Universidad Cenfotec

Maestría en Tecnologías de Bases de Datos

Documento final de Proyecto de Investigación Aplicada 2

Propuesta de una metodología para ofrecer el servicio de ciencia de datos en la PYMES GO-LABS ENTERPRISES

Estudiantes:

Gutiérrez Cerdas Alexander

Jiménez Delgado Efrén Antonio

Octubre, 2017

Gracias a todas las personas que me han impulsado para llegar hasta este punto de mi carrera académica y en especial a mis padres por la compresión, paciencia y su disposición por ayudar en todo momento de mi vida. Este logro es principalmente a mi abuela Mercedes que ha estado los últimos años de mi vida alentándome a crecer en todos los sentidos, y a estar pendiente de mi bienestar, hasta en esas noches de vela para finalizar esta bonita experiencia de mi vida. A mis maestros que me apoyaron en todos estos años con recomendaciones, consejos, experiencias y tiempo invertido para mejorar mis conocimientos y terminar de la mejor manera este trabajo.

Efrén Jiménez Delgado

Gracias a las personas que de una u otra forma han colaborado para poder llegar a este momento en mi vida y en especial a mi familia que se ha visto sacrificada de no poder compartir muchas vivencias. El lograr este objetivo es gracias a mis abuelitos, ya que ellos inculcaron en mí ese deseo de lucha y entusiasmo por cada día ser mejor, siendo la educación un pilar fundamental para llegar a ello. A Marcela Cabalceta y doña Ligia Cabalceta que fueron las personas que me ayudaron a tomar este reto y oportunidad de mejorar. También a los profesores que en el transcurrir del tiempo me enseñaron, dieron recomendaciones y disciplinaron con el objetivo de que crecieran mis conocimientos y realizara de muy buena manera este trabajo. Asimismo, al gobierno de Costa Rica por impulsar la educación y brindar una oportunidad al conocimiento por medio de Instituciones como MICITT.

Alexander Gutiérrez Cerdas

Tribunal examinador

1 Contenido

Índice de figuras

Índice de tablas

Resumen

El presente proyecto tiene como objetivo la propuesta de una metodología para ofrecer ciencia de los datos en una PYMES GO-LABS ENTERPRISES la cual se dedica a la industria de desarrollo de software. Esta metodología permitirá que las organizaciones se introduzcan al análisis de datos de una forma más natural, concisa y económica.

Con el logro de los objetivos se desea crear una guía para el desarrollo e implementación de un departamento de ciencia de los datos. Además, se pretende que dicha metodología abarque la definición de roles y funciones, definición de tareas, identificación y definición de conceptos, implantación de práctica de ciencia de datos, recomendación de software y sinergia con otras metodologías existentes en el mercado.

Considerando que el mercado de esta PYME presenta una necesidad real de conocimiento y aplicación de la ciencia de los datos en sus repositorios de datos. Ademas, se pretende que la siguiente herramienta pueda brindar un panorama más claro sobre la introducción, desarrollo y madurez en la aplicación de ciencia de los datos.

Palabras Clave: ciencia de datos, PYME, metodología, inteligencia de negocios, repositorio de datos, guía, departamento base de datos.

1 Capítulo I

1.1 Introducción

En la actualidad, la manera cómo se gestiona la información en las empresas es una herramienta clave para poder sobrevivir en un mercado cambiante, dinámico y global, es por esta razón que se han venido desarrollando un conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos de las PYMES. Dichas organizaciones disponen, como todas las empresas, sin importar su tamaño, de sistemas de información que contemplan pequeños, medianos y gigantes repositorios de información que son convenientemente importantes de analizar y optimizar.

Existen metodologías, herramientas y mejores prácticas en el mercado para desarrollar ciencia de datos en las empresas. El propósito de este documento es analizar algunas de estas y así considerar un modelo que sirva como referencia a la organización en el momento de abordar la implementación de un proyecto de ciencia de datos

Se llevarán a cabo actividades con el fin de alcanzar los objetivos del proyecto, los cuales van desde contextualizarse hasta documentar los resultados de implementación de ciencia de datos en la PYMES GO-LABS ENTERPRISES.

1.2 Generalidades

Este proyecto será realizado en la PYMES GO-LABS ENTERPRISES, la cual es una empresa dedicada al desarrollo de software en ambientes móviles, web y escritorio. También podría aplicarse en aquellas PYMES que hayan implementado pequeños paquetes informático o sistemas más complejos como (sistema de planificación de recursos empresariales o ERP, Gestión de Servicio al Cliente o CRM), o bien hayan desarrollado sus aplicaciones a medida con el fin de obtener una optimización de los recursos con que se cuentan para maximizar sus ganancias.

1.3 Antecedentes del Problema

Las PYMES en Costa Rica cada día se enfrentan a nuevos retos de competitividad, ya sea en el mercado interno o en el exterior del país. Muchas de estas empresas han intentado implementar ciencia de datos, pero han sido un fracaso debido al poco conocimiento sobre esta área y lo que esto conlleva, se hacen inversiones que al final de cuenta no son lo que se esperaba y se les hace imposible obtener beneficios por lo que muchas prefieren no hacerlo y no aprovechar el conocimiento que se puede obtener de los datos.

Pero debido a que es necesario el poder aplicar ciencia de datos, en GO-LABS ENTERPRISES se han interesado sobre la necesidad del mercado costarricense de una guía o metodología que ayude al correcto desarrollo de ciencia de datos, debido a que el conocimiento que se puede adquirir da mucho valor agregado y es muy caro; las grandes compañías son las que pueden realizar ciencia de datos por lo que es evidente la ventaja que reciben de este conocimiento abstraído de los datos, dejando en una posición vulnerable a las PYMES a la hora de competir en el mercado.

1.4 Definición y Descripción del Problema

La preocupación por las PYMES no es un tema nuevo, pero ha sido de gran importancia desde hace muchos años atrás, el Gobierno ha realizado esfuerzos para apoyar este sector de la economía nacional la cual para el 2015 aportaba un cuarto del porcentaje a la economía nacional según el Informe del Estado de la Nación del 2015, esto se ha venido logrando gracias a la mejora en la competitividad de las PYMES tanto en el mercado interno como externo.

Para mejorar la competitividad en las PYMES se deben utilizar estrategias de crecimiento con el apoyo de tecnología, ya que el uso de herramientas es fundamental para proveer una ventaja competitiva como lo son las ciencias de los datos, las cuales permitirán impulsar y optimizar el uso de los recursos con los que se cuenta.

Es de vital importancia que la PYMES GO-LABS ENTERPRISES en estudio adquiera servicios profesionales para explorar los datos y aplicar ciencia de los datos

para poder conocer el comportamiento del negocio en el que se encuentra involucrado y como se pueden obtener mejores rendimientos, hacer proyecciones y realizar estrategias con proyecciones de bajo nivel de riesgo. Ante esta situación es necesario que la PYMES GO-LABS ENTERPRISES cuente con orientación tecnológica de cómo aplicar ciencia de datos, para así aprovechar de manera idónea la tecnología y así minimizar el riesgo de fracasos a la hora de buscar una mejora en la competitividad por medio de la tecnología.

1.5 Justificación

Para solventar la necesidad de la PYME costarricense GO-LABS ENTERPRISES sobre la aplicación de la ciencia de los datos, se pretende promover una metodología que abarque la definición de roles, funciones, tareas, identificación, conceptos, recomendaciones y sinergia con otras metodologías existentes en el mercado. Esto con el fin de que la PYMES pueda solventar de la manera más eficiente, económica y ágil la implementación de la ciencia de los datos en sus repositorios de datos.

La necesidad creciente de ciencia de los datos en las organizaciones ha hecho que compañías, con una gran cantidad de recursos puedan acceder a una comprensión más expedita y profunda sobre los datos, lo cual les brinda una ventaja competitiva y amplia en contra de las PYMES, donde estas viven en una amenaza latente sobre el desconocimiento de la información y patrones existentes en sus datos. Aunque las PYMES conocen e identifican las características positivas de la ciencia de los datos, estas no pueden obtener este conocimiento, debido a que es económicamente elevado, y en el mercado costarricense no existen muchas opciones para llevar a cabo este tipo de proyecto en estas organizaciones.

Se pretende que con la implementación de la metodología propuesta en este trabajo la PYMES costarricense GO-LABS ENTERPRISES pueda iniciar, desarrollarse y madurar en el tema de las ciencias de los datos y que con esto pueda enrumbar su futuro con ventajas competitivas en el mercado. Además, de poder cerrar un poco la brecha entre las grandes compañías y PYMES con el tema referente a la aplicación de la ciencia de los datos.

1.6Viabilidad

El proyecto es una necesidad presentada por la PYMES GO-LABS ENTERPRISES del sector productivo de tecnologías de información de San Carlos de Alajuela, Costa Rica, los cuales están en disposición de colaborar para poner en práctica el uso de la metodología que se desarrollará para la implementación de ciencia de los datos. Esta investigación consiste en la aplicación de todos los conocimientos adquiridos a lo largo de los estudios de postgrados de los investigadores, en donde se podrán en ejecución buenas prácticas y herramientas de software para la implementación de ciencia de datos.

1.6.1 Punto de Vista Técnico

Para el desarrollo de este proyecto se ha recibido la formación necesaria, acompañado de investigaciones sobre temas relacionados y conocimientos adquiridos a lo largo del tiempo en las organizaciones donde nos desempeñamos para poder tener criterio de discernir y recomendar lo que se considere correcto e idóneo para la implementación de ciencia de datos en PYMES GO-LABS ENTERPRISES. Además de ello se cuenta con un tutor que nos orienta en el desarrollo del proyecto.

1.6.2 Punto de Vista Operativo

La propuesta de la metodología tiene el aval por parte de la PYMES GO-LABS ENTERPRISES que servirá como promotor y aplicador de la guía final. La cual será el resultado del trabajo realizado en este documento. Por lo que se cuenta con la disposición del personal necesario, recursos físicos y económicos. Esto a raíz de que es de gran interés para la PYMES GO-LABS ENTERPRISES la conclusión satisfactoria de esta metodología.

1.6.3 Punto de Vista Económico

La propuesta de metodología cuenta con el apoyo económico y de recursos humanos de la PYMES GO-LABS ENTERPRISES, la cual propone la disposición de

los recursos antes mencionados, en el momento que lo solicite cualquier miembro del grupo investigador.

La compañía GO-LABS ENTERPRISES dispondrá de un tiempo aproximado de 4 horas diarias por día hábil, es decir un aproximado a 20 horas semanales por investigador, para dar un total aproximado de 360 horas por investigador. Esto tiene un costo aproximado de $20 por hora, para un total de $6400 por investigador. Además, pondrá a disposición al gerente Carlos Rojas con 40 horas y un costo de $1600 y a Jennifer Madrigal como administrativa con 20 horas y un costo total de $400.

1.7 Objetivos

Se ha usado la taxonomía original de Bloom de 1956 para plantear los objetivos de esta investigación, debido a su aceptación por el sistema educativo costarricense.

1.7.1 Objetivo General

Proponer una nueva metodología, por medio de la identificación de componentes, herramientas y modelos, para el servicio de ciencia de datos de la PYMES GO-LABS ENTERPRISES

1.7.2 Objetivos Específicos

- Identificar los componentes del modelo metodológico para aplicar ciencia de datos para la PYMES GO-LABS ENTERPRISES.
- Describir los componentes del modelo en la metodología de ciencia de datos para la PYMES GO-LABS ENTERPRISES.
- Escoger las herramientas de software que se deben usar en cada componente del modelo de ciencia de datos para la PYMES GO-LABS ENTERPRISES.
- Elaborar una guía de la implementación de los componentes del modelo de ciencia de datos para la PYMES GO-LABS ENTERPRISES.
- Evidenciar cuáles son los resultados de la implementación de la metodología encontrados en el modelo de ciencia de datos para la PYMES GO-LABS ENTERPRISES.

A continuación, en la tabla 1 se describe como son abarcados los objetivos específicos de este proyecto.

Tabla 1 ¿Como son abarcados los objetivos?

Objetivo	Referencia
1. Identificar los componentes del modelo metodológico para aplicar ciencia de datos para la PYMES GO-LABS ENTERPRISES.	Este objetivo es abarcado en el capítulo III, donde se identifican y describen los componentes de la ciencia de datos.
2. Describir los componentes del modelo en la metodología de ciencia de datos para la PYMES GO-LABS ENTERPRISES.	El objetivo está desarrollado en el capítulo III, donde se explican los componentes de la ciencia de datos, además se complementa en el capítulo IV en donde se describen como se usan y en el capítulo V se realiza la implementación.
3. Escoger las herramientas de software que se deben usar en cada componente del modelo de ciencia de datos para la PYMES GO-LABS ENTERPRISES.	El objetivo se abarca en el capítulo IV y V en la fase de desarrollo donde se hace un estudio de las principales herramientas y se implementa su uso.
4. Elaborar una guía de la implementación de los componentes del modelo de ciencia de datos para la PYMES GO-LABS ENTERPRISES.	Este objetivo se desarrolla en el capítulo IV, donde se definen las fases que serán la guía que ayudará a la aplicación de ciencia de datos en la Pyme Go-Labs.
5. Evidenciar cuáles son los resultados de la implementación de la metodología encontrados en el modelo de ciencia de datos para la PYMES GO-LABS ENTERPRISES.	Es abarcado en el capítulo V donde se realiza la aplicación de la metodología y específicamente en la fase de validación se muestran evidencias de la retroalimentación de la Pyme Go-Labs.

1.8 Alcances y Limitaciones

1.8.1 Alcances

Al finalizar este proyecto se obtendrá como resultado una nueva guía metodológica en donde se hará una definición de roles y funciones, definición de tareas, identificación y definición de conceptos, implantación de práctica de ciencia de datos, recomendación de software y sinergia con otras metodologías existentes en el mercado, que serán aprovechadas por la PYMES GO-LABS ENTERPRISES.

1.8.2 Limitaciones

A la PYMES GO-LABS ENTERPRISES no se le entregará la metodología, sino solo el servicio, además las licencias para el software por utilizar deberá ser adquirido por la organización.

Otra de las limitantes es el tiempo de elaboración del proyecto, ya que se cuenta con un periodo establecido por la Universidad Cenfotec. Además, existe otra limitación la cual es que los desarrolladores de la metodología no estarán a tiempo completo, sino será compartido con las tareas laborales que ellos realizan.

Por otro lado, se cuenta con otra limitante la cual es que la puesta en práctica de la nueva metodología dependerá de los recursos y disposición con que cuente la PYME Go-Labs.

1.9 Marco de Referencia Organizacional y Socioeconómico

1.9.1 Historia

Las PYMES a lo largo del tiempo han tenido un impacto grande en la generación de empleo, en el proceso de cadenas productivas en especial en el sector manufacturero y en la definición del modelo económico de la sociedad costarricense.

Pese a ello existe una diferencia importante en el modelo en países desarrollados, donde las PYMES son empresas innovadoras ligadas al sector moderno, del modelo implementado en nuestro país donde la mayoría existen como una forma de subsistencia vinculada a la economía informal (que no pagan impuestos, ni cargas sociales).

Las PYMES en cualquier país del mundo se concentran en sectores tradicionales que tienen bajas barreras de entrada y mínimos niveles de producción, las cuales hacen uso intensivo de la mano de obra; estos factores incrementan las ventajas relativas a la producción en pequeñas escalas.

En nuestro país la PYMES GO-LABS ENTERPRISES no se ha mantenido al margen del progreso tecnológico y está innovando en uso de tecnologías de información para mejorar los procesos productivos y los métodos de organización del trabajo. Así, en términos generales, a la mayoría de las PYMES les hace falta actualizarse para competir globalmente en forma efectiva y vincularse con las tendencias a futuro del comportamiento de los mercados locales y foráneos; además muy pocas poseen herramientas adecuadas para una efectiva toma de decisiones en situaciones de alto riesgo.

1.9.2 Tipo de Negocio y Mercado Meta

Las PYMES son un conjunto de pequeñas y medianas empresas que, de acuerdo con su volumen de ventas, capital social, cantidad de trabajadores, y su nivel

de producción o activos presentan características propias de este tipo de entidades económicas.

Este tipo de organización acostumbra tener perfiles fáciles de adaptar hacia nuevas inversiones; por lo que están dispuestos a probar nuevas herramientas y tecnologías en sus empresas. La mayoría de los casos de éxito de este tipo de organizaciones afirman que la tecnología ha mejorado su capacidad de innovación y conocimiento hasta el punto de crear modelos de negocio digitales completamente nuevos.

Uno de los mayores retos para aquellos que toman las decisiones dentro de las PYMES es conseguir un equipo eficiente y con las cualidades adecuadas para potenciar el crecimiento del negocio. En este ámbito el buen uso de la tecnología y el correcto análisis de datos también pueden ayudar a desarrollar habilidades para el equipo, además de facilitar la toma de decisiones inteligentes.

El mercado que se desea abarcar es la PYMES GO-LABS ENTERPRISES de tecnología en la zona norte de Costa Rica que están presentes en el marco de la Cámara de empresas de tecnología de la zona norte (CETIC-ZN) y que sirva como base para otras de esta misma área.

2 Capítulo II

2.1 Estado de la Cuestión

La ciencia de los datos es un concepto bastante complejo y moderno, de hecho, no existe una definición de consenso, sino que difieren en variadas fuentes, es por tal motivo que para el presente proyecto se tomará el concepto de ciencia de los datos como la extracción de conocimiento a partir de información de datos.

La ciencia de los datos en los últimos años ha tomado una mayor relevancia, debido a que en las organizaciones se ha buscado la forma de aprovechar al máximo los datos para obtener conocimiento, lo cual da como resultado el éxito en estas organizaciones.

Ciencia de los datos es un concepto que involucra muchos campos, incorpora diferentes elementos y se basa en las técnicas y teorías de muchas áreas, como lo es las matemáticas, estadística, ingeniería de datos, reconocimiento de patrones y aprendizaje automático, computación avanzada, visualización, modelado de datos, almacenamiento de datos y la informática de alto rendimiento con el objetivo de extraer el significado de datos y la creación de productos a partir de datos.

La definición de este concepto es relativamente nuevo que se utiliza a menudo de manera intercambiable con inteligencia o análisis de negocios competitivos. La ciencia de datos busca utilizar todos los datos disponibles y relevantes para contar efectivamente una historia que pueda ser fácilmente comprendido sin necesidad de ser un especialista de la ciencia de los datos.

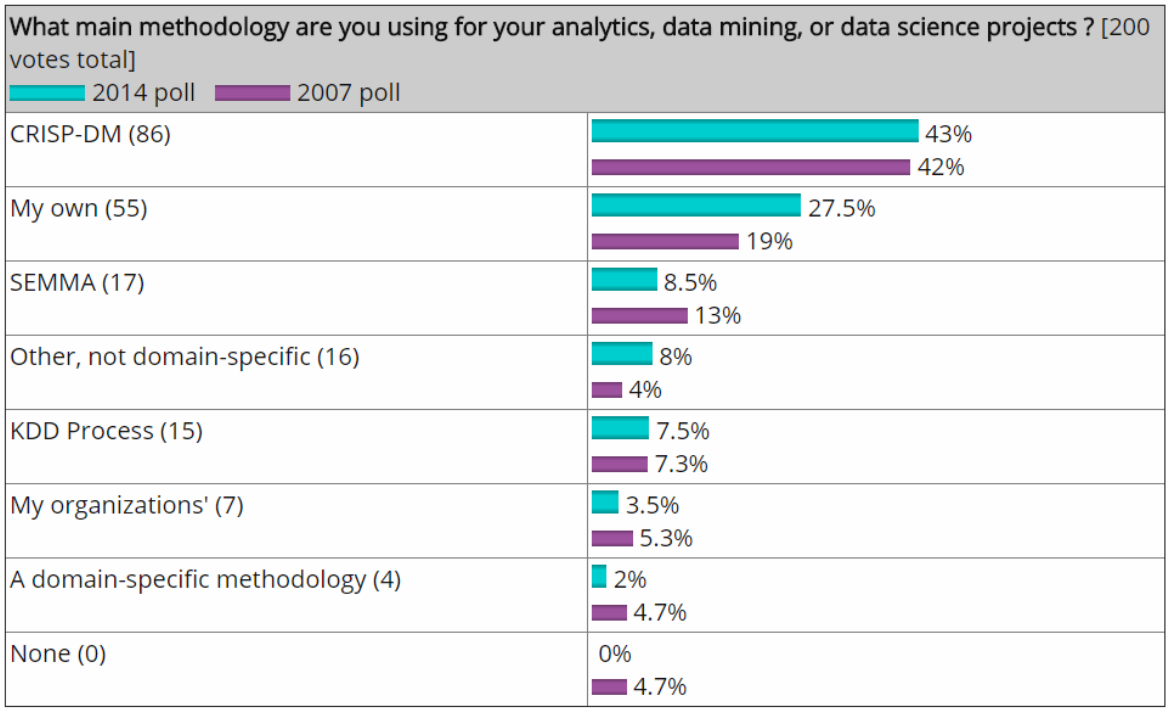
Actualmente existen variadas metodologías referentes a distintos campos de la ciencia de los datos, pero es muy difícil que estas estén relacionadas entre sí, lo cual representa un problema para los implementadores de esta ciencia aplicada en el campo de los negocios. Además, estas metodologías emergentes (véase Figura 1) en rápida evolución buscan su propia solución hacia la construcción de un modelo de desarrollo de ciencia de los datos. Por esto, los científicos de datos necesitan una

respuesta funcional capaz de proporcionar una orientación que realice una sinergia entre los mejores modelos de metodologías propuestas en el mercado. (kdnuggets, 2014)

Distribución regional de los votantes:

- US/Canadá, 45.5%
- Europa, 28.5%
- Asia, 14%
- América Latina, 9.5%
- Otros, 2.5%

Figura 1 Cuál es la principal metodología que es usada en proyectos para analítica, minería de datos o ciencia de datos del 2014 al 2017 (total de votos 200)



Fuente: (kdnuggets, 2014)

Como describe Torsten Priebe y Stefan Markus en su artículo “Business información modeling: A methodology for data-intensive projects, data science and big

data governance” (BIM) es un gran reto para las organizaciones la implementación exitosa de un proyecto de inteligencia de negocios sin una metodología asociada, dado que esto conllevaría a problemas incrementales en el ciclo de vida y consolidación de un proyecto de inteligencia de negocios. Una metodología facilita la armonía entre los procesos cíclicos del modelado y rentabilidad, que debe existir para concluir exitosamente un plan de proyecto de inteligencia de datos.

El uso de la metodología CRISP-DM en la aplicación de proyecto de minería de datos es la más aceptada en el mercado según el libro “Step-by-step data mining guide” escrito por Pete Chapman, el cual realiza una valoración con casos de estudio interesantes sobre la implementación de esta metodología y como los proyectos se deben adaptar a los seis pasos propuestos, para una culminación exitosa del proceso de minería.

El concepto de SCRUM es muy empleado en las empresas para el desarrollo de proyecto ágiles, como se muestra en el trabajo de investigación bajo el nombre “How to Make Business Intelligence Agile: The Agile BI Actions Catalog” (Krawatzeck, Barbara, & Duc Anh, 2015), el cual explica cómo llevar a cabo proyectos de inteligencia de negocios con la metodología SCRUM. Y además cataloga todos los pasos a seguir para implementar dicha metodología en un departamento, lo cual está muy alineado a los conceptos propuestos en este trabajo de investigación.

Según un artículo de IBM, el título profesional de “data scientist” o científico de datos surgió paralelamente con el aumento y difusión de la nueva tecnología del Big Data. Aunque no está ligado exclusivamente a proyectos de Big Data, el papel del científico de datos complementa dicha tecnología debido a la amplitud y profundidad de los datos que se examinan, en comparación con los roles tradicionales. Este mismo artículo afirma que un científico de datos representa una evolución del papel de analista de negocios o de datos. Mientras que un analista de datos tradicional investiga los datos de una sola fuente, por ejemplo, un sistema de CRM, un científico de datos explora y analiza datos de fuentes múltiples y dispares. La capacitación formal es similar, con una base sólida típicamente en ciencias de la computación y las aplicaciones, modelado, estadísticas, análisis y matemáticas. Los científicos de datos

deben tener una buena visión de negocios, junto con la capacidad de comunicar los resultados a los líderes de TI y negocios, de manera que pueda ayudar en cómo una organización se acerca a un reto empresarial.

