fase 6 Desarrollo	25 days	Wed 8/28/13	Tue 10/1/13
Planificación de despliegue	5 days	Thu 8/29/13	Wed 9/4/13
Planificación de la monitorización y del mantenimieto	5 days	Thu 9/5/13	Wed 9/11/13
Generación del informe final	5 days	Thu 9/12/13	Wed 9/18/13
Revisión del proyecto	6 days	Thu 9/19/13	Thu 9/26/13
Informe final del proyecto	2 days	Fri 9/27/13	Mon 9/30/13
Presentación Final del Proyecto	2 days	Tue 10/1/13	Wed 10/2/13

Anexo 2. Lista de cotejo de las Variables del Reporte de Variables Múltiples del PIAD, incluidas en el Almacén de Datos del PIAD.

Variables del Reporte de Variables Múltiples del PIAD Incluidas en el Almacén de Datos PIAD		
Criterio	**Variable**	**Aplicado**
Curso Lectivo	Curso Lectivo	SI
	Periodo	SI
Población	Centro Educativo	SI
	Docente	SI
	Estudiante	SI
	Grupo	NO
	Nivel o Año	SI
Rendimiento Académico y Conducta	Asignatura	SI
	Rendimiento	SI
Información familiar del estudiante	Acceso a Internet	SI
	Escolaridad de la madre	SI
	Escolaridad del padre	SI
	Hogar uniparental	SI
	Estado de la vivienda	SI
	Tenencia de la vivienda	SI
	Ingreso mensual per cápita	SI
Información individual del estudiante	Edad	SI
	Sexo	SI
	Prioridad en el comedor.	SI
	Recibe beca	SI
	Recibe bono escolar	SI
	Recibe transporte	SI
Criterios de riesgo	Repitencia	SI
	Adecuaciones	SI
	Llegadas tardías	SI
	Casos de ausentismo	SI
	Deserción	SI

Anexo 3. Listas de Cotejo de la Metodología CRISP por las actividades aplicadas en el proyecto

Metodología CRISP Fase 1 - Comprendiendo el Negocio			
Tarea	**Salida**	**Actividades**	**Realizada**
1. Determinación de objetivos de negocio	**Objetivos del Negocio**	De manera informal describir el problema a ser solucionado	SI
		Especificar todas las preguntas de negocio, tan precisas como sea posible	SI
		Especificar cualquier otras exigencias de negocio (por ejemplo, el negocio no quiere perder a ningún cliente)	SI
		Especificar las ventajas esperadas, en términos de negocio	SI
	Criterios de éxito de negocio	Especificar criterios de éxito de negocio (por ejemplo, mejorar la tasa de respuesta en una campaña de correo en el 10 % y marcar la tasa en el 20%)	SI
		Identificar quien evalúa los criterios de éxito	SI
2. Evaluación de la situación	**Inventario de recursos/Hardware**	Identificar el hardware básico	SI
		Establecer la disponibilidad del hardware básico, para el proyecto de minería de datos	SI
		Comprobar si la planificación del mantenimiento de hardware se opone a la disponibilidad del hardware, para el proyecto de minería de datos.	NO
		Identificar el hardware disponible para ser usado por la herramienta de minería de datos (si el instrumento es conocido en esta etapa)	SI
	Inventario de recursos/Fuentes de Datos y Recursos	Identificar las fuentes de datos	SI
		Identificar el tipo de fuentes de datos (fuentes en línea, expertos, documentación escrita, etc.)	SI
		Identificar fuentes de conocimiento	SI
		Identificar el tipo de fuentes de conocimientos (fuentes en línea, expertos, documentación escrita, etc.)	SI
		Comprobar herramientas disponibles y técnicas	SI
		Describir el conocimiento de generalidades relevantes (de manera informal o formal)	SI
	Inventario de recursos/Fuentes de Personal	Identificar al patrocinador de proyecto (si difiere del patrocinador interno como en la Sección 1.1.1)	SI
		Identificar al administrador de sistema, el administrador de base de datos, y el personal de soporte técnico para futuras preguntas	SI