Como se puede comprender hasta el momento, para que una organización como las PYMES tenga la capacidad de realizar o aplicar ciencia de datos requiere de bastante conocimiento, es por ello que la metodología que vamos a desarrollar brindará a las PYMES una claridad de lo que se necesita o realizar ciencia de datos.

En el artículo de Ivana Dubravac y Vanja Bevanda “Mobile Business Intelligence Adoption (Case of Croatian SMEs)” las PYMES no ven una necesidad de ciencia de datos debido a los obstáculos de falta de conocimiento, bajo presupuesto y seguridad de los datos. Aunque las expectativas en cuanto a la ciencia de datos son muy altas, aunque los obstáculos han limitado el aprovechamiento de ellas, en este caso las PYMES de Croacia, no son un escenario tan alejado al de nuestro país, aunque existen unas pocas que si han tomado el riesgo y han sido un éxito.

3 Capítulo III

3.1 Marco Conceptual

Desde los inicios de la historia en la humanidad se puede mostrar la evolución del hombre bajo la adquisición constante de nuevos conocimientos. Este exponencial desarrollo brinda una mejor manera de vivir, buscando siempre el objetivo de satisfacer todo tipo de necesidad. Gracias a esto surge la relación de causa – efecto que es la base integral de toda experiencia humana.

Esto proviene de los datos que nacen a partir de las acciones que se hayan tomado a través de la historia. Estos datos se convierten en información, que es al final la fuente para la toma de decisiones, en cualquier tipo de escenario como el campo de: la salud, los negocios, el clima, y otros. Esto es un proceso normal que se hace diariamente en la vida del ser humano en estos días.

En términos de computación, actualmente se tienen avances significativos en el almacenamiento los datos y la información de una empresa o persona u organización. De la misma forma se construyen bodegas de datos, donde se guardan todas las transacciones que se hayan realizado a través del tiempo, formando así una base de datos histórica, con la cual se puede realizar análisis de la información.

Con el conocimiento a partir de la información presente en las bodegas y bases de datos las organizaciones son conscientes de todos los movimientos que se tienen, gustos de los clientes, frecuencia de compras, temporadas de transacciones, qué tipo de productos se vende con otros productos. Claramente se puede notar que se está hablando como ejemplo de un supermercado, algún negocio de venta de productos tangibles, pero este tipo de conocimiento abarca todas las áreas.

Ahora bien, todo este conocimiento va a dar como resultado el objetivo mismo de la información, la toma de decisiones. Se realizan estudios y transformaciones de datos para que se pueda tener una mejor capacidad para tomar decisiones a partir de los resultados. El descubrimiento de la información muchas veces oculta en los mismos

datos información que en distintas ocasiones no se tiene en cuenta y que en determinados casos puede ser valiosa y crítica para un mejor conocimiento del negocio y aportar mayor base a la toma de decisiones en cualquier tipo de escenario.

Seguidamente para tener una mayor comprensión de este trabajo, se definirán los principales conceptos aplicados en la ciencia de los datos:

- **Un administrador de base de datos** (DBA) es la persona o equipo de personas profesionales responsables de control y manejo del sistema de base de datos. Generalmente tienen experiencia en DBMS, diseño de bases de datos, sistemas operativos, comunicación de datos, hardware y programación. (Brock, 2015)
- **Los datos** son hechos objetivos aislados sin significado ni explicación. Es la materia prima para la creación de información, en general las organizaciones consideran que sí existe una estrategia para manejo de sus datos, aunque solamente un porcentaje tiene la estrategia bien definida. (Wagner, 2005)
- **Información**. Es el resultado de la organización y tratamiento que se aplica a los datos para producir un significado adicional al que brindan de manera aislada. (Miciruel Ftichll, 1997).
- **Diagrama de flujo de información:** Un diagrama de flujo es un diagrama que describe un proceso, sistema o algoritmo informático. Se usan ampliamente en numerosos campos para documentar, estudiar, planificar, mejorar y comunicar procesos que suelen ser complejos en diagramas claros y fáciles de comprender.
- **Conocimiento.** Este representa un mayor grado de abstracción y síntesis del significado de la información al asociar el contexto en el que se inscribe. (Marsh, 2000)
- Las Pymes son empresas de pequeño tamaño en cuanto a los ingresos que generan y los empleados con los que cuentan. Este es un concepto que se tiene considerado en el mundo, aunque se difiere entre países entre los rangos que permiten clasificar a una empresa como Pyme. En Costa Rica el 99% de las empresas están identificadas como Pyme, representan una enorme fuente de empleo. Las PYMES generan el 51% de los empleos en el sector privado, lo que

significa más de 500 mil empleos directos para las familias costarricenses según Martha Castillo, vicepresidenta de la CICR (Cámara de Industrias de Costa Rica).

- **La minería de datos** reúne unas cuantas áreas como la Estadística, la Inteligencia Artificial, la Computación Gráfica, las Bases de Datos y el procesamiento masivo y usando como materia prima las bases de datos. La idea de Minería de Datos no es nueva. Ya desde los años sesenta los estadísticos manejaban términos como data fishing, data mining con la idea de encontrar relaciones existentes en los datos en bases de datos con ruido.

  A principios de los años ochenta se empezaron a fortalecer los términos de la Minería de Datos. A principios de los años ochenta sólo existían un par de empresas dedicadas a este estudio. Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios. La tecnología informática se ha convertido en fundamental para las grandes organizaciones. Actualmente permite registrar con lujo de detalle, los elementos de todas las actividades con facilidad.

  Las bases de datos permiten almacenar cada transacción, y otros elementos que reflejan la interacción de la organización con todos sus integrantes, ya sean otras organizaciones, sus clientes, sus divisiones o sus empleados. Estos sistemas de información permiten obtener resúmenes, reportes y distintas formas de ver el estado de cuentas de la empresa para así tener un mayor conocimiento y cubrimiento del estado del negocio. Seguramente, muestren también reportes que sugieren estrategias a futuro o condiciones del nicho de mercado del negocio, que en definitiva proveen argumentos para generar la planeación y estrategias de la empresa. La Minería de Datos surge del análisis de grandes volúmenes de información, con el fin de obtener conocimiento que apoye la toma de decisiones y que contribuya a la construcción de la experiencia a partir de millones de transacciones que registra una corporación en sus sistemas de información.

  La Minería de Datos es más efectiva cuando los datos tienen características que permitan una interpretación de acuerdo con la experiencia

humana, espacio y tiempo. Por ejemplo, qué productos se venden mejor en la temporada de vacaciones, en qué regiones es productivo sembrar maíz. La tecnología ofrece analizar estos grandes volúmenes de datos y reconocer patrones en tiempo y espacio, que resultará en un modelo claro para soportar la toma de decisiones. (Marsh, 2000)

Los objetivos de la Minería de Datos son:

- Descripción de clases: provee una clasificación concisa y resumida de un conjunto de datos y los distingue unos de otros. La clasificación de los datos se conoce como caracterización, y la distinción entre ellos como comparación o discriminación.
- Asociación: es el descubrimiento de relaciones de asociación o correlación en un conjunto de datos. Las asociaciones se expresan como condiciones atributo valor y deben estar presentes varias veces en los datos.
- Clasificación: analiza un conjunto de datos de entrenamiento cuya clasificación de clase se conoce y construye un modelo de objetos para cada clase. Dicho modelo puede representarse con árboles de decisión o con reglas de clasificación, que muestran las características de los datos. El modelo puede ser utilizado para la mayor comprensión de los datos existentes y para la clasificación de los datos futuros.
- Predicción: esta función de la minería predice los valores posibles de datos faltantes o la distribución de valores de ciertos atributos en un conjunto de objetos.
- Agrupación: identifica grupos en el conjunto de datos, donde un agrupamiento es una colección de datos “similares”. La similitud puede medirse mediante funciones de distancia, especificadas por los usuarios o por expertos. La Minería de Datos trata de encontrar patrones con características parecidas, que sean escalables a grandes bases de datos y a bodegas de datos multidimensionales.

- Análisis de series a través del tiempo: analiza un gran conjunto de datos obtenidos con el correr del tiempo para encontrar en él regularidades y características interesantes, incluyendo la búsqueda de patrones secuenciales, periódicos, modas y desviaciones.

- **La Inteligencia de Negocios** es el conjunto de productos y servicios que permiten a los usuarios finales acceder y analizar de manera rápida y sencilla, la información para la toma de decisiones de negocio a nivel operativo, táctico y estratégico. Faculta a la organización a tomar mejores decisiones más rápidas. Este concepto se requiere analizar desde tres perspectivas: Hacer mejores decisiones más rápido, convertir datos en información, y usar una aplicación relacional para la administración. Admite que la alta dirección de las empresas pueda analizar y monitorear tendencias, patrones, metas y objetivos estratégicos de la organización.

Para que una empresa logre llegar a la fase del conocimiento pleno para la toma de decisiones, es necesario el uso de herramientas y procesos detallados a continuación:

- El Data Warehouse es una tecnología para el manejo de la información construido sobre la base de optimizar el uso y análisis de la misma utilizado por las organizaciones para adaptarse a los vertiginosos cambios en los mercados. Su función esencial es ser la base de un sistema de información gerencial, es decir, debe cumplir el rol de integrador de información proveniente de fuentes funcionalmente distintas (bases Corporativas, Bases propias, de sistemas externos, etc.) y brindar una visión integrada de dicha información, especialmente enfocada hacia la toma de decisiones por parte del personal jerárquico de la organización.

  Es un sitio donde se almacena de manera integrada toda la información resultante de la operatoria diaria de la organización. Además, se almacenan datos estratégicos y tácticos con el objetivo de obtener información estratégica y táctica que pueden ser de gran ayuda para aplicar sobre las mismas técnicas de análisis de datos encaminadas a obtener información oculta.

El objetivo de los DWS es consolidar información proveniente de diferentes bases de datos operacionales y hacerla disponible para la realización de análisis de datos de tipo gerencial. Los datos del DW son el resultado de transformaciones, chequeos de control de calidad e integración de los datos operacionales. Se incluyen también totalizaciones y datos pre-calculados con base en datos operaciones.

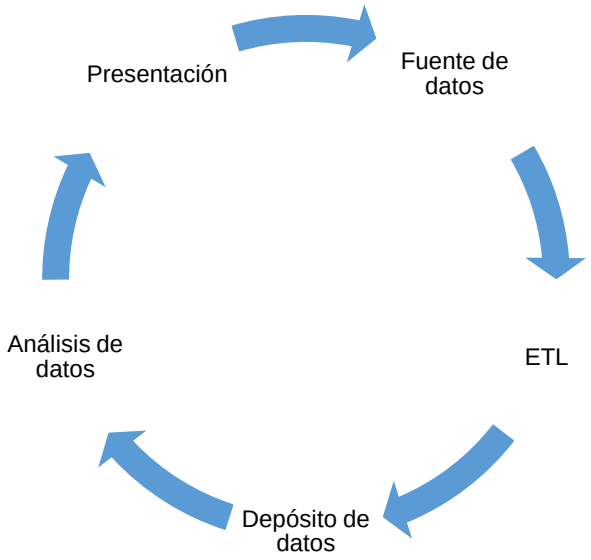
- Procesamiento transaccional en línea llamado también Análisis Multidimensional actúa sobre un depósito de datos organizando en “hipercubos” o cubos multidimensionales, sobre los llamados elementos de análisis (por ejemplo, el número de defectuosos, máximo de ventas netas, promedio de inasistencias), y también bajo ciertas dimensiones (por ejemplo, producto, centro de costo, máquina, año).

  La OLAP trata cuestionamientos tales como: ¿a cuánto asciende las ventas netas por producto, por plaza y por mes (dimensiones)?, ¿cuál es la cantidad de errores detectados por máquina y por año (dimensiones)? La presentación de datos analizados con OLAP se denomina visualización y utiliza herramientas y formas de representación gráfica como: tridimensionalidad, histogramas, regresiones, etc. La visualización permite efectuar análisis de tendencias, puntos de equilibrio y sensibilidad.
- Mercados de datos si bien existen diversas estructuras de datos, a través de las cuales se pueden representar los datos de la bodega de datos, solamente se entrará en detalle acerca de los cubos multidimensionales, por considerarse que esta estructura de datos es una de las más utilizadas y cuyo funcionamiento es el más complejo de entender. Un cubo multidimensional o cubo, representa o convierte los datos planos que se encuentran en filas y columnas, en una matriz de N dimensiones. Los objetos más importantes que se pueden incluir en un cubo multidimensional son los siguientes indicadores: Sumas que se efectúan sobre algún hecho o expresiones basadas en sumas pertenecientes a una tabla de hechos.
- Los procesos ETL es un término estándar que se utiliza para referirse al movimiento y transformación de datos. Se trata del paso que permite a las

organizaciones mover datos desde múltiples fuentes, reformatearlos y cargarlos en otra base de datos (denominada data mart o data warehouse) con el objeto de analizarlos. También pueden ser enviados a otro sistema operacional para apoyar un proceso de negocio.

Seguidamente en la figura 2 se representa de forma gráfica el ciclo de los elementos que son necesarios para la inteligencia de negocios.

Figura 2 Ciclo de inteligencia de negocios



Fuente: propia.

3.2Marco Teórico

Los primeros antecedentes de la inteligencia de negocios aparecen a inicios del siglo V antes de Cristo con un enunciado que lleva el nombre de “El arte de la guerra” escrito por Sun Tzu en el cual se demuestra la importancia de la inteligencia y la información para tratar de solucionar los problemas.

A inicios de los 60’s la Inteligencia de Negocios inicia con una dirección fija hacia las tecnologías de información, en las cuales ya se provee a usuarios una metodología de fácil acceso a modelos de decisión en la toma de decisiones (Benki & Papastathopoulos, 2009). Además, en esta misma década (60) inician los conceptos de base de datos e inteligencia de negocios sobre esta tecnología.

En los 80’s Bill Inmon y Ralph Kimball diseñan un concepto de depósito de datos, el cual proporciona una visión concreta y compleja en la cual se considera como la base de los desarrollos de la inteligencia de negocio. Se inicia con el desarrollo de los primeros softwares de reportería, con el problema que las bases de datos y los reportes todavía eran complejos para el usuario. A finales de los 80’s se formaliza el concepto de inteligencia de negocios por Howard Dresner quien considera el BI como un conjunto de metodologías, cuyo objetivo general es aumentar la eficiencia de las organizaciones. También nacen los sistemas de información para la toma de decisiones o sistemas estratégicos como evolución a los sistemas de reporterías básicas.

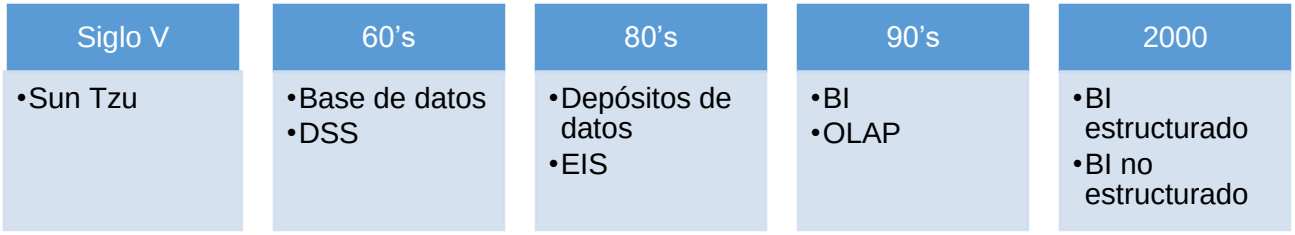
En los 90’s se contaba con un poco más de claridad sobre los que era la inteligencia de negocios, lo que promovió el surgimiento de gran cantidad de herramientas sobre el BI, lo que permitía que el acceso a la información fuera más sencilla y comprensible por casi todos los usuarios. Se utilizó en gran medida los cubos OLAP con ciertas limitantes que forzaban a las organizaciones a desarrollar con mucho esfuerzo software interno para satisfacer las necesidades del negocio.

Al inicio del año 2000 se empieza a consolidar varias herramientas de BI y se amplía la cantidad de fuentes de información en proceso de inteligencia de negocios, con el uso de fuentes no solo de estructuradas, sino de no estructuradas. Con el

continuo auge de las redes sociales inicia la necesidad de los sistemas de inteligencia de negocios más conectados a este tipo de estructuras para poder presentar los resultados de BI, en sistemas web y con los avances en la tecnología permitir a los usuarios acceder a su información a través de diversos dispositivos móviles.

Se extendió la necesidad por desarrollar una cantidad de sistemas de software especializados en inteligencia de negocios, principalmente en ciertas herramientas de BI, como herramientas en las fases de la inteligencia de negocios en ámbitos como ETL, definición, administración y visualización.

Figura 3 Historia de la inteligencia de negocios



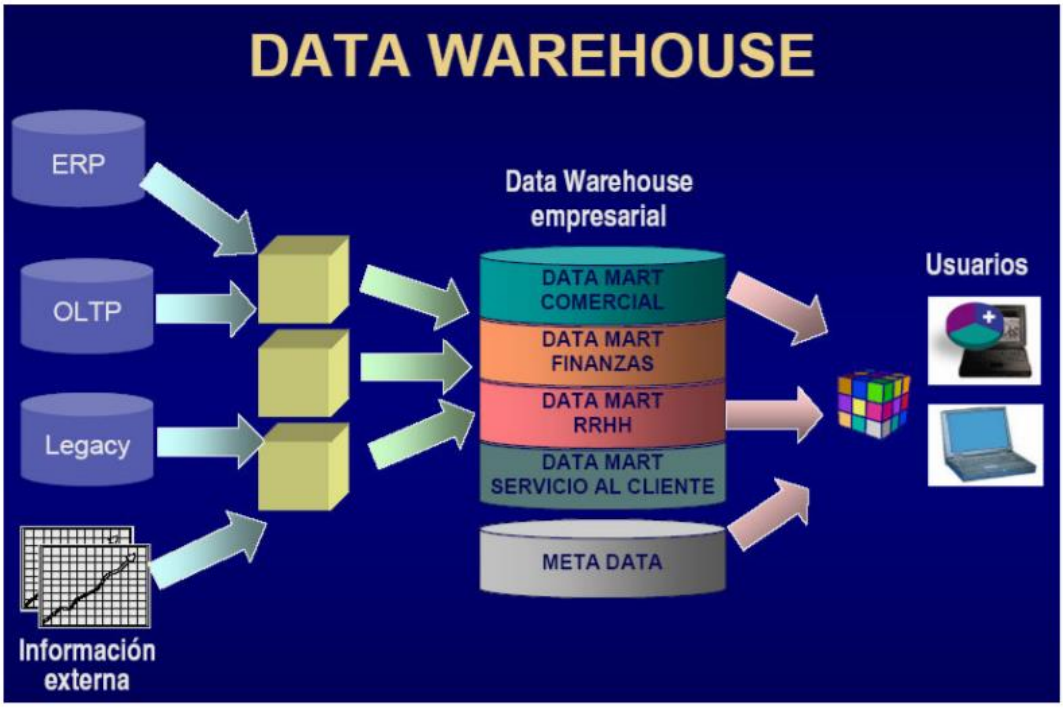
Fuente: Propia

3.2.1 Componentes de ciencia de los datos

En esta sección se hace mención de los componentes básicos que conforman una solución de inteligencia de negocios con la finalidad de tener una mejor comprensión de este trabajo de tesis.

En la figura 4 se muestra los componentes principales de la ciencia de datos.

Figura 4 Componentes de ciencia de datos



Fuente: Datamart paso a paso (Huamantumba, 2007).

A continuación, se procede a definir los componentes de la ciencia de datos, los cuales son:

1. Fuentes de información, de las cuales se partirá para alimentar de información el data warehouse.
2. Proceso ETL de extracción, transformación y carga de los datos en el data warehouse. Antes de almacenar los datos en un data warehouse, éstos deben ser transformados, limpiados, filtrados y redefinidos. Normalmente, la información que se tenga en los sistemas transaccionales no está preparada para la toma de decisiones.
3. El propio data warehouse o almacén de datos, con el metadato o diccionario de datos. Se busca almacenar los datos de una forma que maximice su flexibilidad, facilidad de acceso y administración.
4. El motor OLAP, que debe proveer la capacidad de cálculo, consultas, funciones de planeamiento, pronóstico y análisis de escenarios en grandes volúmenes de datos.

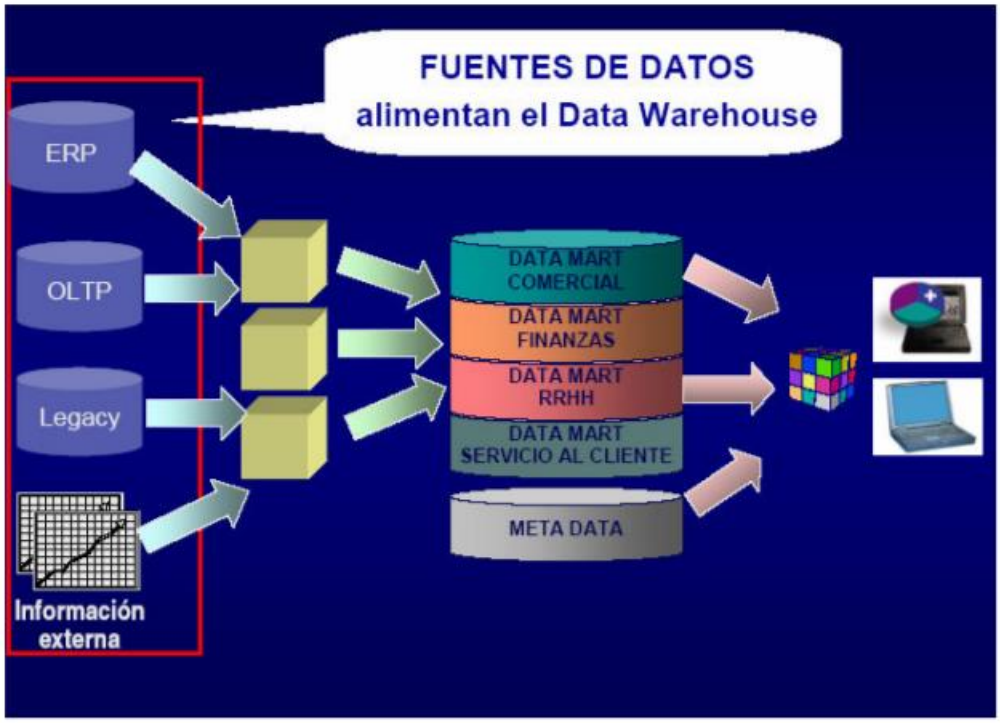
En la actualidad existen otras alternativas tecnológicas al OLAP como son las bases de datos tabulares o las bases de datos en memoria como lo es HANA.

5. Las herramientas de visualización, que permiten el análisis y la navegación de los datos.

3.2.1.1 Las fuentes de datos

En la figura 5 se representa de forma gráfica lo que es una fuente de datos.

Figura 5 Componente fuente de datos



Fuente: Datamart paso a paso (Huamantumba, 2007)

Las fuentes de datos son aquellas que poseen los datos en su mayor nivel de detalle, generalmente se generan directamente de la acción diaria del negocio, pero puede tener diferentes presentaciones. Estas fuentes son las que almacenan la información que el cliente considera relevante según lo que ha definido pudiendo apoyarse o no de Tecnologías de Información. Por lo tanto, las fuentes de datos se pueden clasificar en 2 grupos:

Fuentes de Datos Estructuradas. Son aquellas fuentes de información que tienen cierto orden y son fácilmente manipuladas para todo el proceso de BI.

Fuentes de Datos Estructuradas. Son principalmente textos.

En la tabla 2 se muestran las diferentes fuentes de datos y su respectiva descripción, para que así se tenga una mejor comprensión de estas.