		Identificar a analistas de mercado, los expertos en minería de datos, y estadísticos, y comprobar su disponibilidad	SI
		Comprobar la disponibilidad de expertos de dominio para fases posteriores	SI
	Requerimientos	Especificar el perfil del grupo objetivo	SI
		Capturar todas los requerimientos en la planificación	SI
		Capturar los requerimientos de comprensibilidad, exactitud, desarrollar habilidades, mantenimiento, y repetibillidad del proyecto de minería de datos y los modelos resultantes.	SI
		Capturar los requerimientos de seguridad, restricciones legales, de privacidad, información, y planificación de proyecto	SI
	Presunciones	Aclarar todas las presunciones (incluyendo las implícitas), y las hechas por ellos explícitamente (por ejemplo, dirigir las cuestiones de negocio, a un número mínimo de clientes con la edad por encima de 50 es necesaria)	SI
		Listar las presunciones sobre calidad de datos (por ejemplo, exactitud, disponibilidad)	SI
		Listar las presunciones sobre factores externos (por ejemplo, cuestiones económicas, productos competitivos, avances técnicos)	SI
		Aclarar presunciones que conducen a cualquiera de las estimaciones (por ejemplo, el precio de un instrumento específico es asumido para ser menor que 1,000 $)	SI
		Listar todas las presunciones en cuanto a, si es necesario entender y describir o explicar el modelo (Por ejemplo, como el modelo y los resultados son presentados a la dirección / patrocinador)	SI
	Restricciones	Comprobar restricciones generales (por ejemplo, cuestiones legales, presupuesto, escalas de tiempo, y recursos)	SI
		Comprobar el correcto acceso a fuentes de datos (por ejemplo, restricciones de acceso, la contraseña requerida)	SI
		Comprobar la accesibilidad técnica de datos (los sistemas de operaciones, el sistema de administración de datos, el formato de archivo y de base de datos)	SI
		Comprobar si el conocimiento relevante es accesible	SI
		Comprobar restricciones de presupuesto (gastos fijos, gastos de implementación, etc.)	SI

	Riesgos	Identificar riesgos de negocio (por ejemplo, el competidor aparece primero con mejores resultados).	SI
		Identificar riesgos de organización (por ejemplo, el departamento que solicita el proyecto no tiene financiación para el proyecto).	SI
		Identificar riesgos financieros (por ejemplo, aumentar la financiación depende de los resultados iniciales de minería de datos).	SI
		Identificar riesgos técnicos	SI
		Identificar los riesgos que dependen de datos y de las fuentes de datos (por ejemplo, la mala calidad y cobertura).	SI
	Plan de Contingencias	Determinar condiciones en las que cada riesgo puede ocurrir.	SI
		Desarrollar planes de contingencia.	SI
	Terminología	Comprobar la disponibilidad previa de glosarios; si no comience a bosquejar glosarios.	SI
		Hablar a expertos de dominio, para entender su terminología.	SI
		Familiarizarse con la terminología de negocio.	SI
	Costos y Beneficios	Estimar el costo para la colección de datos.	SI
		Estimar el costo de desarrollo y realización de una solución.	SI
		Identificar beneficios (por ejemplo, mejorar la satisfacción del cliente, ROI, y el aumento de las ganancias).	SI
		Estimar gastos de operación.	SI
3. Determinar objetivos de minería de datos	**Objetivos del Data Warehouse**	Traducir las preguntas de negocio a objetivos de minería de datos (por ejemplo, una campaña de control de comercialización, requiere la segmentación de clientes para decidir a quién acercarse en esta campaña; el nivel/tamaño de los segmentos debería ser especificado).	SI
		Especificar datos tipo de problema de minería de datos (por ejemplo, la clasificación, la descripción, la predicción, y clustering). Para más detalles sobre tipos de problema de minería de datos, vea el Apéndice 2.	NO
	Criterios de éxitos del Data Warehouse	Especificar los criterios para evaluar el Data Warehouse (por ejemplo, la exactitud del modelo, el funcionamiento y la complejidad).	SI
		Definir el patrón de pruebas para los criterios de evaluación.	SI

		Especificar las reglas que dirigen criterios de evaluación subjetivos (por ejemplo, habilidad de explicar el modelo y los datos, y la comprensión de mercadeo proporcionada por el modelo).	SI
4. Producción del plan del proyecto	**Plan del Proyecto**	Definir el plan de proceso inicial y hablar de la viabilidad con todo el personal incluido	SI
		Combinar todos los objetivos identificados y técnicas seleccionadas en un procedimiento coherente, que solucione las cuestiones del negocio y encuentre los criterios de éxito de negocio	SI
		Estimar el esfuerzo y los recursos necesarios, para alcanzar y desarrollar la solución. (Es útil considerar la experiencia de otras personas estimando escalas de tiempo para proyectos de minería de datos. Por ejemplo, es a menudo presumido que el 50-70% del tiempo y el esfuerzo en un proyecto de minería de es usado en la Fase de Preparación de Datos, mientras que sólo un 20-30% es usado en la Fase de Comprensión de Datos, y sólo un 10-20% es gastado en cada una de las Fase de Modelado: Evaluación, y comprensión del Negocio, y el 5-10% en la Fase de Desarrollo.)	SI
		Identificar pasos críticos	SI
		Marcar los puntos de decisión	SI
		Marcar los puntos de revisión	SI
		Identificar las principales iteraciones	SI
	Evaluación Inicial de herramientas y técnicas	Crear una lista de criterios de selección para herramientas y técnicas (o usar uno existente si está disponible)	SI
		Escoger herramientas y técnicas posibles	SI
		Evaluar la adecuación de técnicas	SI
		Revisar y priorizar técnicas aplicables, según la evaluación de soluciones alternativas.	SI

Metodología CRISP Fase 2 - Comprensión de los Datos			
Punto	**Salida**	**Actividades**	**Realizada**
1. *Recolección de datos iniciales*	**Informe de la recolección de datos inicial**	Planear que información es necesaria (por ejemplo, sólo para atributos determinados, o la información adicional específica)	SI
		Comprobar si toda la información necesaria (para resolver los objetivos de la minería de datos) está en realidad disponible	SI
		Especificar los criterios de selección (por ejemplo, ¿Qué atributos son necesarios para los objetivos específicos de minería de datos? ¿Qué atributos han sido identificados como no pertinentes? ¿Cuántos atributos podemos manejar con las técnicas escogidas?)	SI
		Elegir tablas/archivos de interés	SI
		Elegir datos dentro de una tabla/archivo	SI
		Pensar cuanto tiempo de una historial habría que usar (por ejemplo, si 18 meses de datos están disponibles, sólo 12 meses pueden ser necesarios para el ejercicio)	SI
		Si los datos contienen libre entradas de texto, ¿tenemos que codificarlos para modelar o necesitamos agruparlos en entradas específicas?	SI
		¿Cómo podemos encontrar atributos omitidos?	SI
		¿Cómo podemos mejorar la extracción los datos?	SI
2. *Descripción de datos*	**Informe de descripción de datos / Actividades Análisis Volumétrico de datos**	Identificar datos y métodos de captura	SI
		Acceder a las fuentes de datos	SI
		Usar análisis estadísticos, si es apropiado	SI
		Reportar las tablas y sus relaciones	SI
		Compruebe el volumen de datos, el número de múltiplos y la complejidad	SI
		Notar si los datos contienen entradas de texto libres	SI
	Informe de descripción de datos / Atributo tipos y valores	Comprobar la accesibilidad, y disponibilidad de atributos	SI
		Comprobar los tipos de atributos (numérico, simbólico, la taxonomía, etc.)	SI
		Comprobar el rango de valores de los atributos	SI
		Analizar los atributos correlativos (correlaciones de atributo)	SI
		Comprender el significado de cada atributo y clasificar (describir) el valor en términos de negocio	SI
		Para cada atributo, calcular la estadística básica (por ejemplo, calcular la distribución, el promedio, el máximo, el mínimo, la desviación estándar, la varianza, la moda, la inclinación, etc.)	NO
		Analizar la estadística básica, y relacionar los resultados con su significado, en términos de negocio	NO