Tabla 2 Fuentes de datos

Fuente	Descripción
Archivos de texto plano	Son archivos de tipo texto los cuales pueden ser estructurados o no estructurados.
Hojas de cálculo	Son hojas que pueden simular la estructura de una tabla
XML	Es un lenguaje de marcado de texto compuesto por etiquetas que permite estructurar los datos.
Bases de datos transaccionales OLTP	Son almacenes de datos que están diseñados para trabajar de manera eficiente con transacciones de escritura y lectura. Generalmente son sustentados por aplicaciones que funcionan a nivel de operación del negocio.
Bases de datos de aplicaciones especializadas	Las bases de datos alimentadas por aplicaciones especializadas como pueden ser CRM, ERP, SCM, BPM, etc. Pueden servir como fuente de información.
MDX	La información que pueda traer un cubo de información pudiera servir también de fuente de información
Repositorios de documento	Son almacenes de datos compuestos de muchos documentos, en diferentes formatos.

Fuente: Propia

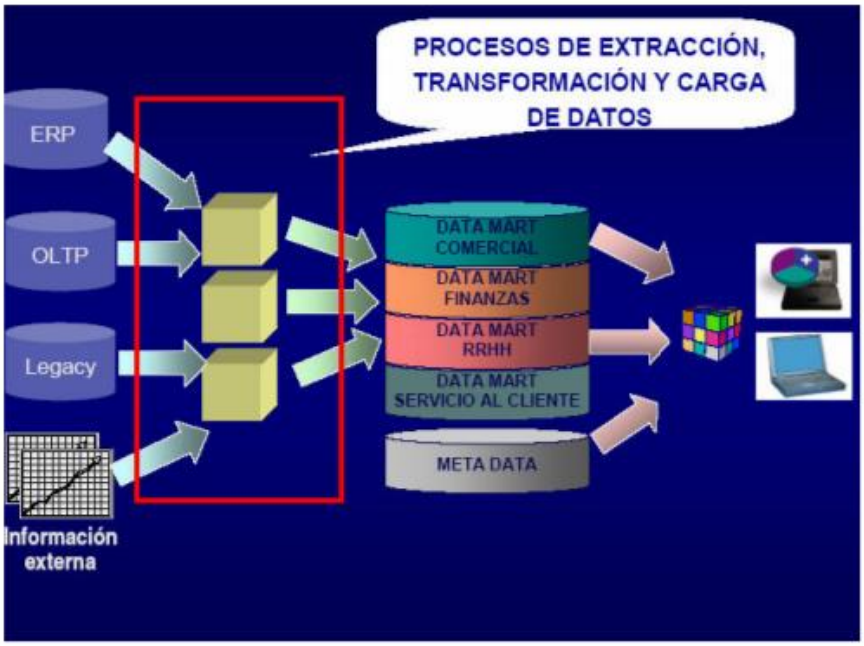
En la tabla 2 se describe las diferentes fuentes de datos que dependiendo de la complejidad que conlleven los procesos de negocio puede que este requiera de una o varias fuentes de datos, además lo ideal para cualquier proceso es que las fuentes de datos se encuentren bien documentadas para poder trabajar sobre ellas, aunque en la realidad, es muy poco probable encontrar fuentes de datos documentadas.

Las fuentes de información son de mucha importancia porque no solo proporcionan los datos necesarios para cubrir las necesidades de información de los usuarios finales, sino que también ayudan a comenzar a modelar el destino de dicha información que principalmente sería una bodega de datos.

3.2.1.2 ETL

El ETL es uno de los principales componentes, este conforma el segundo paso en la creación de un data warehouse como se muestra en la figura 6.

Figura 6 Componente ETL



Fuente: Datamart paso a paso (Huamantumba, 2007)

El proceso ETL en pocas palabras consiste en la extracción, transformación y carga de la información desde las fuentes de datos hasta la bodega de datos. Esto con la finalidad que se garantice que los datos almacenados en el destino cumplan con ciertas validaciones y formatos que permitan asegurar la integridad, consistencia y no redundancia de la información. La representación gráfica permite comprender el valor que aporta una herramienta ETL. Como se observa encerrado en el cuadrado en la

figura 6, es el componente que recoge todos los datos de las diferentes fuentes de datos (un ERP, CRM, hojas de cálculo sueltas, una base de datos SQL, un archivo JSON de una BBDD NoSQL orientada a documentos, etc.) y ejecuta las siguientes acciones (principales, y entre otras):

- ✓ Validar los datos
- ✓ Limpiar los datos
- ✓ Transformar los datos
- ✓ Agregar los datos
- ✓ Cargar los datos

El proceso ETL se ajusta a los requerimientos definidos por el usuario final, quien es el que expresa qué preguntas quiere responder con la información y quién puede indicar de dónde provienen los datos y en qué condiciones debe estar para poder ser convertidos en información.

3.2.1.2.1 Los procesos ETL básicos

El proceso ETL está compuesto por tres etapas, las cuales son:

- **Extracción**: esta es la primera etapa y corresponde a la obtención de los datos que luego serán manipulados para ser cargados en el DW.
- **Transformación**: esta es la segunda etapa, acá la información que es extraída hacia el área de datos temporales realiza distintos pasos de transformación, como la limpieza de la información o selección de los campos necesarios para la carga del DW, también se pueden combinar distintas fuentes de datos y realizar otras operaciones.
- **Carga**: esta es la etapa final del proceso, los datos están en forma para ser cargados dentro del DW. En ésta y en las anteriores etapas se pueden generar distintos tipos de registros.

3.2.1.2.2 Características de un ETL

Las características más importantes que ha de incluir una herramienta ETL según Gartner son las siguientes:

- Conectividad / capacidades de Adaptación (con soporte a orígenes y destinos de datos): habilidad para conectar con un amplio rango de tipos de estructura de datos, que incluyen bases de datos relacionales y no relacionales, variados formatos de ficheros, XML, aplicaciones ERP, CRM o SCM, formatos de mensajes estándar (EDI, SWIFT o HL7), colas de mensajes, emails, websites, repositorios de contenido o herramientas de ofimática.
- Capacidades de entrega de datos: habilidad para proporcionar datos a otras aplicaciones, procesos o bases de datos en varias formas, con capacidades para programación de procesos batch, en tiempo real o mediante lanzamiento de eventos.
- Capacidades de transformación de datos: habilidad para la transformación de los datos, desde transformaciones básicas (conversión de tipos, manipulación de cadenas o cálculos simples), transformaciones intermedias (agregaciones, sumarizaciones, lookups) hasta transformaciones complejas como análisis de texto en formato libre o texto enriquecido.
- Capacidades de Metadatos y Modelado de Datos: recuperación de los modelos de datos desde los orígenes de datos o aplicaciones, creación y mantenimiento de modelos de datos, mapeo de modelo físico a lógico, repositorio de metadatos abierto (con posibilidad de interactuar con otras herramientas), sincronización de los cambios en los metadatos en los distintos componentes de la herramienta, documentación, otros.
- Capacidades de diseño y entorno de desarrollo: representación gráfica de los objetos del repositorio, modelos de datos y flujos de datos, soporte para test y debugging, capacidades para trabajo en equipo, gestión de workflows de los procesos de desarrollo, etc.
- Capacidades de gestión de datos (calidad de datos, perfiles y minería).

- Adaptación a las diferentes plataformas hardware y sistemas operativos existentes: Mainframes (IBM Z/OS), AS/400, HP Tandem, Unix, Wintel, Linux, Servidores Virtualizados, etc.
- Las operaciones y capacidades de administración: habilidades para gestión, monitorización y control de los procesos de integración de datos, como gestión de errores, recolección de estadísticas de ejecución, controles de seguridad, etc.
- La arquitectura y la integración: grado de compactación, consistencia e interoperabilidad de los diferentes componentes que forman la herramienta de integración de datos (con un deseable mínimo número de productos, un único repositorio, un entorno de desarrollo común, interoperabilidad con otras herramientas o vía API), etc.
- Capacidades SOA.

Según las características descritas anteriormente, Gartner realiza el grafico de la figura 7, en donde clasifica las herramientas que existen en el mercado para la integración de datos.

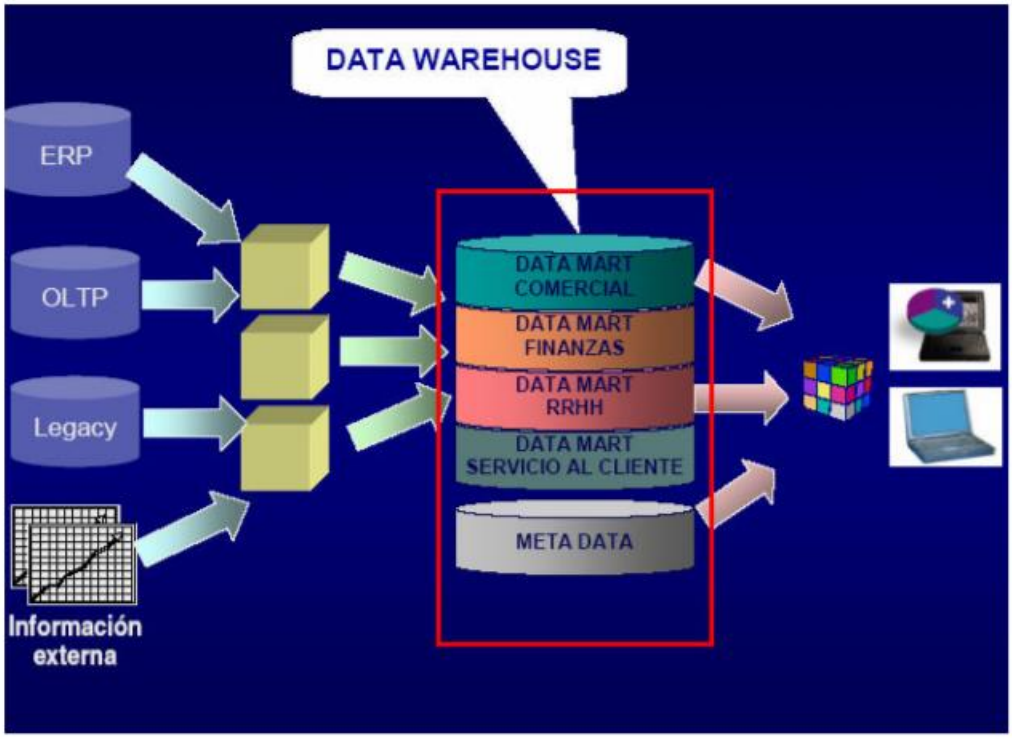
Figura 7 Cuadrante Mágico de Gartner sobre herramientas de integración de datos

Fuente: Cuadrante mágico de herramientas para integración de datos Gartner (2016)

3.2.1.3 Almacén de datos o Data Warehouse

En la figura 8 se muestra lo que es el componente data warehouse encerrado en un cuadro rojo.

Figura 8 Componente repositorio de datos



Fuente: Datamart paso a paso (Huamantumba, 2007)

Los almacenes de datos poseen los conceptos principales de:

1. Colección de datos que están orientados a temas, integrados, no volátiles y que varían en el tiempo y cuya finalidad es servir de soporte en la toma de decisiones. Estos datos contienen granularidad de los datos corporativos. (Inmon, 2002)
2. Es el conglomerado de datos organizacionales en áreas de desarrollo y presentación, donde dichos datos provienen de la operación y son manipulados para el análisis que el usuario final requiera. (Kimball & Ross, 2002)

De acuerdo con estas definiciones se sabe que el Data Warehouse debiera contener las principales áreas del negocio las cuales de preferencia deberían estar identificadas en el modelo de negocio.

Según la definición más tradicional del término DW, especificada por Bill e Inmon a principios de la década de los 90, los DW se caracterizan por ser:

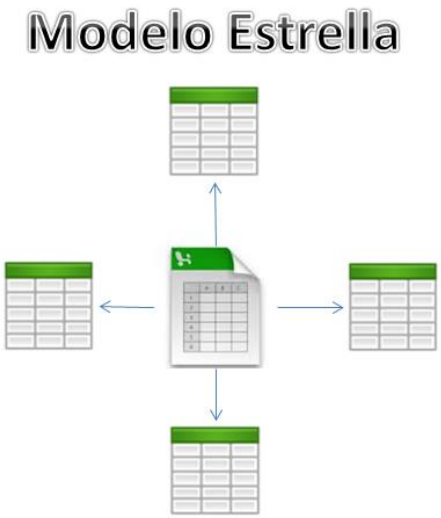
- Orientados al sujeto: Los datos almacenados brindan información sobre un sujeto o asunto en particular en lugar de concentrarse en la dinámica de las transacciones de la organización.
- Integrado: Los datos cargados en el DW pueden provenir de diferentes fuentes y son integrados para dar una visión global coherente.
- Variables en el tiempo: El DW se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones, lo que implica que todos los datos deben estar asociados con un período de tiempo específico
- No volátiles: Los datos son estables en el DW, se agregan y modifican datos, pero los datos existentes no son removidos.

Para cada una las áreas del negocio se deben identificar las medidas, las tablas de dimensión y las tablas de hechos. Las medidas son atributos utilizados como unidades de medida sobre las entidades de la base de datos que conforman los valores o indicadores por analizar. Las tablas de dimensión son aquellas tablas que contienen atributos de los datos, que permiten darles sentido a los datos numéricos, por ejemplo, Clientes, Productos. Las tablas de hechos son tablas que contienen la información numérica de los indicadores por analizar, es decir la parte cuantitativa de la información.

Estas tablas se pueden organizar en dos tipos de modelado:

1. Estrella. Las tablas de dimensiones no se encuentran normalizadas por lo que una tabla de hechos contiene diversos campos identificadores con sus respectivas llaves foráneas a cada tabla de dimensión. Este tipo de modelado permite la creación de jerarquías lo que permite navegar por la información (Boussaid, Ben Messaoud, Choquet, & Anthoard, 2006).

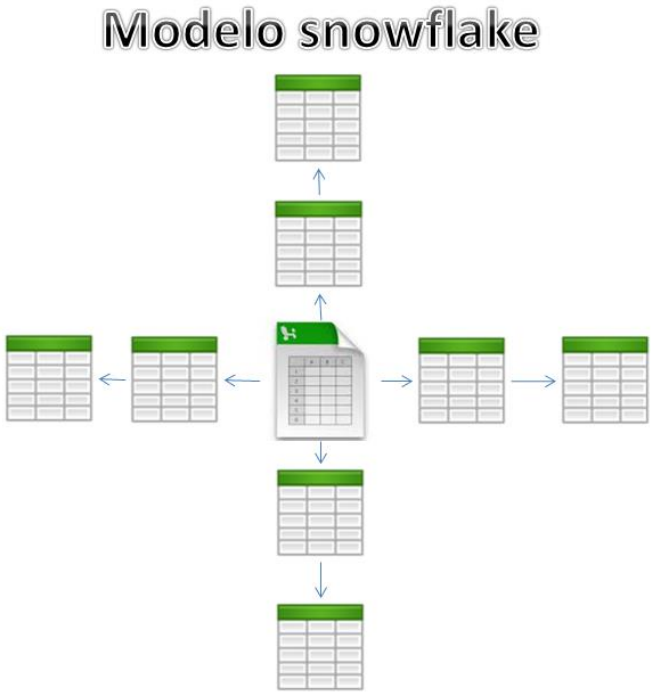
Figura 9 Modelo de depósitos de datos estrella



Fuente: Diseño propio

2. Snowflake Las tablas de dimensiones se encuentran normalizadas por lo que cada tabla contiene un campo identificador, un campo de descripción y un campo que funcione como llave foránea que la une a otra tabla de dimensión. Esta estructura permite que las tablas se unan a la tabla de hechos mediante este campo identificador. Las llaves de la tabla de hechos se encuentran ligadas a varias dimensiones de manera directa e indirecta (Boussaid, Ben Messaoud, Choquet, & Anthoard, 2006).

Figura 10 Modelo de depósito de datos copo de nieve



Fuente: Diseño propio

3.2.1.4 Indicadores clave de Rendimiento (KPI)

Para poder entender lo que son los KPI (Key Performance Indicators) primeramente se debe comprender lo que es una métrica. Una métrica es una medida numérica que representa una parte de los datos del negocio en relación con una o varias dimensiones a través de sus jerarquías. Entonces un KPI es una métrica que está ligada a los objetivos de la empresa y normalmente se presentan en forma de porcentajes y tienen la finalidad de facilitar al usuario de negocio el identificar si están funcionando los planes (González, 2006).

Una vez que han analizado su misión, han identificado los grupos de poder y han definido sus objetivos, las organizaciones necesitan un sistema para medir su progreso hacia la consecución de los objetivos. Los KPI son los instrumentos adecuados para llevar a cabo mediciones del progreso de los planes. Los KPI deben ser cuantificables y deben medir las mejoras en aquellas actividades que son críticas

para conseguir el éxito de la organización. Los KPI deben estar relacionados con los objetivos y con las actividades fundamentales de nuestra organización

Para la definición de KPI se debe buscar que estos cumplan con 7 características:

1. Métricas no financieras.
2. Requieran ser medidas constantemente.
3. Son dirigidas para los altos mandos.
4. Claramente indica que acción debe ser tomada.
5. Se identifica claramente a los responsables de cada KPI.
6. Tienen un impacto significativo.
7. Impulsa a tomar acciones apropiadas.

3.2.1.5 Cubo

Un cubo permite a los datos ser modelados y vistos en diferentes dimensiones para lo cual es obligatorio que trabaje bajo un modelo multidimensional, basado en dimensiones y hechos. Dicho de otra forma, un cubo procesa la información de acuerdo a un diseño específico que concuerda con los requisitos de información determinado por el negocio de tal manera que permite percibir las necesidades establecidas, por ejemplo, se puede ver las ventas por empleado y por producto durante el mes de septiembre, de esta manera se mezclan 3 dimensiones y un hecho.

Los cubos son estructuras que representan los datos como una matriz en la cual sus ejes corresponden a los criterios de análisis y en los cruces se encuentran los valores por analizar. Estos cubos constan de dimensiones y medidas. Las dimensiones están relacionadas con los criterios de análisis de los datos, son variables independientes, representan los ejes del cubo y están organizadas en jerarquías. Las medidas son los valores o indicadores por analizar, estas corresponden a datos asociados a relaciones entre los objetos del problema, que son variables dependientes y se encuentran en la intersección de las dimensiones.

3.2.1.6 Minería de Datos

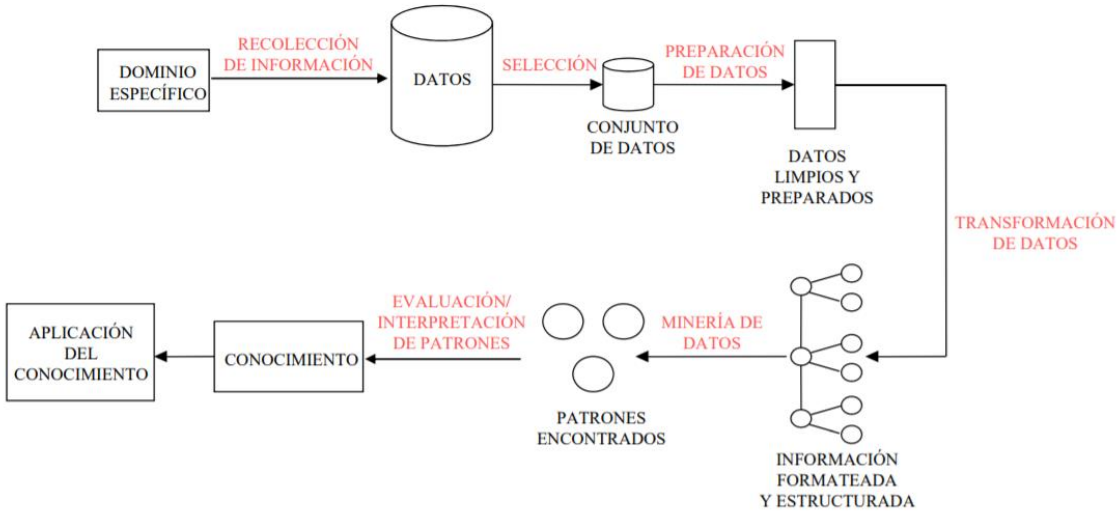
La minería de datos consiste en un proceso cuyo objetivo es la extracción de conocimiento a partir de un conjunto muy grande de datos. El proceso general es conocido como KDD (Knowledge Discovery from data) y está conformado de diversas fases entre las cuales encontramos a la minería de datos.

El proceso KDD está compuesto por fases las cuales son:

1. Limpieza de Datos.
2. Integración de Datos.
3. Selección de Datos.
4. Transformación de Datos.
5. Minería de Datos.
   5.1. Definición de objetivos.
   5.2. Recolección de datos e integración.
   5.3. Análisis de exploración.
   5.4. Selección de atributos.
   5.5. Desarrollo del modelo y validación.
   5.6. Predicción e interpretación.
6. Evaluación de Patrones.
7. Presentación del Conocimiento.

A continuación, en la figura 11 se muestra las fases del proceso para el descubrimiento del conocimiento en las bases de datos. Las fases estas indicadas de color rojo en la imagen.

Figura 11 Fases del proceso de descubrimiento de conocimiento en bases de datos



Fuente: (Febles, 2005)

> Para que este proceso se lleve a cabo con éxito se requiere de un gran compromiso por parte de los expertos en el dominio de los datos y los mineros de datos. Mientras que el experto en el dominio de los datos responde todas las posibles dudas que le puedan surgir al minero de datos, el minero de datos hace uso de métodos matemáticos que lleven a un aprendizaje inductivo a partir de los datos (Vieira Braga, 2009).

Como lo menciona Vieira, el proceso de minería requiere de gran compromiso por parte de los expertos, debido a que se deben saber realizar las siguientes 6 principales tareas:

1. Clasificar: Asignar a cada entidad un grupo.
2. Estimación: Asignar valores numéricos a variables.
3. Predicción: Clasificación de entidades de acuerdo con su comportamiento esperado en el futuro.
4. Agrupado de afinidad: Evaluación de relaciones entre elementos de datos.

5. Agrupamiento: Dividir un universo de datos en conjuntos pequeños que tengan similitudes.
6. Descripción: Caracterizar lo que se haya descubierto a lo largo del proceso de minería de datos.

La minería de datos puede tener dos enfoques:

a) Descriptivo: permite la identificación de patrones en términos de reglas matemáticas que sean fácilmente entendidas por los expertos en el dominio de los datos y que de esta manera representen conocimiento para ellos.
b) Predicción: es la utilización de variables para predecir valores desconocidos o futuros.

Para lograr que estos enfoques se lleven a cabo existen variados algoritmos matemáticos que permiten obtener los modelos deseados, para mencionar algunos son: C4.5, K-Means, Support Vector Machines y otros.

> "El proceso puede ser mediante un aprendizaje guiado en el cual existe una previa clasificación de los datos o aprendizaje no supervisado en el cual no existe ninguna clase de clasificación de los datos" (Vercellis, 2009).

Una aportación en el campo de la minería de datos se dio a partir de un estudio enfocado en las tareas de marketing (Kumar Kar, Kumar, & Kumar De, 2010), en el cual se hace un resumen de como algunas tareas de minería de datos ayudan a mejorar el resultado de los esfuerzos de mercadotecnia. El estudio sobre tareas de marketing de Kumar Kar, Kumar, & Kumar De en el año 2010 menciona lo siguiente:

- ✓ Agrupamiento. Se agrupan los datos en clases de acuerdo con sus similitudes, esta técnica puede ser utilizada para segmentar un mercado.
- ✓ Clasificación. Se realiza a partir de la teoría de decisión Bayesiana, redes neuronales, etcétera. Esto puede permitir desarrollar publicidad selectiva acorde a una segmentación previa.
- ✓ Asociación de patrones. Sirve para predecir patrones basados en secuencias de datos a partir de secuencias de entrenamiento. Usualmente esto se realiza

mediante la composición de alguna técnica de clasificación. En marketing permite la predicción de preferencias de los clientes, qué productos o publicidad pueden ser interesantes.

- ✓ Sumarización: Son métodos que permiten agrupar la información de acuerdo con ciertas variables que le dan sentido a métricas establecidas con las que se relaciona. Permite la identificación de la utilidad por segmentos lo que facilita que se distribuya los recursos acordes a este análisis.
- ✓ Modelo predictivo: Es un proceso mediante el cual un modelo es creado y elegido para hacer mejores predicciones de una probabilidad. Este tipo de procesos permiten que se identifiquen las probabilidades de ocurrencia de respuesta de los clientes de tal manera que se les pueda proporcionar promociones especiales.
- ✓ Análisis de liga: Es una metodología para mapear y medir el flujo de la información mediante la interacción de sus nodos. Permite tener visibilidad que persona es un líder y puede influenciar a los seres a su alrededor.