		Decidir si el atributo es relevante, para los objetivos específicos de la minería de datos	SI
		Determinar si el significado del atributo es usado coherentemente (conscientemente)	SI
		Entrevistar a expertos de dominio para obtener su opinión sobre la importancia de los atributos	SI
		Decidir si es necesario equilibrar los datos claves (basado en las técnicas que modelan a ser usado)	SI
		Analizar relaciones claves	SI
		Comprobar la cantidad de coincidencias entre valores de atributos claves, a través de tablas	SI
	Informe de descripción de datos / Revisión de Objetivos/Presunciones	Actualizar la lista de presunciones, si es necesario	SI
3. *Exploración de datos*	**Informe de exploración de datos**	Analizar en detalles las propiedades de atributos interesantes (por ejemplo, la estadística básica, las sub-poblaciones interesantes)	SI
		Identificar las características de las sub-poblaciones	SI
		Considerar y evaluar la información y conclusiones en el informe de descripciones de datos	SI
		Formar una hipótesis e identificar acciones	SI
		Transformar la hipótesis en un objetivo de minería de datos, si es posible	SI
		Aclarar objetivos de minería de datos, o hacerlos más exactos. Una búsqueda "ciega" no es necesariamente inútil, pero una búsqueda más dirigida hacia objetivos de negocio es preferible.	SI
		Realizar un análisis básico para verificar la hipótesis	SI
4. *Verificación de la calidad de datos*	**Informe de calidad de datos**	Identificar valores especiales y catalogar su significado	SI
		Comprobar la cobertura (por ejemplo, si todos los valores posibles son representados)	SI
		Comprobar las claves	SI
		Verificar que los significados de los atributos y valores contenidos, se satisfacen simultáneamente	SI
		Identificar atributos omitidos y campos en blanco	SI
		Establecer el significado de datos que faltan o fallan	SI
		Comprobar los atributos con los valores diferentes que tienen significados similares (por ejemplo, la grasa baja, la dieta)	SI
		Comprobar la ortografía y el formato de valores (por ejemplo, mismo valor pero a veces comienza con una letra minúscula, a veces con una letra mayúscula)	SI
		Comprobar las desviaciones, y decidir si una desviación es "ruido" o puede indicar un fenómeno interesante	SI

		Comprobar la plausibilidad de valores, (por ejemplo, todos los campos que tienen el mismo o casi los mismos valores)	SI
		Si los datos son almacenados en archivos planos, comprobar que delimitador es usado y si esto es usado coherentemente en todos los atributos	NO
		Si los datos son almacenados en archivos planos, comprobar el número de campos en cada registro para ver si ellos coinciden	NO
		Comprobar consistencia y superabundancia entre fuentes diferentes	SI
		Planear para tratar el ruido	SI
		Descubrir el tipo de ruido, y que atributos son afectados	SI

Metodología CRISP Fase 3 - Preparación de los Datos			
Tarea	**Salida**	**Actividades**	**Realizada**
1. *Datos seleccionados*	**Razonamiento para inclusión/exclusión**	Recoger datos adicionales apropiados (de diferentes fuentes - internos así como externos)	SI
		Realizar las pruebas de importancia y correlación, para decidir si los campos son incluidos	SI
		Reconsiderar Criterios de Selección de Datos (Vea la Tarea 2.1) a la luz de las experiencias de calidad de los datos y en la exploración de datos (esto es, puede desear incluir/excluir otros juegos de datos)	SI
		Reconsiderar Criterios de Selección de Datos (Vea la Tarea 2.1) a la luz de experiencia de modelado (esto es, la evaluación del modelo puede mostrar que otros conjuntos de datos son necesarios)	SI
		Seleccionar diferentes subconjuntos de datos (por ejemplo, atributos diferentes, sólo los datos que encuentran ciertas condiciones)	SI
		Considerar el uso de técnicas de muestreo (por ejemplo, una solución rápida puede implicar la prueba dura, y el entrenamiento del conjunto de datos o la reducción del tamaño de la conjunto de datos de prueba, si la herramienta no puede manejar conjunto de datos llenos. Esto puede también ser útil para tener muestras ponderadas, para dar la distinta importancia a atributos diferentes o valores diferentes del mismo atributo.)	SI
		Documentar el razonamiento para la inclusión/exclusión	SI
		Comprobar técnicas disponibles para el muestreo de datos	SI
2. *Limpieza de datos*	**Informe de la limpieza de datos**	Reconsiderar como tratar con cualquier tipo de ruido observado	SI
		Corregir, remover, o ignorar el ruido	SI
		Decidir cómo tratar con valores especiales y su significado. El área de valores especiales puede dar lugar a muchos resultados extraños, y ser examinados con cuidado. Los ejemplos de valores especiales podrían surgir por los resultados tomados de una revisión donde algunas cuestiones no fueron preguntadas o no fueron contestadas. Esto podría terminar en un valor de 99 para datos desconocidos. Por ejemplo, 99 para estado civil o afiliación política. Los valores especiales también podría surgir cuando los datos son truncados por ejemplo., 00 para gente de 100 años o para todos los coches con 100,000 kilómetros en el odómetro.	SI
		Reconsiderar Criterios de Selección de Datos (Vea la Tarea 2.1) a la luz de las experiencias de los datos limpiados (esto es, si puede o desea incluir/excluir otros conjuntos de datos).	SI
3. *Construcción de datos*	**Registros generados / Construir datos**	Comprobar los mecanismos de construcción disponibles con la lista de herramientas sugeridas para el proyecto	SI