3.2.1.7 Algoritmos en la ciencia de los datos

La ciencia de datos es un campo interdisciplinario que involucra múltiples conocimientos para extraer un mejor entendimiento de los datos en sus diferentes formas por lo que dada esta situación es muy importante conocer muchos algoritmos. Un algoritmo es un conjunto prescrito de instrucciones bien definidas, ordenadas y finitas que permite realizar una actividad mediante pasos sucesivos que no generen dudas a quien deba realizar dicha actividad.

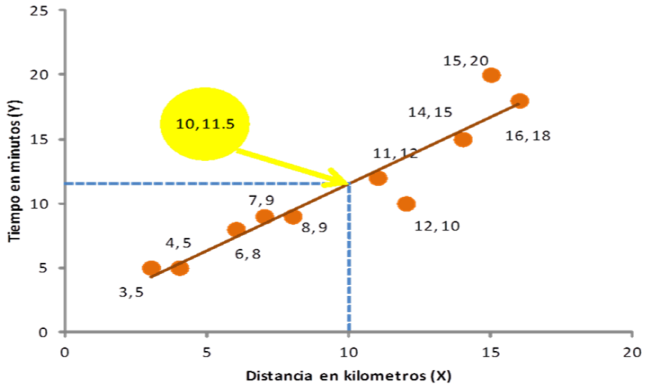
A continuación, se procederá a explicar los principales algoritmos más conocidos y utilizados en el campo de la ciencia de los datos.

3.2.1.7.1.1 Regresión lineal

Método de análisis estadístico implementado en una gran variedad de campos, especialmente las ciencias Sociales y naturales. El método permite la predicción de valores futuros con base en un análisis de datos inicial.

El modelo está basado en variables dependientes (endógenas) e independientes (exógenas), este modelo busca obtener una descripción y evaluación entre estas variables; si se cuenta con dos o más independientes, es regresión múltiple.

Figura 12 Gráfica de regresión lineal

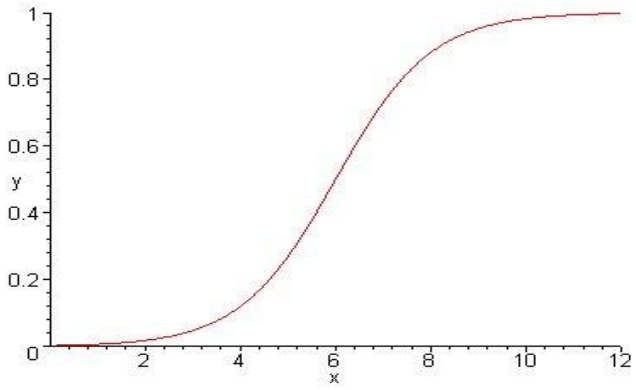


Fuente: Algoritmos en Ciencia de los Datos para la aplicación en pequeñas y medianas empresas (Tecnológico de Costa Rica, 2017)

3.2.1.7.1.2 Regresión logística

En estadística, la regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento que ocurra como función de otros factores. Resulta necesario para los casos en los que se desea predecir la presencia o ausencia de una característica o resultado según los valores de un conjunto de predictores. Es similar a un modelo de regresión lineal, pero está adaptado para modelos en los que la variable dependiente es dicotómica.

Figura 13 Gráfica de regresión logística

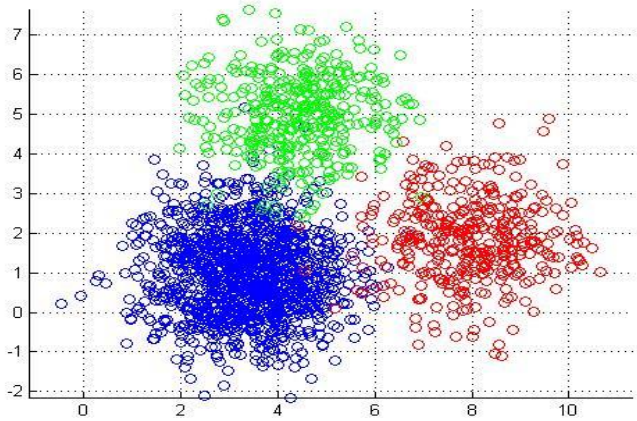


Fuente: Algoritmos en Ciencia de los Datos para la aplicación en pequeñas y medianas empresas (Tecnológico de Costa Rica, 2017)

1.1.1.1.1.1 K-MEAS

Es un tipo de aprendizaje no supervisado, que se utiliza cuando hay datos sin etiquetar (es decir, los datos sin categorías definidas o grupos). El objetivo de este algoritmo es encontrar grupos en los datos, con el número de grupos representados por la variable K. El algoritmo funciona de forma iterativa para asignar a cada punto de datos a uno de K grupos en función de las características que se proporcionan. Cuando se usan heurísticas como el algoritmo de Lloyd es fácil de implementar incluso para grandes conjuntos de datos. Por lo que ha sido ampliamente usado en muchas áreas como segmentación de mercados.

Figura 14 Gráfica de K-Means



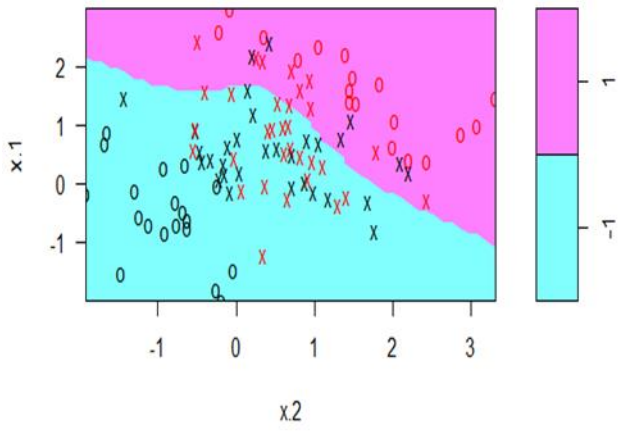
Fuente: Algoritmos en Ciencia de los Datos para la aplicación en pequeñas y medianas empresas (Tecnológico de Costa Rica, 2017)

3.2.1.7.1.3 Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial (MSV) son una moderna y efectiva técnica de IA, que ha tenido un formidable desarrollo en la última década del presente siglo, se desarrollaron inicialmente para solucionar problemas de clasificación, pero se han ampliado para problemas de regresión. Los resultados finales a los que se puede llegar luego del empleo de las MSV pueden ser cualitativos o cuantitativos, para el análisis cuantitativo se emplean MSV para regresión. La idea básica de MSV consiste en realizar un mapeo de los datos de entrenamiento $x \in X$, a un espacio de mayor dimensión F a través de un mapeo no lineal $\varphi: X \rightarrow F$, donde se puede realizar una regresión lineal.

En la figura 15 se muestra un ejemplo de un representacion grafica realizada utlizando el algoritmo de máquinas de soporte vectorial.

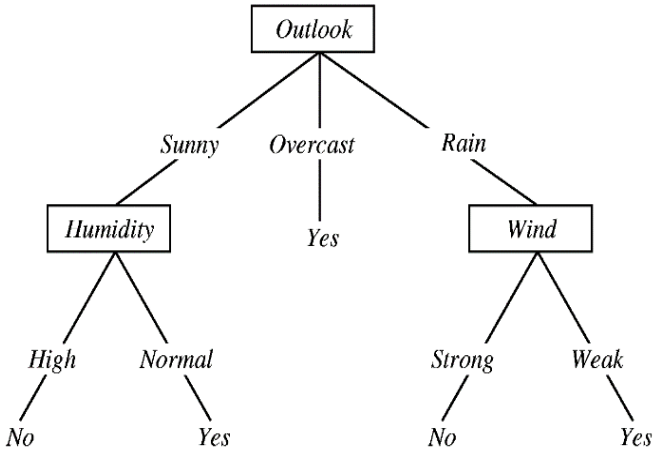
Figura 15 Gráfico de Máquina de soporte vectorial



Fuente: Algoritmos en Ciencia de los Datos para la aplicación en pequeñas y medianas empresas (Tecnológico de Costa Rica, 2017)

3.2.1.7.1.4 Árboles de decisión

Figura 16 Árbol de decisión



Fuente: Algoritmos en Ciencia de los Datos para la aplicación en pequeñas y medianas empresas (Tecnológico de Costa Rica, 2017).

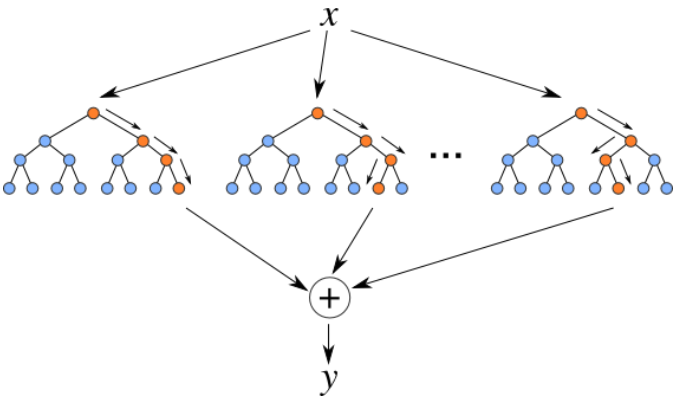
Son un algoritmo categorizado como aprendizaje con base en similitudes, los árboles de decisión son de los algoritmos más sencillos y fáciles de implementar que a su vez está entre los más poderosos. Este algoritmo genera un árbol de decisión de forma recursiva al tomar en cuenta el criterio de la mayor proporción de ganancia, es decir, elige la propiedad que mejor clasifica a los datos.

3.2.1.7.1.5 Bosques aleatorios

Predictores de árboles binarios de tal manera que cada árbol depende de los valores de un vector aleatorio muestreado independientemente y con la misma distribución para todos los árboles en el bosque. De tal manera que cada árbol depende de los valores de un vector aleatorio muestreado independientemente y con la misma distribución para todos los árboles en el bosque.

La figura 17 es un ejemplo gráfico para comprender de mejor forma lo que es un bosque aleatorio.

Figura 17 Bosque Aleatorio



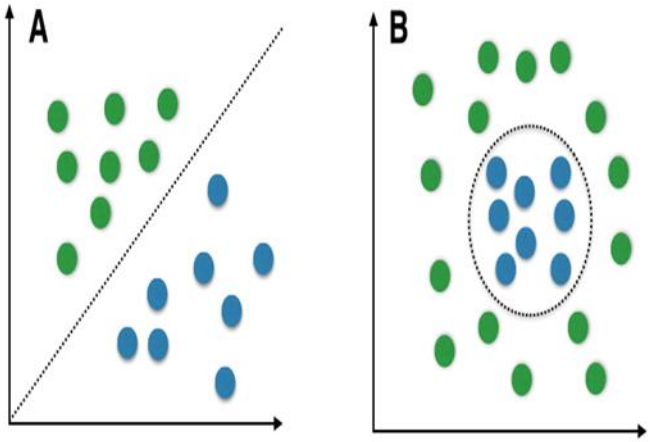
Fuente: Algoritmos en Ciencia de los Datos para la aplicación en pequeñas y medianas empresas (Tecnológico de Costa Rica, 2017).

3.2.1.7.1.6 Clasificador Bayesiano Ingenuo

En teoría de la probabilidad y minería de datos, un clasificador Bayesiano ingenuo es un clasificador probabilístico fundamentado en el teorema de Bayes y complementado con algunas hipótesis de simplificación. Donde su técnica de clasificación y predicción supervisada construyen modelos que brindan la probabilidad de posibles resultados. Su técnica debe ser supervisada porque necesita tener ejemplos.

En la figura 18 se muestra un claro ejemplo de la clasificación que se puede realizar con el clasificador bayesiano, en donde en el gráfico A se muestra cómo se separan los elementos, y en el gráfico B como trata de agrupar los elementos de color azul debido a que se encuentran en el centro y son los que tienen mayor similitud.

Figura 18 Clasificador Bayesiano Ingenuo

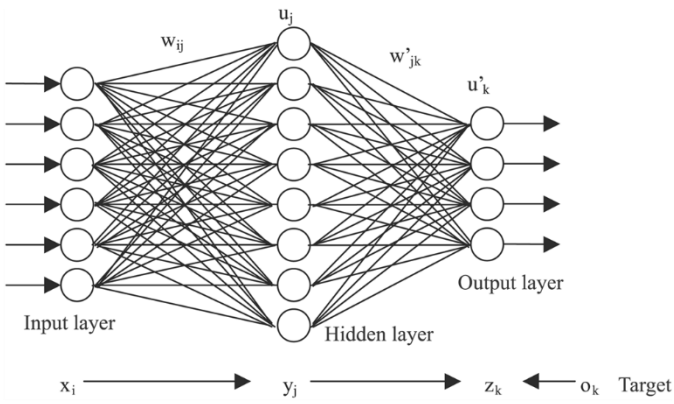


Fuente: Algoritmos en Ciencia de los Datos para la aplicación en pequeñas y medianas empresas (Tecnológico de Costa Rica, 2017).

3.2.1.7.1.7 Redes neuronales artificiales

Son un modelo computacional basado en un gran grupo interconectado de unidades neuronales simples (neuronas artificiales), de forma aproximadamente análoga al comportamiento que tiene los axones de las neuronas en los cerebros biológicos. Son utilizadas en la ciencia artificial, ciencia de los datos y otras disciplinas de investigación, que se basa en una gran colección de unidades simples conectadas. Cada nodo representa una neurona artificial y cada flecha representa una conexión desde la salida de una neurona a la entrada de otra.

Figura 19 Redes neuronales artificiales

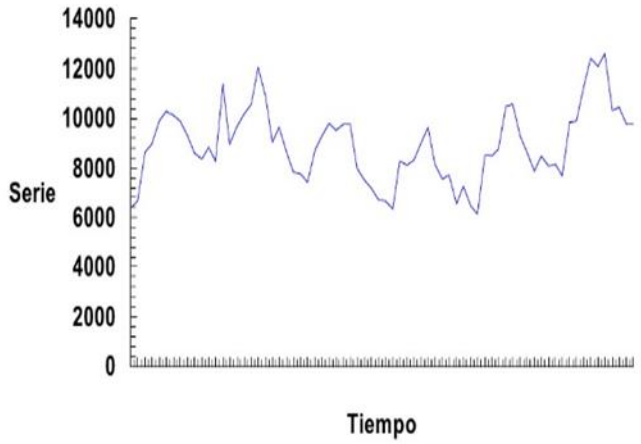


Fuente: Algoritmos en Ciencia de los Datos para la aplicación en pequeñas y medianas empresas (Tecnológico de Costa Rica, 2017).

3.2.1.7.1.8 Series de tiempo

Básicamente una serie de tiempo se le denomina a cualquier variable que conste de datos reunidos, registrados u observados sobre incrementos sucesivos de tiempo. Por lo tanto, se concluye que es una secuencia ordenada de observaciones sobre una variable en particular. Una serie estacionaria es aquella cuyos momentos al origen y a la media, no varían a través del tiempo. Estas situaciones se presentan cuando los patrones de demanda que influyen sobre la serie son relativamente estables.

Figura 20 Series de tiempo



Fuente: Algoritmos en Ciencia de los Datos para la aplicación en pequeñas y medianas empresas (Tecnológico de Costa Rica, 2017).

3.2.1.8 Presentación.

El propósito de tener un Data Warehouse es que genere información al usuario final para que pueda tomar las mejores decisiones pertinentes para el negocio, por lo tanto, existen varias formas de presentar la información de manera que sea posible su comprensión.

Entrá las maneras en las que se puede presentar están:

3.2.1.8.1 Cuadros de mandos.

Son utilizados para rectificar las operaciones con la estrategia del negocio. Este tipo de presentación de información se realiza principalmente por medio de los KPI y está dirigida a los altos mandos. Las herramientas de este tipo como lo especifica un artículo sobre Pymes japonesas (Aoki & Hasebe, 2012), son importantes para monitorear la labor de una estrategia y dichas estrategias se deben evaluar constantemente para que se adapten al ambiente. En este mismo estudio se indica un proceso para la implementación de un BSC:

- ✓ Generar caso de la compañía.
- ✓ Identificar problema.
- ✓ Generar y planear la solución.
- ✓ Revisión y aprobación del cliente para su implementación.
- ✓ Validación (este paso puede hacer que se repita nuevamente todo el proceso).

En la figura 21 se muestra un ejemplo de lo que es un cuadro de mando de ventas, el cual está compuesto por los principales KPI que puede tener una organización en el área de ventas

Figura 21 Ejemplo de cuadros de mandos de ventas

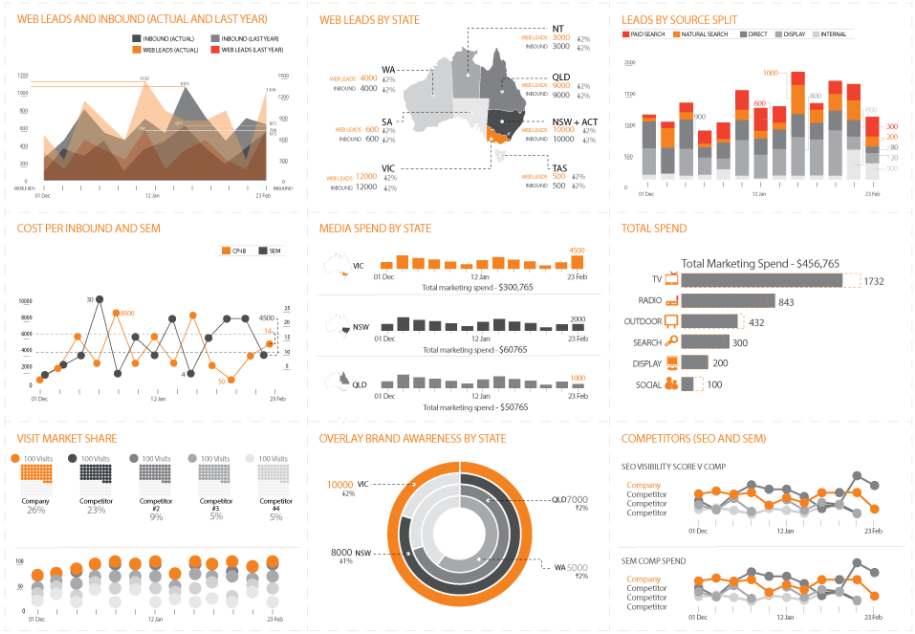
Fuente: (Konta Asesores, 2014)

3.2.1.8.2 Dashboards.

Está destinado a medir el desempeño de los procesos de negocio para asegurar que cumplan los objetivos establecidos, el dashboard contiene métricas y KPI' para medir el cumplimiento de los objetivos proveyendo información para la toma de decisiones sobre acciones que ayuden el negocio al cumplimiento de sus objetivos.

La figura 22 es un ejemplo de un dashboard con excelentes técnicas de visualización.

Figura 22 Ejemplo de un dashboard



Fuente: (Datalabs, 2017)

3.2.1.8.3 Reportes.

Permite la visualización de la información en forma de tablas que contienen información para ser analizada por los especialistas del negocio, este tipo de presentación de la información permite presentar la información con diferentes niveles de detalle. Así mismo esta información dependiendo de la herramienta que se utilice para presentarla, puede llegar al usuario de diferentes maneras, como se muestra en la tabla 3.

Tabla 3 Diferentes formas de presentación

Presentación	Descripción
Archivos	Los reportes se pueden generar en archivos de algún tipo, generalmente y por comodidad se utiliza Excel.
Correo	Los reportes pueden ser enviados por correo electrónico.
Presentación Web	Pueden conectarse a un portal empresarial o al front end web de la herramienta implementada
Móviles	Existen herramientas que permiten que los reportes sean consultados desde los móviles con conexión a internet y que contengan las aplicaciones móviles necesarias

3.3 Marco Metodológico

Las metodologías permiten llevar a cabo el proceso de ciencia de los datos en forma sistemática y no trivial. Estas metodologías ayudan a entender el proceso de descubrimiento de conocimiento para proveer una guía de planificación y ejecución en los proyectos. Algunos modelos conocidos como metodologías son en realidad un modelo de proceso: un conjunto de actividades y tareas organizadas para llevar a cabo un trabajo. La diferencia fundamental entre metodología y modelo de proceso radica en que el modelo de proceso establece qué hacer, y la metodología especifica cómo hacerlo.

Metodología KDD se define como la extracción de conocimiento está principalmente relacionada con el proceso de descubrimiento conocido como “Knowledge Discovery in Databases” (KDD), que descubre conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información. No es un proceso automático, es un proceso repetitivo que explora volúmenes muy grandes de datos para determinar relaciones. Es un proceso que extrae información de calidad que puede usarse para dibujar conclusiones basadas en relaciones o modelos dentro de los datos.

Metodología CRISP-DM se encuentra dentro de las metodologías más utilizadas para la elaboración de proyectos de minería de datos, está basado en actividades ordenadas en seis fases que recorren todo el proceso de minería de datos, desde la definición de los objetivos del negocio que se pretende obtener, hasta la vigilancia y el mantenimiento del modelo que se proponga e implemente. Cada una de esas fases se ha subdividido a su vez en tareas ordenadas en un esquema jerárquico, desde un mayor a un menor nivel de detalle.

Metodología SCRUM se busca identificar cada uno de los actores, fases, roles en los que se permita la aplicación a cada uno de los componentes de la ciencia de los previamente identificados, y por medio de los resultados obtenidos en cada fase, se logre la integración de los mismos para obtener una correcta visión, diseño,

implementación, operación, seguimiento, mantenimiento y mejora continua del departamento de la ciencia de los datos.

Para el desarrollo de la metodología propuesta se realizó una combinación de principios y fases de cada una de las metodologías mencionadas anteriormente, tomando como base principal la metodología KDD para la búsqueda de nuevos conocimientos, en la parte donde ya se procede con lo que es la minería de datos se utiliza lo que es la metodología CRISP- DM, además para el manejo del proyecto se usó SCRUM, buscando darle agilidad y que el negocio vaya obteniendo resultados durante el proceso.

3.3.1 Tipo de Investigación

Esta investigación es de tipo aplicada, ya que consiste en la solución de un problema el cual es contar con una metodología que guíe para aplicar ciencia de los datos en la PYMES GO-LABS ENTERPRISES.

3.3.2 Alcance Investigativo

Esta investigación debe realizarse sobre un alcance de tipo de exploratorio debido a que se tiene varias dudas sobre el tema, y todavía no existe con anterioridad una metodología que incluya todo lo necesario para aplicación de ciencia de los datos, en organizaciones similares de TI. Al finalizar esta exploración el resultado final deberá darnos una metodología clara para realizar el proceso de aplicación de ciencia de datos. Cabe mencionar que este proyecto no incluye mediciones en cuanto al contexto de nivel de madurez de la aplicación de ciencia de datos.

3.3.3 Enfoque

Esta investigación se realizará por medio de un enfoque alternativo en donde no se realizará una escogencia entre el enfoque cualitativo y cuantitativo, solo se centrará en dejar clara la posición ante lo que se investigue y así evitar que sea subjetivo basado en rúbricas y dando la posibilidad de ser estudiado. Para ello nos basamos en las ideas de los autores como (Chavarría, 2011) y (Padrón, 1992), que insisten en la

necesidad de reconocer que los enfoques cualitativos y cuantitativos no pueden existir de manera separada.

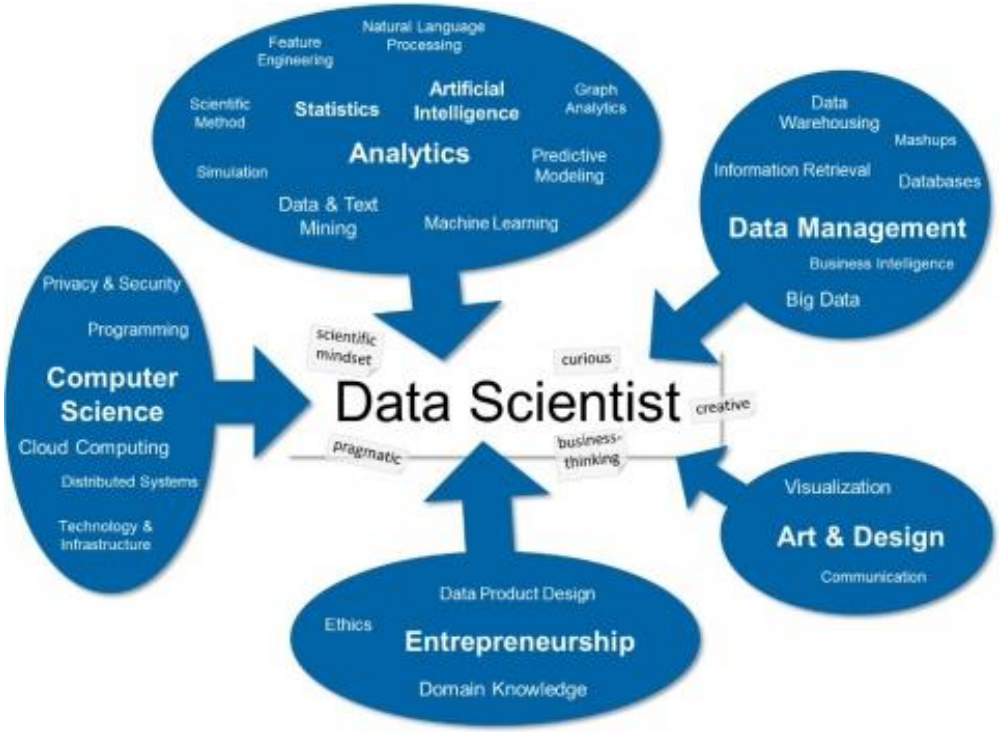
De esta manera se comprenderá que es necesario para la aplicación de las ciencias de los datos en organizaciones, para el caso de esta investigación será en la PYMES GO-LABS ENTERPRISES. Para ello se explorará y describirá las diferentes perspectivas teóricas que se aplican en el proceso de aplicar ciencia de los datos.

Debido a ello se va a elaborar con base en tres dimensiones las cuales son: ontológica, epistemológica y axiológica de la investigación.

La dimensión epistemológica trata de la creación de una metodología para la implementación de ciencia de los datos en la PYMES GO-LABS ENTERPRISES, esto implicará que se deba tomar una postura como involucrado en donde se investigará, analizará y diseñará una metodología de implementación de ciencia de los datos, la cual toma en cuenta detalles de eficacia, costos, facilidad de implementación, principales métodos y tecnologías que se utilizan en la actualidad.

La dimensión ontológica permitirá comprender cómo se aplica o desarrolla lo que es la ciencia de los datos, cómo se estructura la ciencia de datos y qué elementos son necesarios para poder realizar ciencia de datos. Se observa en la Figura 23 la manera muy abstracta lo que compone la ciencia de datos.

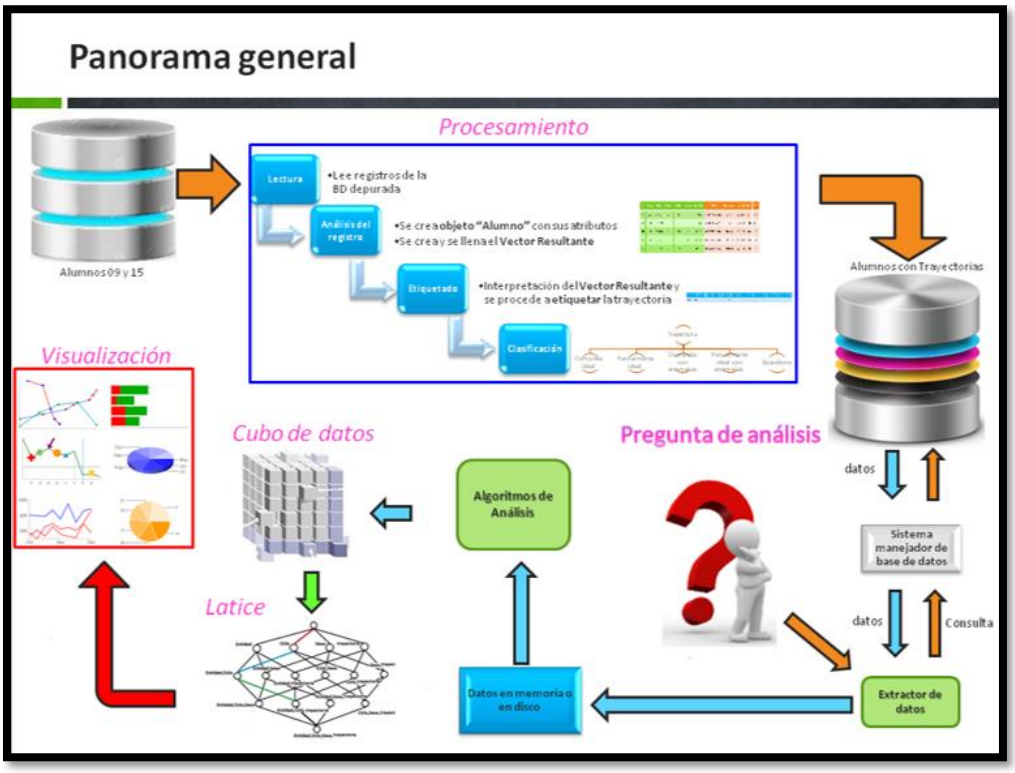
Figura 23 Ciencia de datos



Fuente: INLab FIB (Obiols, 2015)

Pero en la figura 24 se observa cuál es proceso que se debe desarrollar para poder obtener y aplicar la ciencia de los datos

Figura 24 Aplicación de ciencia de datos



Fuente: Red de Computación del IPN (Guzmán Arenas, Martínez Luna, & Orantes Jiménez, 2015)

La dimensión axiológica de esta investigación consistirá principalmente en el costo de la solución, facilidad para implementación e integración en cuanto a la metodología para aplicación de ciencia de datos, esto es un poco complejo porque no existe una referencia para ciencia de datos como un todo, sino como cada uno de los elementos y por eso para el proyecto se hará un estimado y se tomará en cuenta la información de cada uno de los elementos por separado para tratar de unificarlos.

Dada la situación que no existen datos para la aplicación de ciencia de datos como un todo, en la tabla 3 se trata de estimar un tiempo y un valor para cada uno de los elementos de la ciencia de los datos.

Tabla 4 Esfuerzo por tiempo según rubro

Rubro	Valor	Opciones
ETL	25%	Cantidad de horas: 240 horas= 25% 380 horas = 15% 480 horas = 1%
Data Warehouse	25%	Cantidad de horas: 240 horas= 25% 380 horas = 15% 480 horas = 1%
Minería de datos	50%	Cantidad de horas 480 horas = 50% 960 horas =25% 1440 horas = 1%

Fuente: Diseño propio.

3.3.4 Diseño

Se tomará como base el diseño exploratorio en donde primero se investiga y analiza el proceso de aplicar ciencia de datos, para posteriormente determinar cuál es el proceso necesario y que mediciones se realizan para determinar si el modelo diseñado cumple las expectativas esperadas sobre la aplicación de ciencia de datos.

3.3.5 Población y Muestreo

No existe una muestra debido a que se va a aplicar en una sola organización por lo que la población corresponde a la organización es decir el 100% es abarcado para la realización de la aplicación de la metodología.

3.3.6 Instrumentos de Recolección de Datos

Se utilizará dos instrumentos de recolección de datos, una es la observación que permite tener una perspectiva de la metodología y el otro sería el grupo focal para tener una perspectiva de algunas personas del perfil requerido sobre lo que les parece la metodología de aplicación de ciencia de datos

3.3.7 Técnicas de Análisis de la Información

Para el análisis de la información recolectada por medio de la observación se realizará por medio de diagrama de flujo y para la obtenida por medio del grupo focal se analizará por medio de la técnica de árboles causa efecto.

3.3.8 Estrategia de Desarrollo de la Propuesta

Para el desarrollo de la propuesta se utilizará metodologías ágiles como SCRUM y CRISP-DM, además de apoyarse en ciertos principios de gestión en las cuales las responsabilidades se repartirán. En este caso para fundamentar que la metodología diseñada es adecuada, se implementará de la mano con el desarrollo del proyecto en la PYMES GO-LABS ENTERPRISES en donde cada proceso diseñado se pondrá en práctica con el fin de verificar su usabilidad y funcionamiento de la aplicación de ciencia de los datos.

4 Capítulo IV: Guía de la implementación de los componentes del modelo.

Este capítulo contiene el diseño de la Metodología DASCI-SC, la cual surge debido a la necesidad de que no existe una metodología que explique cómo debe realizarse el desarrollo o la aplicación de la ciencia de los datos en Pymes, siendo la ciencia de los datos de gran importancia para mejorar el funcionamiento de las pequeñas organizaciones.

4.1 Definición de la metodología “DASCI-SC”

La metodología que se propone de tesis tiene la intención de ajustarse a la realidad de las PYMES en organizaciones tecnológicas. Esto a raíz de lo que se indica en el artículo (Gameiro, 2011), en el cual se establece que el nivel de madurez de manejo de la información de la empresa es el requerimiento básico en la toma de decisiones, por lo que, a mayor nivel de madurez de manejo de información, se requiere una mejor toma de decisiones. De acuerdo con la investigación realizada, diversas fuentes consultadas concluyen en que se debe mantener la sencillez en la solución que se proponga para una Pyme, así mismo no debe de representar un incremento en los costos de la empresa. Debe de ser amigable y que permita la optimización de los procesos actuales. Según el estudio realizado por Sadok y Lesca, existen 7 condiciones necesarias de aceptación para una buena implementación de una solución de inteligencia de negocios (Sadok & Lesca, 2009):

- Debe ser una solución simple y amigable.
- Debe propiciar el almacenamiento de los datos.
- Debe estar basado en el uso de fuentes de datos estructurados y no estructurados.
- Debe considerarse la exploración e interpretación de datos de varias fuentes.
- Debe ofrecer un mecanismo que permita la reducción de tiempos y costos.

- Debe fortalecer la implementación del conocimiento para la interpretación de los datos.
- Debe proporcionar resultados de manera eficaz y veraz.

  Otro aspecto importante por tomar en cuenta para la propuesta de metodología DASCI-SC.

  - Una solución de ciencia de datos, debe considerarse como no muy complicada y se debe contemplar una capacitación que permita un mejor entendimiento de la solución a implementar.
  - Se debe elegir el software correcto, que no genere más problemas de los que solucione, así como se debe guardar especial cuidado con la calidad de los datos y debe dar tiempos de respuesta satisfactorios.
  - Debe contener las funcionalidades que requiera el usuario final.

4.2Fases de la metodología DASCI-SC

Las fases de la metodología se presentan en la figura 25, en donde se observa un ciclo, donde las fases son dependientes una de la otra, esto permite que al final de realizarse el ciclo se tenga una retroalimentación y se proceda de nuevo a mejorar lo que en la ejecución del ciclo no dio el resultado esperado por parte del negocio.

Figura 25 Ciclo de Metodología

Fuente: Diseño propio

De acuerdo con el nivel de manejo de la información por parte de la empresa, se recomienda la implementación de cada fase.

Esta metodología al estar diseñada para el sector Pyme se requiere de un conocimiento del entorno donde están inmersas las empresas que se encuentran en este mercado. En Costa Rica la distribución de las PYME por sector económico en el año 2016 se concentraron en los sectores económicos de servicios (43.6%), comercio (39.6%), industria (11.9%) y tecnologías de información (4.8%) manteniendo una distribución muy similar a la del año 2015.

Figure 26 Distribución de PYME según sector económico, años 2015-2016

Fuente DIGEPYME, MEIC con datos del DEE-INEC (MEIC, 2016)

En la tabla 4 se muestra la clasificación de los niveles de madurez del manejo de la información que se pueden en las organizaciones de acuerdo con la aplicación de la ciencia de los datos.

Tabla 5 Nivel de madurez según manejo de datos

Nivel de madurez de manejo de información	Descripción
Alto	Tiene un manejo de datos para la toma de decisiones como medio para orientar hacia los posibles resultados
Medio	Tiene un manejo de datos que da orientación a las decisiones, pero es bastante limitado.
Bajo	No tiene ningún manejo de datos

Fuente: propio.

4.2.1 Fase 1 Negocio

En esta sección el objetivo primordial, es tener un conocimiento del negocio y de ser necesario trabajar a la par de los expertos del negocio para establecer los conceptos primordiales en los cuales se intentará ayudar a la organización sobre su rumbo para establecer las estrategias que permitan orientarse a objetivos que planteen y definan las métricas necesarias que apoyen en la determinación de los cumplimientos de los objetivos ya establecidos. El ejecutar esta fase permite llegar a un entendimiento del negocio.

4.2.1.1 Planeación estratégica

La planeación estratégica permite que las empresas puedan definir claramente quiénes son, a dónde se quiere ir y cómo llegar. Todo ello lo deben lograr por medio de la definición de los objetivos y metas por alcanzar en un periodo de tiempo que darán las pautas iniciales de un camino por seguir.

4.2.1.1.1 Misión

La misión permite delimitar quiénes somos y qué hacemos y así describir en un pequeño párrafo los principales servicios y/o productos que se ofrecen, así como el mercado al que se dirigen. Es un reflejo del presente y es la carta de presentación con los clientes. La misión permite realizar la priorización de objetivos y actividades que la ciencia de datos deber responder. También esto aclara al experto de ciencia de datos sobre lo que busca y necesita la organización en dicho momento.

4.2.1.1.2 Visión

Con la misión de la empresa, se puede establecer su visión. La visión es una vista al futuro de lo que se espera lograr con la empresa y el crecimiento que esta tenga, todo descrito en un pequeño párrafo. Esta definición es crucial para la definición de los “cómo” que guiarán a la empresa. Esto dará orientación al experto y los guiará para determinar las necesidades de información que la ciencia de datos en un futuro deberá poder apoyar.

4.2.1.1.3 Objetivos

A partir de la visión se puede definir los objetivos que servirán de base para hacer que la visión de la empresa se vuelva una realidad. Los objetivos se definen en frases cortas que se escriben de manera infinitiva. Son descritos de manera genérica y su finalidad es meramente de guía y para tener plasmado el camino que se debe seguir para cumplir la visión, de esta manera se puede consultar cada vez que sea necesario.

Los objetivos son los que indican cuáles serán las principales mediciones que se deben de realizar, y también cuáles son las primeras necesidades de información.

4.2.1.1.4 Metas

Las metas están relacionadas con los objetivos, son frases cortas que complementan los objetivos incluyendo valores numéricos que faciliten la medición del cumplimiento de los objetivos por medio del cumplimiento de las metas. Las metas son más específicas que los objetivos.

De las metas se obtendrá lo que se debe medir por medio de KPI específicos para cada área que se evaluará en el desempeño.

4.2.1.1.5 Factores Críticos de éxito

Los factores críticos de éxito son aquellos que forzosamente deben o no presentarse para el cumplimiento de las metas y por ende para el logro de los objetivos definidos.

Estos serán los obstáculos a los cuales el experto de la ciencia de datos se enfrentará para lograr objetivos, para ello se debe realizar un manejo de riesgos para que no exista imprevistos o el impacto de algún elemento que lleve al fracaso.

4.2.1.2 Procesos de negocio

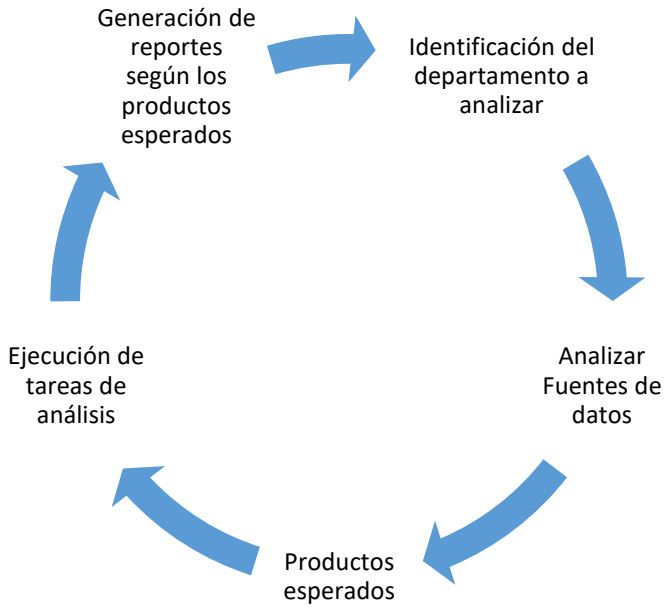
Los procesos de negocio son las actividades de la empresa que guardan cierta relación y que requieren de un insumo para poder generar algún resultado. La definición de los procesos de negocio es de gran utilidad para la definición de estrategias y para generar las mejoras necesarias para la operación de la empresa.

Estos conjuntos de actividades están íntimamente ligados con la operación de la empresa tomando en cuenta las principales funciones que debe cumplir de acuerdo con la misión definida.

En los procesos de negocios es en donde se obtendrán los datos de las operaciones diarias que serán el insumo de datos en donde se describen quiénes son los dueños de datos y la forma en que son utilizados por parte de la organización, de estos se obtendrá marcadores importantes del estado actual de la organización. En este proceso se recomienda utilizar diagramas de flujos de información.

Como se puede observar en la siguiente figura, se debe seguir todo un flujo de la información dentro de la organización para garantizar indicadores efectivos en cada uno de los departamentos.

Figura 27 Ciclo de flujo de información



Fuente: Diseño propio

4.2.1.3 Necesidades de información

El conocer los procesos de negocio dará claridad en cuáles son las necesidades de información que se requieren en la empresa, lo cual también permitirá que se definan las fuentes de información necesarias para la construcción de una solución de inteligencia de negocios. Las necesidades están relacionadas con los insumos que alimentan a los procesos de negocio para el logro de los objetivos planteados por la empresa.

Para realizar esta tarea de identificación es necesario analizar el flujo de la información vinculado con los objetivos y metas del negocio planteados. Es necesario realizar una agrupación para abarcar cada objetivo y meta de este.

4.2.1.4 KPI

Los indicadores de desempeño clave (KPI), permiten medir el estado en el que se encuentra la empresa, para determinar si se están logrando los objetivos definidos. Este indicador es esencial para la toma de decisiones con el fin de realizar ajustes en las estrategias ejecutadas, para estar en un proceso de mejora continua se debe tomar en cuenta los datos del histórico entre unos 3 a 5 años, debido a que las estrategias van cambiando con el tiempo.

4.2.1.5 Lista de prioridades

Una vez que ya se cuenta con el conocimiento del negocio se procede a realizar una lista de prioridades de los indicadores de desempeño clave, para tener claro el orden en el cual se trazará la solución de inteligencia de negocios, de tal manera que se pueda segmentar la solución con la finalidad de poder hacer un desarrollo ágil. Las prioridades las deben definir los líderes de negocio por lo cual con ellos es que se debe negociar como se deben hacer las entregas y que se considerará un entregable.

La priorización debe realizarse de acuerdo con las perspectivas que comprende el cuadro de mando integral como son aprendizaje, innovación, procesos, clientes rentabilidad o financiero.

4.2.2 Fase 2 Planeación

En esta sección se definirá los aspectos relacionados con la planeación para el desarrollo de la solución de inteligencia de negocios. Esta fase de la metodología es recomendada para aquellas empresas cuyo nivel de madurez según manejo de información, es medio o alto, esto debido a que es a partir de estos niveles es recomendable invertir tiempo en organizar los proyectos que se presenten. Cuando una empresa tiene un nivel de manejo de información bajo, tal vez no sea tan recomendable la inversión del tiempo en la planeación del desarrollo de la solución debido a su baja complejidad. En la tabla 5 se especifica a mayor detalle la clasificación de los niveles de madurez según el manejo de información.

4.2.2.1 Alcance

Primeramente, se definirá el alcance del proyecto, estableciendo qué aspectos estarán incluidos en el desarrollo, tratando de delimitar de manera adecuada el proyecto para evitar problemáticas innecesarias como carga errónea de datos, limpieza poco depurada de datos, etc. En este apartado se deberán incluir los entregables que se comprometerán según el marco regulatorio que se establece en Costa Rica bajo la Ley N. º 8262 en los artículos número uno y tres de la “Ley de fortalecimiento de las pequeñas y medianas empresas y sus reformas.”

4.2.2.2 Actividades

Una vez definidos los entregables se deben definir las actividades que se deben realizar para poder generar dichos entregables como lo son el análisis, diseño y desarrollo, estas actividades deberán incluir un orden que permita conocer cómo se va a realizar cada actividad establecido por el modelo KDD y CRISP-DM. De la misma manera las actividades deberán ser precedencia una de otras para conocer la dependencia que exista entre las actividades.

4.2.2.3 Recursos

Definidas las actividades se deben definir los recursos que se necesitan para la ejecución de dichas actividades. De esta manera se debe designar responsables para las actividades, así como roles y responsables para gestión del proyecto. También se debe especificar qué recursos materiales se requieren, así como los recursos técnicos necesarios.

4.2.2.3.1 Tablas de recursos

Se debe realizar una tabla para definir cuáles serán las herramientas y personal a utilizar en la implementación de la metodología, esto con el fin de tener un presupuesto de costo que trate de evitar contratiempos. Además, ayudará a crear un presupuesto previo para conocer el monto por invertir entre los distintos recursos.

La siguiente tabla permite especificar y comprender de mejor forma cuales son los recursos necesarios.

Tabla 6 Costos de Recursos

Recurso	Precio	Licencia	Horas

Fuente: Diseño propio

4.2.2.3.2 Perfiles de los recursos necesarios

Es necesario definir cuáles son las cualidades que preferiblemente se requieren para aplicar ciencia de datos en las PYMES, por lo cual es necesario los siguientes roles.

4.2.2.3.2.1 Perfil del analista de datos:

Se requiere un profesional que puede resolver problemas de distintas disciplinas, haciendo uso de análisis de datos en la extracción de conocimiento desde distintas fuentes de información. Además, debe tener una gran capacidad de

comunicación y capacidad analítica para encontrar patrones, por medio de la aplicación de análisis estadístico, métodos cuantitativos y modelos matemáticos y computacionales; con el fin de generar información para el negocio y dar valor agregado a los procesos que se realizan en la organización y a tomar mejores decisiones basadas en datos.

Habilidades:

- ✓ El perfil debe poseer conocimiento en matemáticas enfocada a la estadística (descriptiva e inferencial).
- ✓ Debe conocer el manejo de repositorio de datos (big-small data) por medio de depósito de datos de distintas fuentes (SQL).
- ✓ Enfoque en inteligencia artificial y modelos matemáticos asistidos por computación.
- ✓ Conocimiento de herramientas en el proceso de inteligencia de negocios y análisis de datos
    - Repositorio de datos (Datos estructurados y no estructurados)
    - Visualización (Pentaho, Power BI, Tableau)
    - Análisis de datos (R, WEKA, Python)
- ✓ Debe tener la habilidad de programar.

4.2.2.3.2.2 Perfil del experto del negocio

Se requiere un profesional que conozca el funcionamiento de la empresa, este debe saber hacia dónde está orientada la organización, lo que quiere y cómo es posible llegar a obtener los resultados esperados. Debe de conocer cuál es la estrategia organizacional y cuál es el plan estratégico, debe poder manejar el negocio y conocer que reportes, KPIs e información se requiere para mejorar los procesos de negocio.

Habilidades

- ✓ Facilidad de comunicación.
- ✓ Destreza para identificar y proponer oportunidades de negocios.
- ✓ Decidir las características funcionales del producto o servicio.
- ✓ Capacidad para transmitir la visión del proyecto al equipo.

- ✓ Proteger los intereses del negocio.
- ✓ Sugerir cambios y adaptaciones de forma acertada.
- ✓ Resolver los conflictos de interés del negocio.

4.2.2.4 Restricciones y supuestos

En los proyectos de aplicación de ciencia de datos siempre va a existir una posibilidad alta hacia hechos y restricciones de los cuales no se tiene total certeza de los eventos que pueden ocurrir durante su ciclo de vida. Es por ese motivo que se debe estimar los diferentes supuestos y restricciones para determinar el camino por seguir de un proyecto.