		Decidir si esto es lo mejor para realizar la construcción dentro de la herramienta o fuera de ella (esto es, que es más eficiente, exacto, repetible)	SI
		Reconsiderar Criterios de Selección de Datos (Vea la Tarea 2.1) a la luz de las experiencias de construcción de datos (puede incluir/excluir otros conjuntos de datos)	SI
	Registros generados / Atributos derivados	Decidir si cualquier atributo puede ser normalizado (por ejemplo, usando un algoritmo de agrupamiento (clustering) con el periodo y el ingreso, en ciertas divisas, el ingreso se controlará)	SI
		Considerar agregar nueva información sobre la importancia de los atributos, para agregar nuevos atributos (Por ejemplo, atributos con peso, normalización ponderada)	SI
		¿Cómo se puede construir o imputar atributos faltantes? Decidir el tipo de construcción (por ejemplo, la combinación, el promedio, la inducción).	SI
		Agregar atributos nuevos a los datos de acceso	SI
	Registros generados / Transformaciones de atributo individual	Especificar los pasos de transformaciones necesarias en términos de facilitar las transformación disponibles (por ejemplo, cambiar un binning de un atributo numérico)	SI
		Realizar pasos de transformación	SI
		Comprobar por técnicas disponibles si es necesario (por ejemplo, mecanismos para construir prototipos, para cada segmento de datos segmentados).	SI
4. *Integración de datos*	**Datos combinados**	Comprobar si las aplicaciones de integración son capaces de integrar las fuentes de entrada, como se requiere	SI
		Integrar fuentes y resultados almacenados	SI
		Reconsiderar Criterios de Selección de Datos (Vea la Tarea 2.1) a la luz de las experiencias de integración de datos (incluir/excluir otros conjuntos de datos)	SI
5. *Formateo de datos*	**Datos Formateados**	Reorganizar los atributos	SI
		Reformatear valores internos	SI
		Estos son cambios puramente sintácticos hechos para satisfacer las exigencias de la herramienta específica de modelado	SI
		Reconsiderar Criterios de Selección de Datos (Vea la Tarea 2.1) a la luz de las experiencias de limpieza de datos (incluir/excluir otros conjuntos de datos)	SI