La tabla 5 muestra las similitudes y diferencias entre estos dos términos:

Tabla 7 Comparación entre los supuestos y restricciones

	Supuestos	Restricciones
Característica	Condición, Circunstancia o evento.	Condición, Circunstancia o evento.
Impacto	Permite al proyecto proceder.	Restringe y limita la ejecución del proyecto.
Proceso	Debe ser analizado y monitoreado para asegurar validez y relevancia a medida que el proyecto procede.	Debe ser identificado e incorporado en los planes del proyecto para asegurar que este sea realista.

Fuente: Introducción a la Gerencia de Proyectos. (Jaramillo Parra, 2015)

4.2.2.5 Riesgos

Es necesario identificar los riesgos que pueden interferir con el cumplimiento de los tiempos establecidos para las actividades definidas, estos riesgos tienen que ser evaluados para determinar las probabilidades de que ocurran y el impacto que puede tener sobre el proyecto. Así mismo se debe asignar un responsable para cada riesgo, así como establecer actividades que permitan mitigar los riesgos identificados. Para el

manejo de los riesgos existen diferentes metodologías que permiten la gestión de este, se recomienda la guía de PMBOK.

4.2.2.6 Plan

Finalmente se debe plasmar en un documento las actividades con sus tiempos, responsables, porcentajes de progreso, fechas compromisos, dependencias. Este documento deberá facilitar el dar seguimiento a los avances del proyecto, permitiendo actualizar de manera fácil el estado de las actividades.

4.2.3 Fase 3 Desarrollo

Con un plan para la ejecución de las actividades necesarias para desarrollar una solución de inteligencia de negocios. Se precede con la ejecución de dichas actividades que en grandes rasgos se describirán en los siguientes apartados. De acuerdo con el nivel de manejo de información de la empresa, serán las tareas que se recomiendan que se ejecuten. En el caso de un nivel de manejo bajo, solo se requiere del modelado de la información y se presenta en hojas de cálculo, con lo cual se puede jugar con la información a través de tablas dinámicas. En el caso de un nivel de manejo medio, es recomendable dar un paso más después del modelado, es decir se recomienda la construcción del Data Warehouse, así mismo es recomendable implementar cubos que permitan un manejo de la información multidimensional, la presentación de estos resultados se recomienda que sean en web o sistemas de visualización. En el caso de un nivel de manejo alto, se recomienda la implementación de minería de datos como un paso más delante de la creación de cubos, recomendándose la presentación en web o sistemas de visualización avanzados.

4.2.3.1 Preparación Técnica

Se debe realizar una preparación de los ambientes para poder iniciar con la ejecución de las actividades designadas durante la planeación, pero para lograr esa preparación se debe realizar un análisis de las herramientas que sean más adecuadas para el desarrollo de la solución de inteligencia de negocios de acuerdo con los volúmenes de información y a la capacidad de compra que se tenga.

4.2.3.1.1 Selección Herramientas

Cuando se trata de desarrollos ambiciosos, con grandes riesgos y presupuestos holgados, se puede considerar el uso de herramientas propietarias por la solidez que simbolizan, aunque representan altos costos, sin embargo, para aquellos casos en los que se cuenta con un presupuesto muy reducido se puede considerar el utilizar herramientas open source. Cuando se recurre a herramientas open source es necesario hacer una buena selección pues no todas las herramientas disponibles son adecuadas (Gameiro, 2011).

Existen diversas herramientas en el mercado que permiten desarrollar soluciones de inteligencia de negocios. Para elegir las herramientas es necesario considerar los requerimientos para su instalación, las ventajas y desventajas que representa cada herramienta.

Se debe generar el documento que refleje las ventajas y desventajas de usar la herramienta seleccionada y debe contener las firmas de aceptación que sirva de respaldo. Así mismo este documento deberá contener los requisitos de hardware y software para la instalación, así como una guía para la instalación.

4.2.3.1.2 Instalación

Una vez seleccionada la herramienta se debe realizar la instalación de las herramientas tomando en consideración los requerimientos de software y hardware para que esta instalación sea exitosa. En primera instancia se deben cubrir los requerimientos de hardware, por lo que se deben realizar las compras necesarias para cumplir con dichos requerimientos. Con los requerimientos de hardware se procede con la instalación de los pre-requisitos para que funcione adecuadamente la herramienta para finalmente realizar la instalación de la herramienta.

4.2.3.2 Construcción de Data WareHouse

Finalmente se genera los scripts con la estructura del Data Warehouse para construirlo con el manejador de base de datos seleccionado en la fase de preparación técnica, se toma en cuenta la sintaxis que utilice dicha herramienta. Como documentación en este punto se deberá generar el diccionario de datos que permita tener el entendimiento del objetivo que se persigue con cada tabla y cuáles son sus características principales.

4.2.3.3 Datos

El objetivo de esta sección es trabajar con los datos analizados para diseñar y desarrollar el proceso ETL, el cual se terminará poblando el data warehouse. Esta sección es una de las bases para los siguientes pasos de la metodología, pues la salida de este paso será el data warehouse poblado el cual es una entrada para el resto de los procesos.

4.2.3.3.1 Análisis de datos

En este punto se parte del análisis de la fuente de datos para conocer el origen de los datos y comenzar a analizar y planear las transformaciones que son necesarias para que se inserten en el Data Warehouse. Se realizan los mapeos necesarios entre las fuentes de datos y las tablas finales contenidas en el Data Warehouse para tener una visión clara del destino que tendrá cada dato que será utilizado.

4.2.3.3.2 Desarrollo proceso ETL

Ya que se cuenta con el conocimiento del origen y destino de los datos se procede con el desarrollo del proceso ETL, el cual se pide tener bien identificadas estas partes. Se recomienda primeramente hacer una extracción de los datos sin mayores transformaciones e insertarla en una base de datos de paso (staging área). En el cual se tenga la copia fiel de los datos extraídos en la base de datos de paso, se procederá a realizar las transformaciones necesarias para lo cual se puede apoyar de tantas tablas temporales como se considere necesario. Ya que se tienen los datos como se desean para introducir al DWH. Se debe tomar en cuenta que un ETL cuya finalidad

sea la de poblar dimensiones, deberá considerar la actualización de dichas dimensiones y la inserción de nuevos registros, esto es recomendable para tener siempre los datos disponibles, aunque el proceso ETL falle. En el caso de poblar una tabla de hechos generalmente solo se realizan inserciones de los nuevos registros. Se debe tomar en cuenta la creación de un proceso ETL para la población histórica de las tablas de hechos del Data Warehouse y otro proceso ETL para las cargas periódicas las cuales se planean generar. Se debe documentar estos procesos para saber el camino que siguen los datos y las transformaciones que se le hacen a los datos.

4.2.3.3.3 Pruebas

En este punto, se cuenta con los procesos ETL a los cuales se le realizan las pruebas necesarias para comprobar que funcionan adecuadamente. Una vez que se han probado y se valida que funcionan adecuadamente se considera que están listos para ser usados en un ambiente productivo.

4.2.3.3.4 Población histórica DWH y calendarización ETL

Finalmente se realiza la carga de información en el Data Warehouse y se utiliza el ETL de carga histórica y se realiza un proceso que permita programar el proceso ETL para las cargas periódicas, para que sea ejecutado con la frecuencia que se considere necesario y que de esta manera esté poblando el Data Warehouse con la información más reciente.

4.2.3.4 Minería

Cuando se tiene identificado algún problema que requiera la identificación de patrones de la información se puede hacer uso de la minería de datos para apoyar en la toma de decisiones de las empresas.

La minería de datos debe basarse en el proceso de las siguientes etapas de la minería que son tomadas como guía de la metodología de CRISP-DM o Cross Industry Standard Process for Data Mining

4.2.3.4.1 Análisis del problema

En el análisis del problema se debe comprender entender los objetivos y requisitos desde un punto de vista de negocio, como paso previo a la definición del problema de Minería de Datos.

4.2.3.4.2 Entendimiento del problema

Se recolectan los datos, se exploran, se detectan problemas con la calidad de los mismos, y se obtienen los primeros insights, subconjuntos de datos, primeras hipótesis.

4.2.3.4.3 Exploración de los datos

Se realiza un análisis preliminar de los datos, obteniendo unas primeras conclusiones sobre su forma, tendencias, etc., para guiar a decidir qué camino seguir.

4.2.3.4.4 Modelo de Minería

En esta etapa se debe de seleccionar y aplicar las técnicas de modelado, también se debe calibrar los parámetros en búsqueda de los mejores resultados.

4.2.3.4.5 Evaluación

Se evalúan los modelos y se revisan los pasos seguidos para la construcción de los modelos, en relación a los objetivos de negocio. Para ello es necesario contar con el personal experto para determinar que modelos se acercan más a la realidad de la empresa, de manera tal que se puedan tomar la totalidad de los datos para generar los reportes necesarios para la toma de decisiones

4.2.3.4.6 Resultados

El modelo se descubre y se aplica. También puede llegar a desplegarse en un sistema o entorno de producción, o al menos se genera un entregable que el cliente pueda usar.

4.2.3.5 Cubo MOLAP

Para poder tener un análisis de los datos más potente y significativo es importante explotar el Data Warehouse con herramienta de análisis multidimensional,

para lo cual se genera cubos de información que permiten tener los datos lineales del Data Warehouse en un esquema multidimensional lo que permite generar consultas más enriquecedoras para la toma de decisiones.

4.2.3.5.1 Dimensiones y Jerarquías

Dentro de un modelo multidimensional uno de los puntos importantes por desarrollar son las dimensiones y las jerarquías involucradas en el modelo. Las dimensiones son los elementos cualitativos que permitirán evaluar la información. Son aquellos elementos que le dan sentido a los hechos. Las jerarquías es la manera en la que se puede navegar a través de la información. Dentro del modelo multidimensional es importante definir de manera correcta las dimensiones y sus jerarquías para poder consultar los indicadores necesarios en distintos niveles de agregación.

4.2.3.5.2 Indicadores

Los Indicadores son las métricas utilizadas en el modelo multidimensional y son todos aquellos elementos que nos permiten hacer mediciones de manera cuantitativa. Estos indicadores surgen a partir de las tablas de hechos del Data Warehouse y consiste en todos aquellos datos numéricos, estos son conocidos como indicadores base. Estos indicadores base permiten realizar los cálculos necesarios para generar los indicadores principales de rendimiento, los cuales son claves para la toma de decisiones.

4.2.3.5.3 Cubos

Finalmente se genera la estructura que, uniendo las dimensiones y los indicadores mediante un esquema de estrella, formará un cubo. Una vez formada esta estructura se podrán realizar las consultas multidimensionales que enriquecerán los reportes que se puedan generar para la toma de decisiones.

4.2.3.6 Presentación

El usuario final podrá explotar el Data Warehouse a partir de los reportes que se le presenten los cuales generalmente se presentan en formato web. Existen tres

maneras de presentar el resultado del proceso de inteligencia de negocios al usuario final.

4.2.3.6.1 Reportes

En los reportes se incluye la información con cierto grado de detalle. Generalmente se presentan como tablas pivote en las cuales se pueden mezclar las dimensiones e indicadores que se lean de los cubos.

4.2.3.6.2 Dashboards

Es un resumen de los indicadores de rendimiento, mostrando sus resultados de manera visual mediante gráficas, tacómetros, entre otros. De manera que a primera vista se pueda observar el estado general de la empresa. Esta más enfocado a medir el desempeño de los procesos. Generalmente un dashboard engloba la información por un tema específico.

4.2.3.6.3 Cuadros de mando

Contiene un resumen de la información de los indicadores de rendimiento clave, mostrando mediante semáforos su estado. Este tipo de reportes es generado con una temporalidad más amplia debido a que permite comparar períodos de tiempo, por ejemplo, cuatrimestres. Está más enfocado en medir el estado del cumplimiento de la estrategia de la empresa. Debe incluir todos los aspectos que permitan conocer el estado global de la empresa.

4.2.4 Fase 4 Validación

Cuando el usuario final es capaz de ver la información mediante un reporte, dashboard o cuadros de mando, entonces puede comenzar a validar si la información que visualiza es correcta. Este proceso es de suma importancia para saber si el proceso completo es correcto.

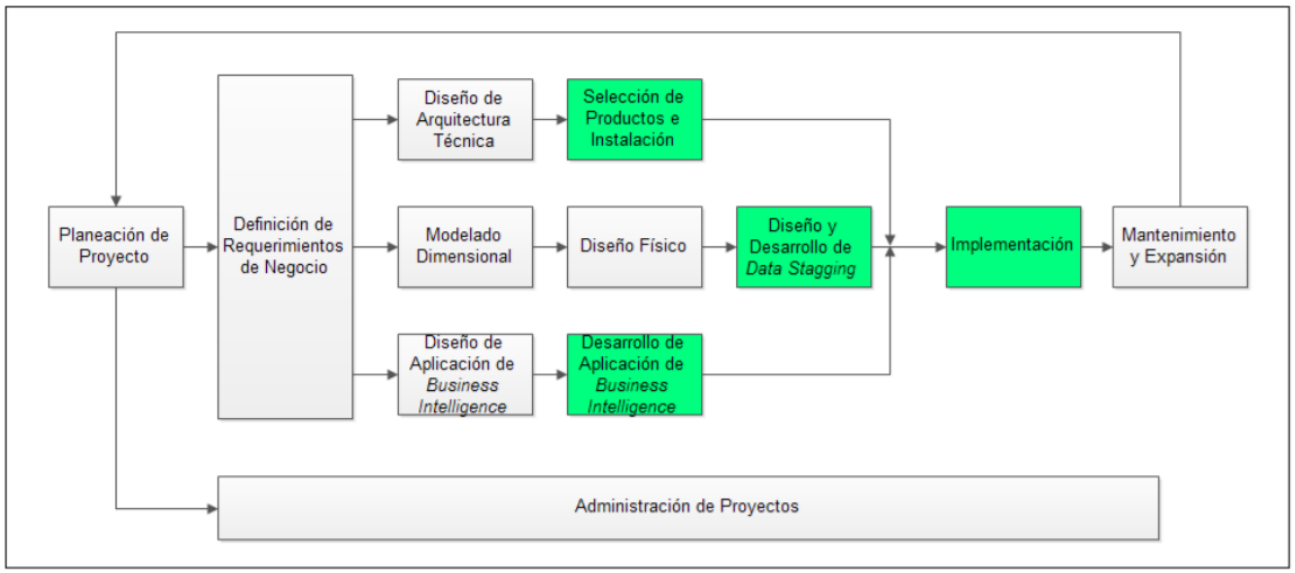
Una vez que el usuario indica que la información es correcta se procede con la implementación, en caso contrario se hacen los ajustes necesarios para hacer las correcciones necesarias. Es recomendable hacer validaciones entre cada bloque del

proceso de desarrollo para facilitar el proceso de validación y que este no sea tan pesado al final del desarrollo de la solución.

4.2.5 Fase 5 Implementación

La implementación representa la convergencia de la tecnología, los datos y las aplicaciones de usuarios finales accesible desde el escritorio del usuario. Hay varios factores extras que aseguran el correcto funcionamiento de todos estos elementos, entre ellos se encuentran: la capacitación, el soporte técnico, la comunicación y las estrategias de retroalimentación. Todas estas tareas deben tenerse en cuenta antes de que cualquier usuario pueda ingresar al Data Warehouse. Con el visto bueno del usuario, se puede continuar con la implementación de la solución en un ambiente productivo. Una vez implementado es recomendable monitorearlo y dar soporte a la solución hasta que esta funcione de manera estable.

Figura 28Diagrama de planeación de proyectos



Fuente: (Universidad Autónoma de México, 2014)

Implementación como convergencia de la tecnología, los datos y las aplicaciones

5 Capítulo V

5.1Fase 1 Negocio: Track It

Go-Labs es una empresa de desarrollo de software localizada en la Zona Norte de Costa Rica, precisamente en Ciudad Quesada de San Carlos. Fundada en enero de 2014, rápidamente ha logrado posicionarse a nivel nacional e internacional desarrollando soluciones de software innovadoras para sus clientes tanto a nivel nacional como internacional. Entre sus áreas de especialización se encuentra el desarrollo Web, diseño de bases de datos, aplicaciones móviles en los principales sistemas operativos del mercado entre otros.

El mundo es ahora un gran mercado para las Pymes, debido a que la apertura hace que se deba competir, no sólo con otras empresas del país, sino con empresas de otros países y continentes, lo cual causa diversas alteraciones en las pequeñas y medianas empresas, esto ha llevado a afrontar los nuevos desafíos, para lo que se requiere iniciativas de reestructuración del sector, tecnología apropiada, mano de obra calificada, producción a gran escala y recursos financieros.

Track It es una aplicación móvil para las plataformas Android y IOS para la administración de envíos de paquetes rastreables por medio de códigos QR y geolocalización. Entre las ventajas que posee se encuentra:

- Notificaciones por Email.

- Notificaciones por SMS.

- Historial de ubicación de paquetes.

- Suscripciones de notificaciones de paquetes.

5.1.1 Planeación estratégica

La planificación estratégica es el proceso de determinar cómo una organización puede hacer el mejor uso posible de sus recursos (fuerza de trabajo, capital, clientes, etc.) en el futuro.

Mediante la planificación estratégica se define la estrategia o dirección, estableciendo las posibles vías mediante las cuales se puedan seguir cursos de acción particulares, a partir de la situación actual. Además, esta proporcionará una dirección general en estrategias financieras, estrategias de desarrollo de recursos humanos u organizativos, en desarrollos de tecnología de la información y crear estrategias de marketing para enumerar tan solo algunas aplicaciones.

5.1.2 Misión

“Ser la mejor empresa de software de la región norte”. La misión de esta organización requiere de una mejora en los procesos, ingresos, productos, marketing, etc., para lo cual ha decidido apoyarse en la ciencia de los datos para lograr dicho objetivo.

5.1.3 Visión

“Convertimos ideas en soluciones de software con calidad mundial”. Dada la misión que posee esta organización requiere apoyarse en las últimas tendencias tecnológicas para brindar productos de calidad al mercado apoyado en el uso de la ciencia de los datos.

5.1.4 Objetivos

De acuerdo con la misión y visión de la empresa se definieron los siguientes objetivos:

- ✓ Incrementar las ventas en 15%.
- ✓ Aumentar en 10% la atracción de clientes a la categoría Premium.
- ✓ Incrementar la oferta de paquetes de órdenes en un 20%.

5.1.5 Metas

Siguiendo los objetivos planteados, se definieron las siguientes metas, que permitirán lograr los objetivos establecidos.

- ✓ Incrementar las ventas en un 15%.
- ✓ Incrementar la publicidad segmentada a los clientes en un 20%.
- ✓ Mejorar el servicio al cliente por medio de un aumento de la satisfacción de los clientes en un 10%.

5.1.6 Factores Críticos de éxito

Para que la empresa considere que las acciones que planea ejecutar en el futuro próximo, será necesario proceder de acuerdo con los resultados que se obtengan a partir de la toma de decisiones actuales.

Por lo tanto, uno de los principales criterios será que el administrador general sea capaz de tomar decisiones a partir de los reportes que se generen con el desarrollo del proyecto.

Otro criterio será el cumplimiento de los objetivos de acuerdo con la información generada por las necesidades de información identificadas a partir de los objetivos definidos por la empresa.

Los factores críticos de éxito son aquellos que forzosamente deben o no presentarse para el cumplimiento de las metas y por ende para el logro de los objetivos definidos.

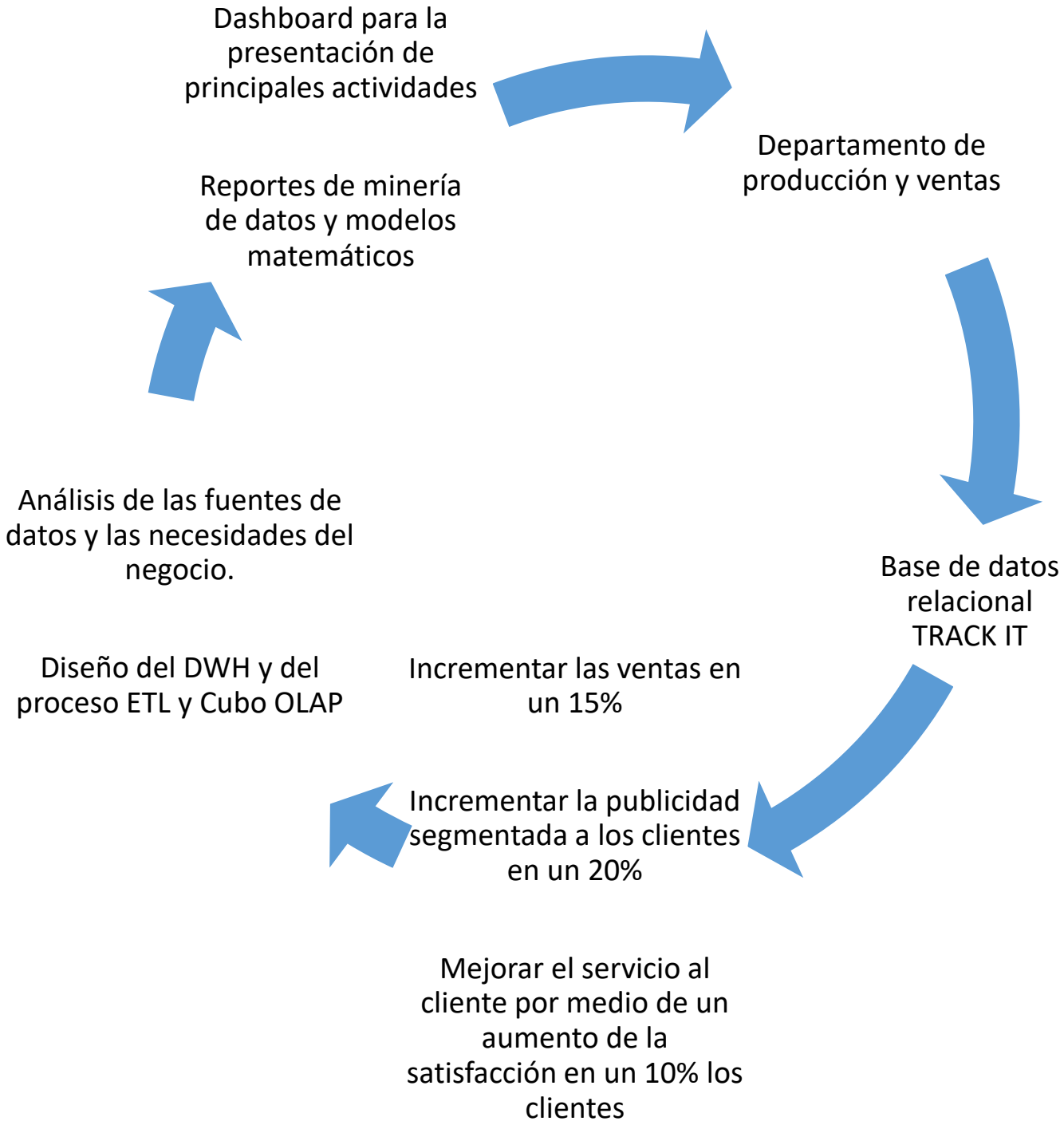
5.1.7 Procesos de negocio

Al ser una organización que ofrece un servicio de control de envió de paquetes de forma rastreable por medios de códigos QRS y geolocalización, que a trevés de un dispositivo móvil pueden hacer el seguimiento de cada paquete.

El producto que se vende es una cantidad de códigos QRS que se venden para ser impresos y pegados en los paquetes de diferentes productos.

El diagrama propuesto para el manejo de la información es el siguiente:

Figure 29 Diagrama de flujo de información



Fuente: Diseño propio.

5.1.8 Necesidades de información

Para la toma de decisiones es necesario previamente recolectar la información adecuada, la cual va a ayudar siendo una base indispensable que guiará para la adecuada toma de decisiones.

Para ello lo ideal es recolectar la información proveniente del sistema de Trackit, el cual proporcionará la información de las ventas, el trazado de los productos, los tiempos de entrega y las rutas tomadas. También se podrá extraer la información de los clientes para poder dar un seguimiento a estos y poder mejorar el servicio al cliente.

5.1.9 KPI

Los principales indicadores de rendimientos que se requieren son:

- Ventas
- Clientes
- Ofertas a los clientes

5.1.10 Lista de prioridades

Se define las prioridades para alcanzar los objetivos de acuerdo con las necesidades del negocio, las cuales se irán abarcando de manera progresiva para así lograrlos en tiempo y forma.

Las prioridades establecidas son:

- Desarrollo de informes necesarios para la evaluación de las ventas.
- Desarrollo de informes necesarios para la mejora de las ofertas a los clientes (mejora de marketing).
- Desarrollo de informes necesarios para la clasificación de clientes.

5.2 Fase 2 Planeación

La fase de planeación permite definir las bases, las actividades y recursos que van ser necesarios para aplicar la ciencia de los datos en las Pymes

5.2.1 Alcance

El proyecto por desarrollar contemplará el análisis, diseño y generación de un prototipo de solución de inteligencia de negocios, tomará en cuenta lo siguiente:

- ✓ Análisis, diseño y construcción de un DWH.
- ✓ Análisis, diseño y construcción del proceso ETL.
- ✓ Análisis, diseño y construcción de un reporte estratégico.

5.2.2 Actividades

Las actividades que se definen en la tabla 8 son tareas específicas para la implementación de la aplicación de ciencia de datos en la PYMES Go-Labs.

Tabla 8 Actividades para la aplicación de ciencia de datos

Actividades	Descripción
Análisis	Se realizará un análisis de las fuentes de datos y las necesidades del negocio.
Diseño	Se realiza el diseño del DWH y del proceso ETL, así como el reporte a generar.
Desarrollo	Se creará el DWH y su proceso de ETL y se creará el reporte.
Pruebas	Se harán las pruebas de implementación necesarias para verificar que el desarrollo funciona de manera adecuada.
Dashboard	Se hará un dashboard para la presentación de principales actividades

5.2.3 Recursos

Los recursos para elaborar este proyecto son:

- ✓ El personal.
- ✓ Datos.
- ✓ Equipo de hardware.
- ✓ Software.

En la tabla 9 se detalla el personal involucrado y el rol de cada uno.

Tabla 9 Recurso personal y sus roles correspondientes

Personal	Rol
Efrén Jiménez Delgado	Consultor de BI
Alexander Gutiérrez Cerdas	Consultor de BI
Carlos Rojas	Jefe de Tecnología
Jennifer Madrigal	Administradora

En la tabla 10 se especificará el precio total y la licencia de las herramientas por utilizar, además del costo del recurso humano.

Tabla 10 Cuadro de recursos técnicos

Recurso	Precio	Licencia	Horas
SQL Server Data Tools 2013	0	Free	
SQL Server 2014 Standard	$3,717	Pago	
R	0	Free	
Power View	$ 10 mensuales	Pago	
Efrén Jiménez Delgado	$ 6400		320
Alexander Gutiérrez Cerdas	$ 6400		320
Carlos Rojas	$ 1600		40
Jennifer Madrigal	$ 400		20

Fuente propia

5.2.4 Restricciones y Supuestos

Para la elaboración del proyecto es necesario que este deba estar orientado de acuerdo con los objetivos del negocio, para los cuales se definieron los siguientes requerimientos.

- ✓ Software: Se deberá contar con el software necesario para la elaboración del proyecto de inteligencia de negocios.
- ✓ Hardware: Se deberá contar el hardware necesario para soportar los requerimientos de las herramientas de software seleccionadas.
- ✓ Accesos: Se dispondrá del acceso a la información necesaria a las fuentes de datos.

- ✓ Comunicación: se debe realizar una precisa comunicación entre los expertos del negocio y los consultores de inteligencia de negocios.

Para lograr los objetivos establecidos por el negocio, se definieron los siguientes supuestos:

- ✓ Se cuenta con la infraestructura idónea para poner el proyecto en producción
  - Hardware.
  - Software.
  - Red.
- ✓ Se cuenta con el apoyo del personal de la empresa para resolver cuestiones de definición de datos.
- ✓ Se cuenta con la información necesaria para la conexión a las fuentes de datos.

También se definieron las siguientes restricciones que permiten demarcar el alcance del proyecto, estas restricciones son:

- ➢ Se utilizará las herramientas de desarrollo seleccionadas.
- ➢ Se desarrollará la solución de acuerdo con las prioridades definidas por el negocio.
- ➢ Se creará el repositorio de datos según las definiciones iniciales.
- ➢ Se hará la limpieza de datos según lo establecido por el repositorio de datos.
- ➢ Se seguirá la metodología propuesta.

5.2.5 Riesgos

La identificación de los riesgos ha permitido definir ciertas acciones que mitiguen o en su efecto se minimice la probabilidad para que estos no sucedan. A continuación, se muestra un listado de riesgos y las posibles acciones de mitigación.

Tabla 11 Manejo de riesgos

Riesgo	Probabilidad	Acción	Responsable
Falta de tiempo de los colaboradores	Media	Concertar citas con colaboradores con anticipación	Negocio, Consultor
Fallas en conexiones de red con fuentes de datos	Media	Comunicar fallas al departamento de redes para dar pronta solución	Negocio (Proveedor Red)
Fallas equipo Hardware de desarrollo	Baja	Se deberán hacer respaldos semanales de avances y se deberá contar	Negocio (Proveedor HW), Consultor
Fallas con instalaciones y configuraciones de Software	Media	Se deberá contactar con los proveedores correspondientes y realizar la Investigación necesaria.	Consultor

5.2.6 Plan

Tabla 12 Plan de trabajo

Descripción de tareas	Duración	Fecha inicio	Fecha finaliza
Planificación del proyecto			
Identificación de componentes del modelo metodológico de ciencia de datos			
1. Identificación de roles y actividades de los técnicos de base de datos.			
2. Identificación de roles y actividades de los administradores de base de datos.			

3. Identificación de roles y actividades del personal en inteligencia de negocios.			
4. Identificación de roles y actividades del científico de los datos.			
5. Identificación de roles y actividades de los usuarios.			
6. Identificación de las prácticas de la metodología KDD.			
7. Identificación de las prácticas de la metodología BIM.			
8. Identificación de las prácticas de la metodología CRISP-DM.			
9. Identificación de las prácticas de la metodología SCRUM.			
10. Identificación de metodologías para crear DW.			
11. Identificación de BD para el repositorio de datos (DW).			
12. Identificación de áreas de BI.			
13. Identificación de datos masivos.			
Describir los componentes del modelo en la metodología de ciencia de datos.			
14. Describir los componentes del modelo en la metodología de ciencia de datos.			
15. Descripción de metodologías ágiles.			
16. Descripción de ETL y procedimientos para la construcción.			
17. Descripción de DW.			
18. Descripción de repositorios masivos.			
19. Descripción de elementos para el BI.			
20. Descripción de perfiles técnicos.			

21. Descripción de perfiles para realizar BI.			
22. Descripción de perfil para el científico de datos.			
Escoger las herramientas de software que se deben usar en cada componente del modelo de ciencia de datos.			
23. Investigación, selección e implementación de las herramientas para la ETL.			
24. Investigación, selección e implementación de las herramientas para construcción del DW			
25. Investigación, selección e implementación de las herramientas para BI.			
26. Investigación, selección e implementación de las herramientas para la aplicación de minería de datos.			
Elaborar una guía de la implementación de los componentes del modelo de ciencia de datos			
27. Desarrollo de la etapa número uno de la metodología repositorios de datos.			
28. Desarrollo de la etapa número uno de la metodología tipos de datos estructurados y no estructurados.			
29. Desarrollo de la etapa número uno de la metodología repositorios de datos pequeños.			
30. Desarrollo de la etapa número uno de la metodología repositorios de datos masivos.			
31. Desarrollo de la etapa número uno de la metodología ágil en el proceso de ETL.			
32. Desarrollo de la etapa número uno de la metodología ágil en el proceso de inteligencia de negocios.			

33. Desarrollo de la etapa número uno de la metodología ágil en el proceso de ciencia de los datos.			
34. Desarrollo de la etapa número dos de la metodología - Sinergia de metodología BIM y SCRUM en inteligencia de negocios.			
35. Desarrollo de la etapa número dos de la metodología - Sinergia de metodología CRISP-DM y SCRUM en la ciencia de los datos.			
36. Desarrollo de la etapa número dos de la metodología - Sinergia de metodología KDD y SCRUM en la ciencia de los datos.			
37. Desarrollo de la etapa número dos de la metodología - Sinergia de metodología KDD y SCRUM en la ciencia de los datos.			
Analizar cuáles son los mayores aportes de la metodología y algunas deficiencias encontradas en el modelo de ciencia de datos.			
38. Elaboración de un informe con las ventajas y desventajas encontradas de la implementación de los repositorios.			
39. Elaboración de un informe con las ventajas y desventajas encontradas de la implementación de los repositorios de datos pequeños.			
40. Elaboración de un informe con las ventajas y desventajas encontradas de la implementación de los repositorios de datos masivos.			
41. Elaboración de un informe con las ventajas y desventajas encontradas de la implementación de KDD en el departamento de ciencia de los			

datos.			
42. Elaboración de un informe con las ventajas y desventajas encontradas de la implementación de BIM en el departamento de ciencia de los datos.			
43. Elaboración de un informe con las ventajas y desventajas encontradas de la implementación de CRISP-DM en el departamento de ciencia de los datos.			
44. Elaboración de un informe con las ventajas y desventajas encontradas de la implementación de SCRUM en el departamento de ciencia de los datos.			
45. Elaboración de un informe con las ventajas y desventajas encontradas de la implementación de la inteligencia de negocios en el departamento de ciencia de los datos.			
46. Elaboración de un informe con las ventajas y desventajas encontradas de la implementación de la ciencia de los datos en el departamento de ciencia de los datos.			
47. Elaboración de un informe con las ventajas y desventajas encontradas de la implementación de KDD en el departamento de ciencia de los datos.			
48. Recopilación final de los hallazgos en los componentes, herramientas y modelos, para el servicio de ciencia de datos.			
49. Recomendaciones y conclusiones finales de los hallazgos en los componentes,			

herramientas y modelos, para el servicio de ciencia de datos.			

5.3Fase 3 Desarrollo

5.3.1 Preparación Técnica

5.3.1.1 Selección Herramientas

Esta sección tiene por objetivo recopilar algunas de las herramientas de Inteligencia de tipo open source y propietarias que se encuentran en el mercado con la finalidad de mostrar que el mercado de herramientas es lo suficientemente amplio y variado.

En este proyecto se usará la plataforma de Microsoft BI para realizar el proyecto debido a que la organización ya contaba con dicha herramienta, pero para otros casos se puede utilizar cualquiera de las plataformas descritas posteriormente.

5.3.1.1.1 Jaspersoft

Es una plataforma de Business Intelligence que está destinada para generar soluciones en empresas pequeñas y medianas. Esta plataforma cuenta con herramientas que permiten hacer el desarrollo completo de una solución de BI, partiendo desde la extracción de la información de las fuentes de datos para ser almacenadas en un repositorio de datos, para que posteriormente estos sean explotados con herramientas de análisis para que finalmente sean visualizados por los usuarios de negocio en diferentes niveles. Para lograr esto la plataforma cuenta con las siguientes herramientas:

5.3.1.1.1.1Jaspersoft ETL.

Permite desarrollar, administrar y documentar los procesos de ETL en una organización. Dichos procesos servirán para poblar el repositorio de datos.

5.3.1.1.1.2 Jaspersoft OLAP

Esta herramienta permite hacer análisis multidimensional sobre los datos. Lo que permite obtener una mejora en la respuesta de las consultas de dicha información debido a las agregaciones que son almacenadas en la metadata del servidor OLAP.

5.3.1.1.1.3 JaspersoftReport Server

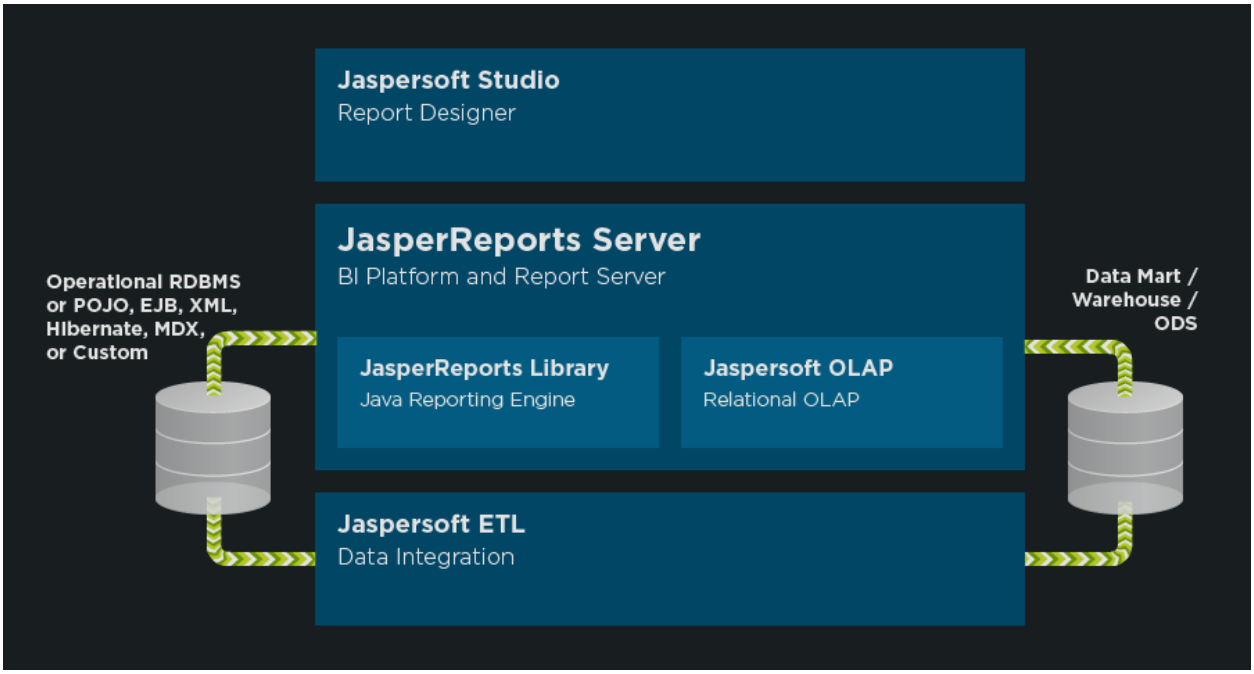
Es el servidor que permite la administración de la publicación de los reportes generados para su visualización en Web.

5.3.1.1.1.4 Jaspersoft Studio

Se utiliza para diseñar y ejecutar plantillas de informes; Crear consultas de informes; Escribir expresiones complejas; componentes visuales de diseño como 50 tipos de gráficos, mapas, tablas, tablas de referencias cruzadas y visualizaciones personalizadas; y mucho más. Integra TIBCO JasperReports Server para crear potentes flujos de trabajo de publicación de informes.

Plataforma

Figura 30 Plataforma de Jaspersoft



Fuente: Jaspersoft

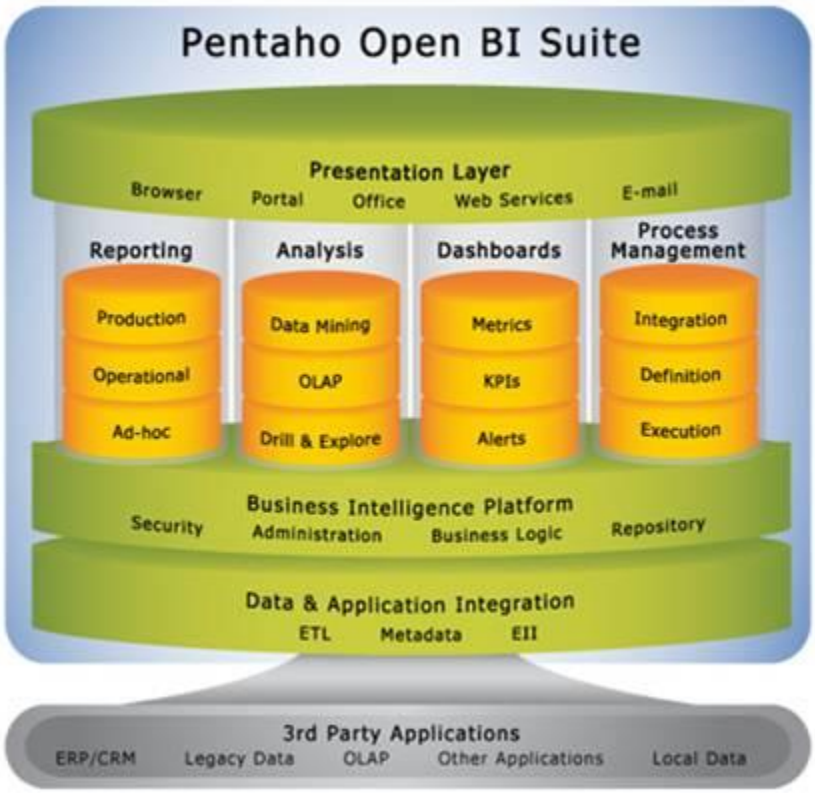
Es una suite de Business Intelligence que ofrece servicios críticos como calendarización seguridad, integración, navegación por contenido y provee las siguientes funcionalidades:

- ✓ Big Data.

- ✓ Data Integration.
- ✓ Reporting.
- ✓ Analytics.
- ✓ Dashboard.
- ✓ Data Mining.

5.3.1.1.1.5 Plataforma de pentaho

Figura 31 Plataforma de Pentaho



Fuente: Pentaho

5.3.1.1.2 SpagoBI

Es una plataforma de Business Intelligence única totalmente código abierto. Cubre todas las áreas de análisis de proyectos de Business Intelligence, con temas y motores innovadores. SpagoBI ofrece una amplia gama de herramientas de análisis, de la siguiente manera:

5.3.1.1.2.1 Plataforma y Arquitectura

Figura 32 Plataforma de SpagoBI

Fuente: spago BI

Esta es la gama de herramientas que posee Spago BI:

- ✓ Reporting.
- ✓ OLAP.
- ✓ Chart.
- ✓ Dashboard.
- ✓ KPI.

- ✓ Cockpits.
- ✓ GEO/GIS.
- ✓ Data Mining.
- ✓ QuerybyExample.
- ✓ Smart Filter.
- ✓ Accesibility Reporting.
- ✓ RT Console.
- ✓ Dossier.
- ✓ ETL.
- ✓ Office.

La Figura siguiente muestra la arquitectura en que la que funciona la plataforma de Spago BI.

Figura 33 Arquitectura SpagoBI

Fuente: spagoBI

Posterior se muestra en detalle una explicación de cada componente de la arquitectura:

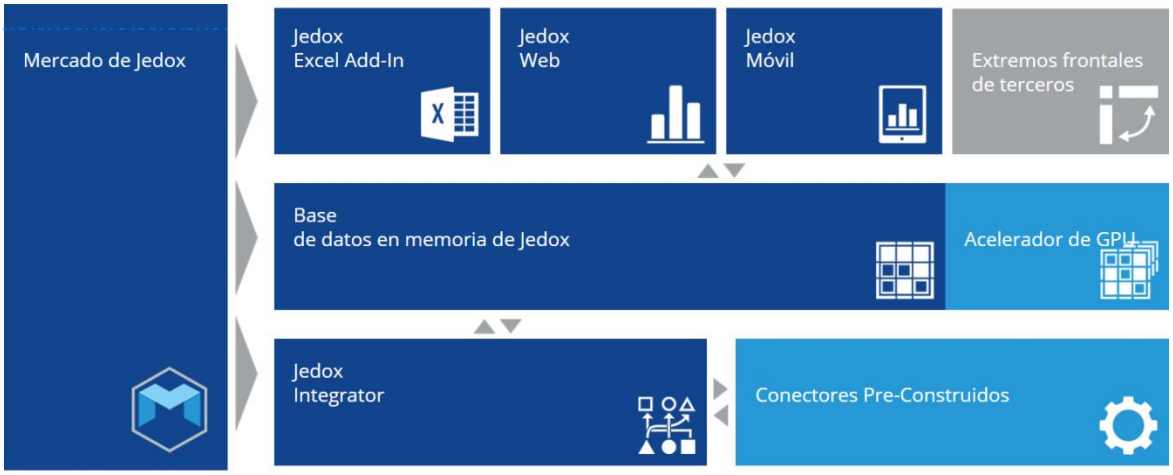
- ✓ **SpagoBI servidor**, el núcleo de la suite incluyendo las herramientas analíticas y características.
- ✓ **SpagoBI Studio**, el entorno de desarrollo integrado.
- ✓ **SpagoBI Meta,** el entorno de metadatos.
- ✓ **SpagoBI SDK**, la capa de integración que permite utilizar SpagoBI con herramientas externas.
- ✓ **Aplicaciones SpagoBI**, una colección de modelos analíticos verticales que se desarrollan utilizando SpagoBI.

5.3.1.1.3 Jedox

El software Jedox Enterprise Performance Management es una plataforma unificada y todos los componentes son desarrollados por Jedox. La plataforma forma parte de la poderosa base de datos en memoria de Jedox con interfaces de usuario intuitivas para la planificación, análisis e informes y componentes para la integración y preparación de datos de autoservicio. La innovadora tecnología de GPU aumenta el rendimiento para dar soporte a organizaciones complejas y con muchos usuarios.

5.3.1.1.3.1 Plataforma

Figura 34 Plataforma Jedox



5.3.1.1.4 Weka

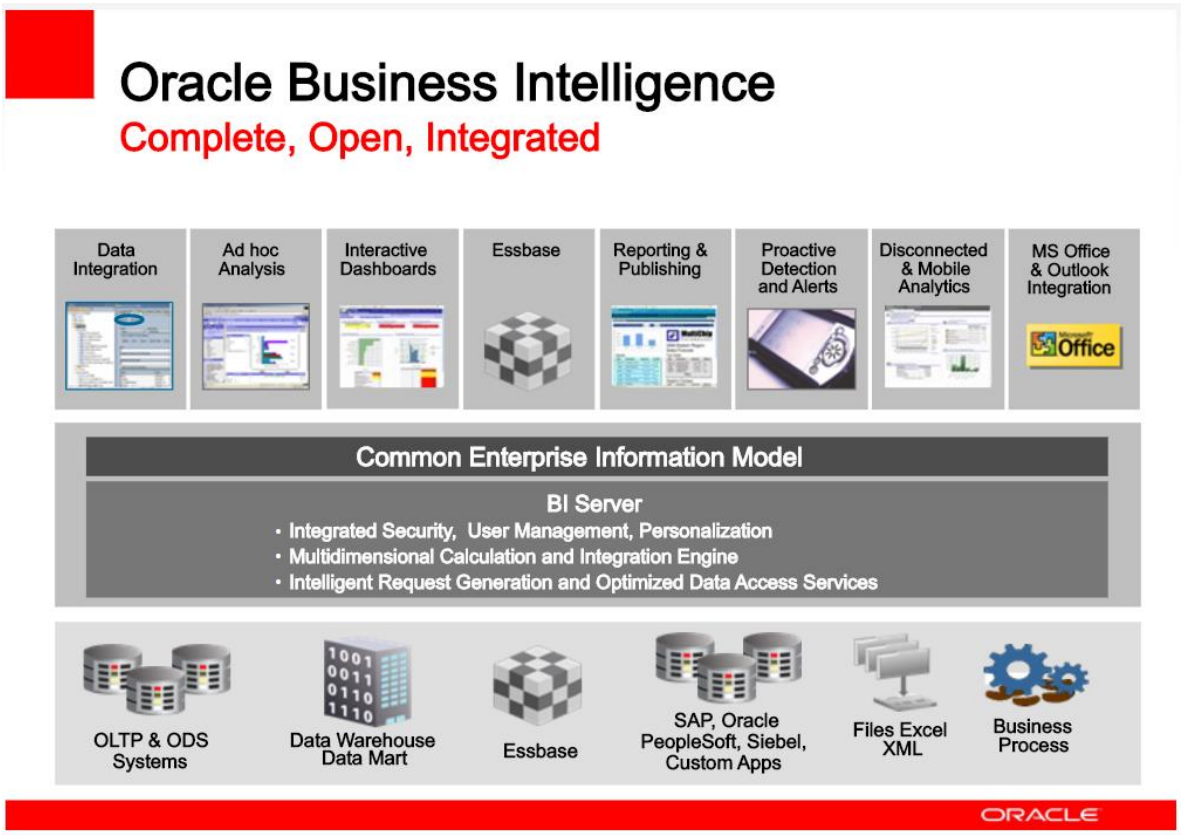
Es una colección de algoritmos de aprendizaje automático para tareas de minería de datos. Los algoritmos pueden ser aplicados directamente a un conjunto de datos o llamados desde su propio código Java. Weka contiene herramientas para pre-procesamiento de datos, clasificación, regresión, agrupamiento, reglas de asociación y visualización. También es adecuado para desarrollar nuevos esquemas de aprendizaje de máquinas.