Metodología CRISP Fase 4 – Modelado			
Punto	**Salida**	**Actividades**	**Realizada**
1. *Seleccionar técnicas de modelado*	**Presunciones de modelado**	Definir cualquier presunción construida por la técnica sobre los datos (por ejemplo, la calidad, el formato, la distribución)	SI
		Comparar estas presunciones con las del Informe de Descripción de Datos.	SI
		Asegurarse que estas presunciones se sostienen, y volver a la Fase de Preparación de Datos, si es necesario	SI
2. *Generar el diseño de prueba*	**Diseño de Prueba**	Comprobar que existe diseños de prueba, separadamente para cada objetivo de minería de datos	NO
		Decidir los pasos necesarios (el número de iteraciones, el número de desviaciones o curvas, etc.)	NO
		Preparar los datos requeridos, para la prueba	SI
3. *Construcción del modelo*	**Parámetros de ajuste**	Determinar los parámetros iniciales	NO
		Documentar las razones para elegir aquellos valores	NO
	Modelos	Ejecutar la técnica seleccionada sobre el conjunto de datos de entrada, para producir el modelo	NO
		Post-procesar los resultados de minería de datos (por ejemplo, editar reglas, mostrar árboles)	NO
	Descripción del modelo	Describir cualquier características del modelo actual que puede ser útil para el futuro	NO
		Ajustar parámetro de entorno (de registro) usado para producir el modelo	NO
		Dar una descripción detallada del modelo, y cualquier rasgo especial	NO
		Para modelos basados en reglas, listar las reglas producidas, más cualquier evaluación de cada regla, o la exactitud y alcance total del modelo	NO
		Para modelos no transparentes, listar cualquier información técnica sobre el modelo (como la topología de las redes neuronales) y cualquier descripción de comportamiento producido por el proceso de modelado (como la exactitud o la sensibilidad)	NO
		Describir el comportamiento del modelo y la interpretación	NO
		Expresar conclusiones respecto a los patrones en los datos (si hay alguno); a veces el modelo revela hechos importantes sobre los datos sin un proceso de evaluación separado (por ejemplo, que la salida o la conclusión son duplicadas en una de las entradas)	NO
4. *Evaluación del*	**Evaluación del modelo**	Evaluar los resultados en lo que concierne a criterios de evaluación	SI

modelo		Probar los resultados según una estrategia de prueba (por ejemplo: Corrida y Prueba, Validación cruzada, bootstrapping, etc.)	SI
		Comparar los resultados de la evaluación y la interpretación	NO
		Crear la clasificación de resultados en lo concerniente a criterios de éxito y evaluación	SI
		Seleccionar los mejores modelos	NO
		Interpretar los resultados en términos de negocio (tanto como sea posible en esta etapa)	NO
		Conseguir comentarios de los modelos por expertos en datos o en el dominio	NO
		Chequear la credibilidad del modelo	SI
		Comprobar los efectos sobre los objetivos de minería de datos	NO
		Comprobar los modelos con una base de conocimiento determinada, para ver si la información descubierta es nueva y útil	NO
		Comprobar la fiabilidad de los resultados	SI
		Analizar el potencial para el desarrollo de cada resultado	NO
		Si hay una descripción verbal del modelo generado (por ejemplo, en forma de reglas), evaluar las reglas: ¿Son lógicas, o son factibles, hay demasiadas reglas o hay demasiado pocas, violan el sentido común?	NO
		Evaluar resultados	SI
		Conseguir ideas específicas de cada técnica de modelado, y ciertos parámetros de ajustes que conduzcan a resultados buenos/malos	NO
	Revisación de parámetros de ajuste	Ajustar parámetros para producir mejores modelos.	NO

Metodología CRISP Fase 5 – Evaluación			
Punto	**Salida**	**Actividades**	**Realizada**
1. *Evaluación de los resultados*	**Evaluación de los resultados del Data Warehouse en lo que respecta a criterios de éxito de negocio**	Comprender los resultados de la minería de datos	SI
		Interpretar los resultados en términos de la aplicación (del uso)	SI
		Comprobar efectos sobre los objetivos de minería de datos	SI
		Comprobar los resultados de minería de datos con la base de un conocimiento determinado para ver si la información descubierta es nueva y útil	SI
		Evaluar y estimar los resultados en lo que respecta a criterios de éxito de negocio (esto es, el proyecto ha alcanzado los Objetivos de Negocio originales)	SI
		Comparar los resultados de la evaluación y la interpretación	SI
		Clasificar los resultados en lo que respecta a criterios de éxito de negocio	SI
		Comprobar el efecto de los resultados sobre el objetivo (fin) de la aplicación inicial	SI
		Determinar si hay nuevos objetivos de negocio, para ser dirigidos más tarde en el proyecto, o en nuevos proyectos	SI
		Expresar recomendaciones.	SI
	Modelo aprobado	Aprobar el modelo del Data Warehouse	SI
2. *Proceso de revisión*	**Revisión de procesos**	Proporcionar una descripción del proceso de minería de datos usado	SI
		Analizar el proceso de minería de datos. Para cada etapa del proceso preguntar: ¿Esto fue necesario?, ¿Esto fue ejecutado óptimamente?, ¿En qué modo podría mejorar?	SI
		Identificar fracasos	SI
		Identificar pasos desviados (de engaños)	SI
		Identificar acciones alternativas posibles y/o caminos inesperados en el proceso	SI
		Revisar resultados de minería de datos en lo que concierne a criterios de éxito de negocio	SI
3. *Determinación de los próximos pasos*	**Lista de acciones posibles**	Analizar e potencial para el desarrollo de cada resultado	SI
		Estimar el potencial para la mejora de proceso actual	SI
		Comprobar los recursos restantes para determinar si ellos permiten iteraciones de proceso adicionales (o si recursos adicionales pueden estar siendo disponibles)	NO
		Recomendar continuar con las alternativas	SI
		Refinar el plan de proceso	SI
	Decisión	Clasificar las acciones posibles	SI
		Seleccionar una de las acciones posibles	SI
		Documentar las razones para la elección	SI

Metodología CRISP Fase 6 - Implementación			
Punto	**Salida**	**Actividades**	**Realizada**
1. *Plan de implementación*	**Plan de Implementación**	Resumir resultados desarrollados	SI
		Construir y evaluar los planes alternativos para el desarrollo	SI
		Decidir para cada resultado de conocimiento o información distinto	NO
		Determinar como el conocimiento o la información serán propagados (generados) a los usuarios	SI
		Decidir cómo será supervisado el uso del resultado y medido sus beneficios (donde sea aplicable)	NO
		Decidir por cada resultado de modelo desarrollado o de software	NO
		Establecer como el modelo o software que serán desplegados dentro de los sistemas de la organización	NO
		Determinar cómo su empleo será supervisado y medido sus beneficios (donde sea aplicable)	NO
		Identificar posibles problemas durante el desarrollo (peligros a ser evitados)	SI
2. *Supervisión y mantenimiento del plan*	**Plan de supervisión y mantenimiento**	Comprobar aspectos dinámicos (esto es, ¿qué cosas podrían cambiar en el entorno?)	SI
		Decidir cómo será supervisada la precisión	SI
		Determinar cuando el resultado de minería de datos o el modelo no deberían ser usados. Identificar criterios (la validez, el límite de la exactitud, nuevos datos, cambios en el dominio de aplicación, etc.), y qué debería pasar si el modelo o el resultado no pueden ser usados. (Actualización del modelo, establecimiento de nuevos proyectos de minería de datos, etc.).	SI
		¿Cambiarán con el tiempo los objetivos de negocio del uso empleo del modelo? Documentar totalmente el problema inicial que el modelo intentaba solucionar.	SI
		Desarrollar el plan de mantenimiento y la supervisión.	SI
3. *Producción de Informe final*	**Informe Final**	Identificar cuáles informes son necesarios (presentación de diapositiva, conclusiones de administración, detalles encontrados, explicación de los modelos, etc.)	NO
		Analizar que tan bien se han encontrado los objetivos de minería de datos iniciales	NO
		Identificar grupos de objetivos para el informe	NO
		Describir en forma general las estructuras y el contenido de informe(s)	SI
		Seleccionar conclusiones para ser incluidas en los informes	SI
	Presentación Final	Decidir el grupo objetivo para la presentación final, y determinar si ellos ya habrán recibido el informe definitivo	SI
		Seleccionar cuáles de los artículos del informe definitivo deberían ser incluidos	SI

		en la presentación final	
4. *Revisión del proyecto*	**Documentación de Experiencia**	Entrevistar a toda la gente significativa involucrada en el proyecto y preguntarles sobre su experiencia durante el proyecto	NO
		Si los usuarios finales trabajan en el negocio con los resultados de minería de datos, entrevistarlos: ¿Están satisfechos? ¿Cómo podría haber sido mejor realizado? ¿Necesitan apoyo adicional?	NO
		Resumir la realimentación y escribir la documentación de experiencia	SI
		Analizar el proceso (las cosas que se trabajaron bien, los errores producidos, las lecciones aprendidas, etc.)	SI
		Documentar el proceso de minería de datos específico (¿Cómo pueden los resultados y la experiencia de aplicación del modelo ser realimentado en el proceso?)	SI
		Generalizar desde los detalles para producir la experiencia útil para proyectos futuros	SI