5.3.1.1.5 Oracle BI

Es la plataforma más completa para la inteligencia de negocios (BI) disponible en la actualidad, cubriendo un amplio espectro de necesidades de inteligencia de negocios, incluidos los tableros interactivos, el análisis ad-hoc, alertas e inteligencia proactivas, publicación e informes avanzados, análisis predictivo en tiempo real, análisis de tecnología móvil, y mucho más.

5.3.1.1.5.1 Plataforma

Figura 35 Plataforma Oracle



Fuente: OracleBI

5.3.1.1.6 MicroStrategy

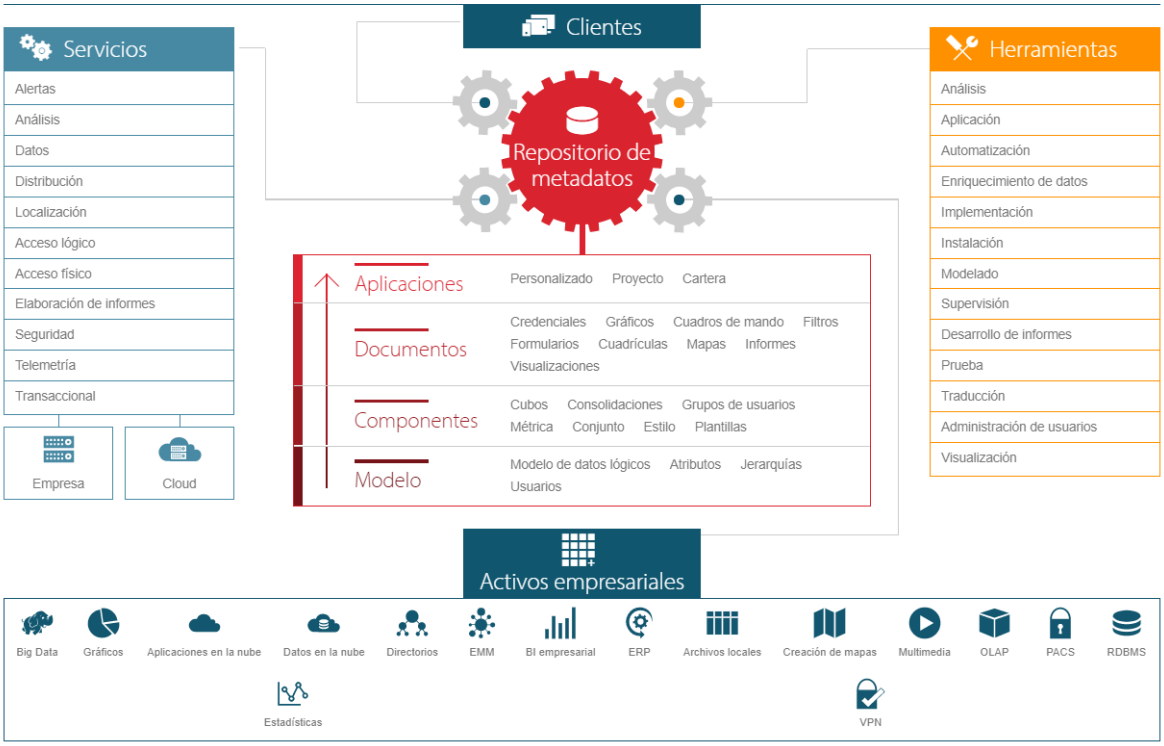
Plataforma que de productos que ayudan a las empresas a soportar las necesidades cambiantes del negocio para grupos de trabajo y aplicaciones departamentales de BI, facilita la migración de aplicaciones BI departamentales a una arquitectura empresarial. Esta combina la inteligencia de negocio tradicional con el

análisis de última generación, la movilidad y la tecnología en la nube, permitiendo a las organizaciones desarrollar e implementar aplicaciones de transformación que maximizan el valor de la información y aceleran los negocios. A diferencia de otros proveedores, toda esta tecnología se ha desarrollado de manera orgánica, por lo que cada pieza del ecosistema MicroStrategy encaja a la perfección. Su principal virtud es la capacidad de ser la única plataforma de movilidad y analítica empresarial del mercado. Algo de más interesante de esto es que posee una plataforma unificada para el análisis de toda su información.

5.3.1.1.6.1 Plataforma

Esta es la plataforma de cómo se integran todas las Áreas.

Figura 36 Plataforma MicroStrategy



Fuente: Microstrategy

5.3.1.1.7 IBM Cognos

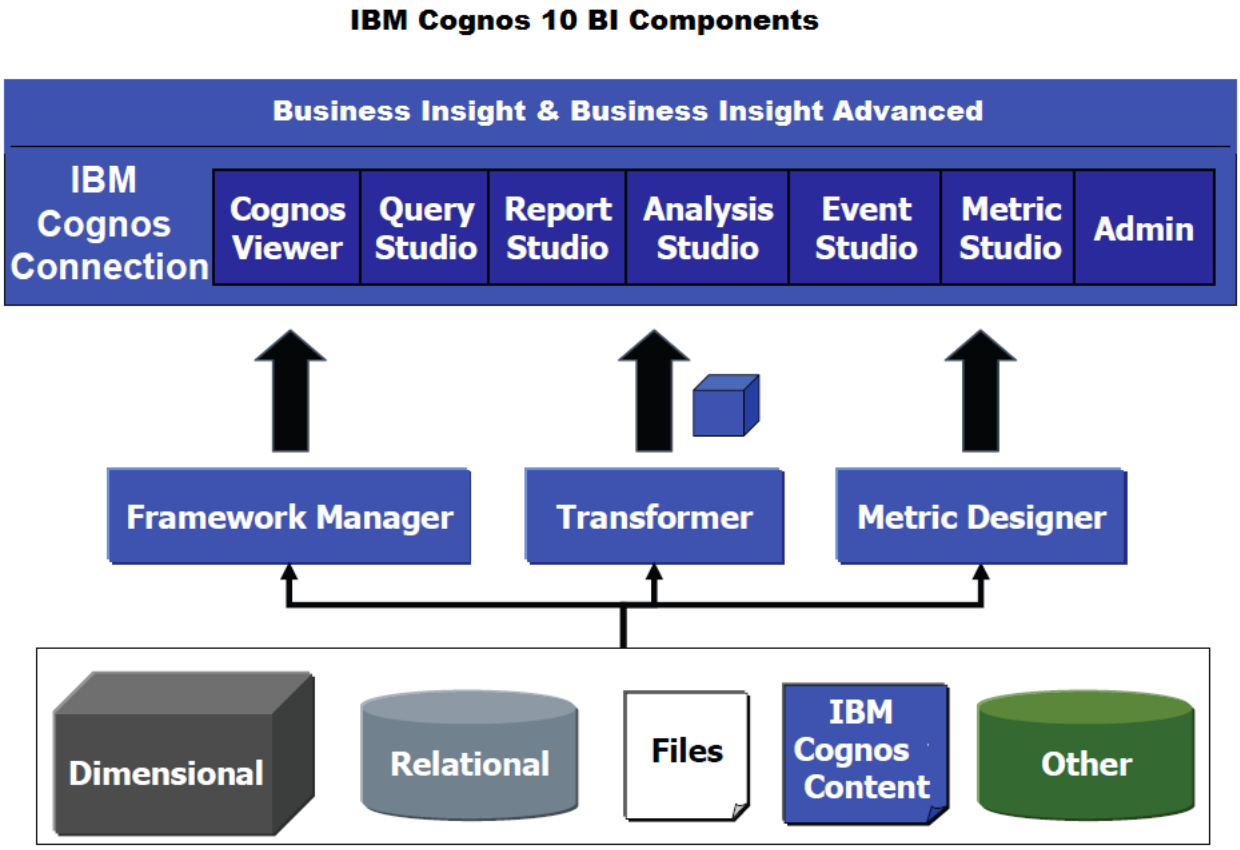
Es una plataforma que permite la entrega completa consistente y a tiempo de la información para todos los usuarios dentro de una infraestructura altamente estable. Además, satisface las necesidades de inteligencia de negocios como reportes, dashboards, cuadros de mandos, análisis y planeación reduciendo la complejidad del ambiente de Business Intelligence.

Esta conformada por:

- IBM Cognos BI.
- IBM Financial Performance and Strategy Management
    - IBM Cognos TM1.
    - IBM Cognos Planning.
    - IBM Cognos Controller.
    - IBM Cognos Business ViewPoint.
- IBM Analytics Applications.
- IBM Advanced Analytics (SPSS).
- IBM Cognos Express.

5.3.1.1.7.1 Plataforma

Figura 37 Plataforma IBM Cognos



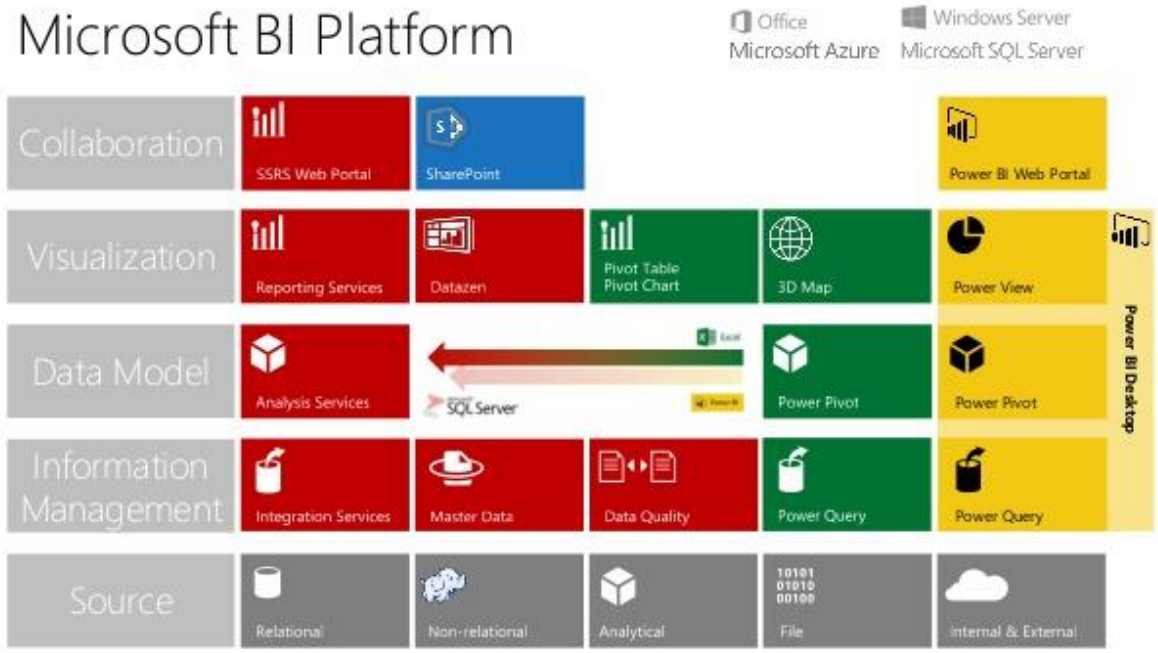
Fuente: COGNOS

5.3.1.1.8 Microsoft

Microsoft BI Platform, es una plataforma completa para el análisis de datos e inteligencia de negocios, está basada en Microsoft SQL Server y proporciona capacidades de reportes, análisis e integración de datos sin precedente.

5.3.1.1.8.1 Plataforma

Figura 38 Plataforma Microsoft BI



Fuente: Microsoft

5.3.1.1.9 QlikView

Provee un motor de ETL y no requiere de datos pre-agregados. Permite el análisis a cualquier nivel de detalle. Permite una conexión automática de las tablas para lo cual es necesario que el modelo de datos contenga los mismos nombres en todas las tablas cuyos conceptos se deban ligar (Grabova, Darmont, Chauchat, & Zolotaryova, 2010).

Está enfocado en soluciones de inteligencia de negocios de autoservicio lo que quiere decir que buscan dar autonomía al usuario final para que pueda generar los reportes que requiera en el momento que así lo decida.

5.3.1.1.10 R

Es un lenguaje de programación especialmente indicado para el análisis estadístico. R fue inicialmente diseñado por Robert Gentleman y Ross Ihaka, miembros del Dpto. de Estadística de la Universidad de Auckland, en Nueva Zelanda. Sin embargo, una de las grandes ventajas de R es que actualmente es, en realidad, fruto del esfuerzo de miles de personas en todo el mundo que colaboran en su desarrollo.

R se considera la versión libre de otro programa propietario, llamado S o S-Plus, desarrollado por los Laboratorios Bell. Aunque las diferencias entre R y S son importantes, la mayoría del código escrito para S funciona en R sin modificaciones.

El código de R está disponible como software libre bajo las condiciones de la licencia GNUGPL. La página principal desde la que se puede acceder tanto a los archivos necesarios para su instalación como al resto de recursos del proyecto R es http://www.r-project.org.

5.3.1.1.10.1 Las funcionalidades de R están divididas en un buen número de "paquetes":

- ✓ El R Base que contiene, entre otros, los paquetes básicos requeridos para hacer funcionar el programa (funciones fundamentales).
- ✓ Otros paquetes que incluyen utils, stats, datasets, grid, tools, etc., contenidos en el R básico.
- ✓ Otros paquetes recomendados como class, cluster, rpartial, etc.
- ✓ Actualmente existen casi 5000 paquetes en CRAN desarrollados por contribuidores de todo el mundo.

5.3.1.2 Instalación

Utilizando el diseño de arquitectura técnica como marco, es necesario seleccionar los componentes específicos de la arquitectura como la plataforma de hardware, el motor de base de datos, la herramienta de ETL o el desarrollo pertinente, herramientas de acceso, entre otros. Para ello se analizó la plataforma con la que cuenta la PYMES GoLabs y además qué herramientas se tienen y cuáles hacen falta para el proceso de desarrollo del proyecto, se consulta a los líderes de la PYMES que si están dispuestos a invertir y cuál es el presupuesto con el que se cuenta, para así analizar el uso de herramientas de uso libre. Una vez evaluados y seleccionados los componentes determinados se procede con la instalación y prueba de los mismos en un ambiente integrado de Data Warehouse.

Se realizó la instalación de la Plataforma de Microsoft en bases de datos, ETL, Cubo y visualización y además se trabajó con software libre R para la minería de datos.

5.3.2 Construcción de Data Warehouse

Este punto de la metodología es en donde se genera el diccionario de datos y se describen las estructuras para entender el objetivo que se persigue con cada tabla y sus características principales.

5.3.2.1 Construcción de modelo de Data Warehouse

La fuente principal de la información con la que contamos es la base de datos del sistema TrackIt que utiliza Go-Labs para registrar las actividades diarias. Esta base de datos se encuentra en SQL Server, y para su análisis ha sido cargada en SQL Server 2014 express edition.

Para iniciar con el análisis de los datos fue proporcionado un respaldo de la base de datos, la cual fue restaurada en el laboratorio. Se visualiza el contenido de dicha base de datos, lo cual permite observar que existen muchos objetos que no son necesarios para cumplir con los objetivos planteados.

Luego se realizó una revisión de cada una de las tablas en búsqueda de datos que sean útiles para el logro de los objetivos planteados. Estos objetos están relacionados con lo siguiente:

- Cliente: persona que realiza compras de órdenes de track.
- Paquetes: son la cantidad de traqueos que posee una orden.
- Trackeo: es la trama o conjunto de bites con información de la localización de un paquete o mercancía.

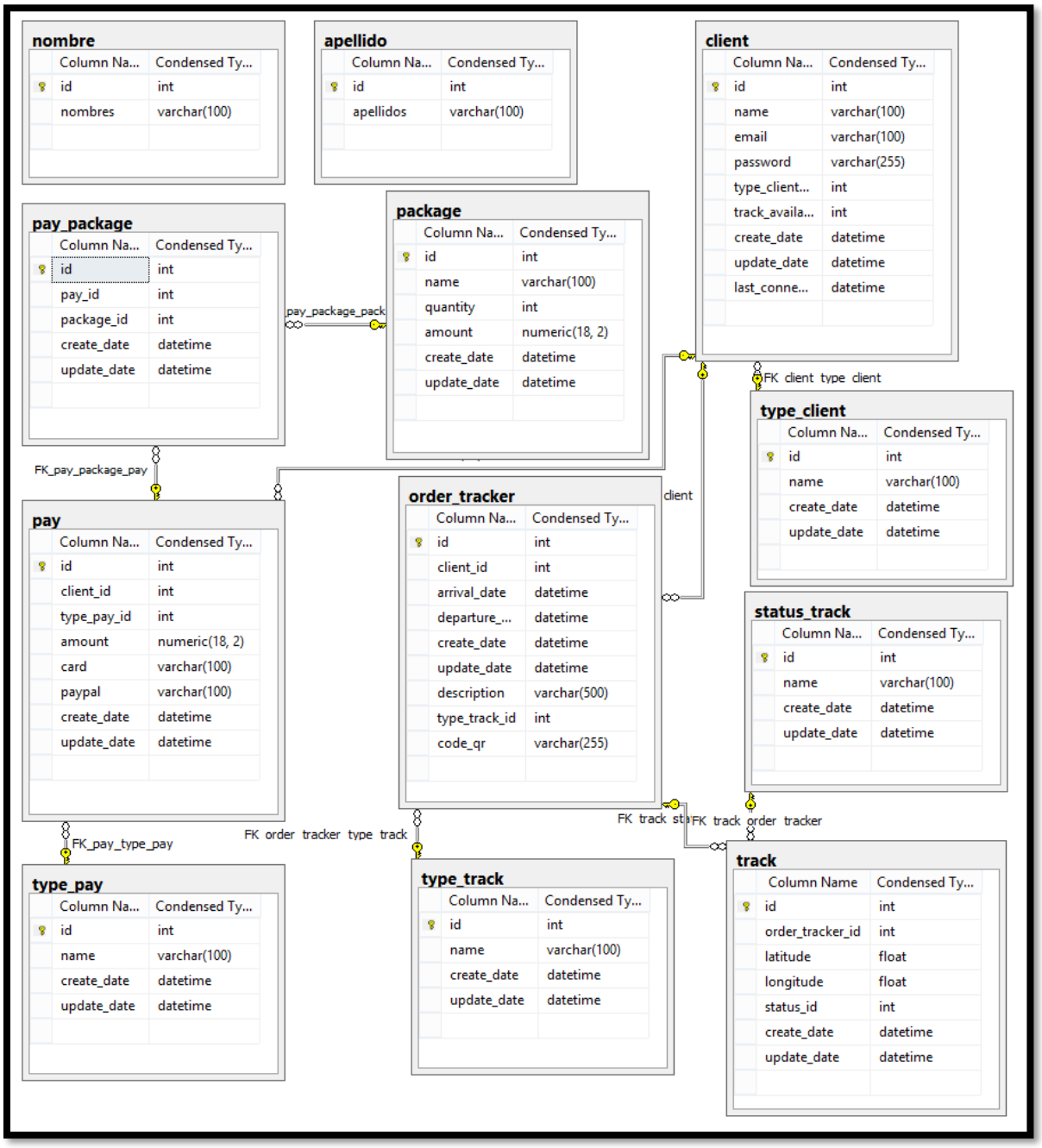
5.3.2.2 Base de datos transaccional

Esta base de datos será la principal fuente de datos para realizar el análisis de la información, esta contiene toda la información del día con día del sistema Track it. Los datos almacenados en esta base de datos son de mucha importancia para abstraer la información útil para la estrategia de negocio.

5.3.2.3 Diagrama de Entidades de la Base de Datos Transaccional

Diagrama de la base de datos con sus entidades y atributos.

Figura 39 Diagrama transaccional

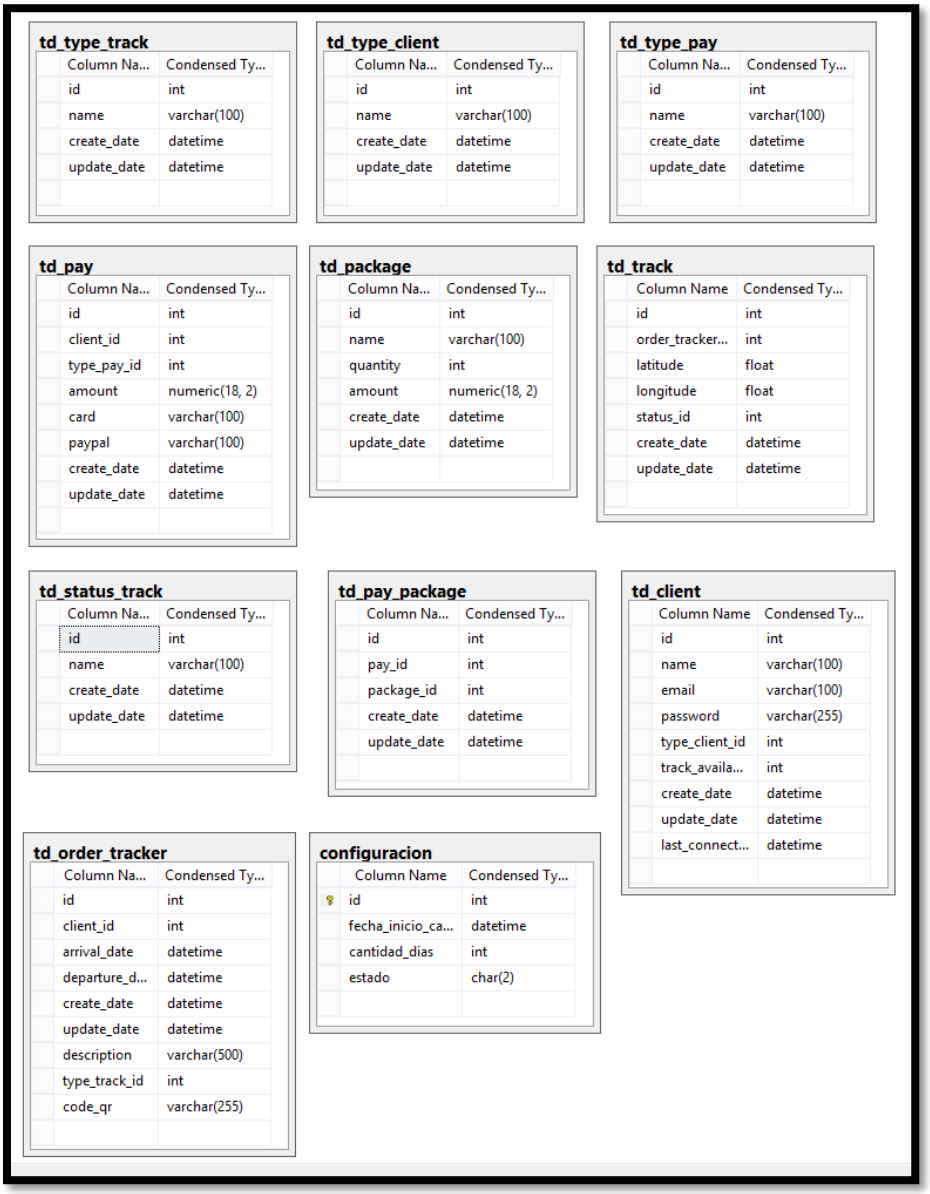


5.3.2.4 Base de datos de área intermedia o Staging Área (SA)

Se diseña una base de datos donde se cargará solo la información necesaria, cabe mencionar que esta base se le realizarán dos tipos de cargas de datos, la carga inicial que incluye los datos de un largo periodo y la carga diaria que se hará con datos del día anterior cuando se necesite pasar al Data Warehouse.

5.3.2.5 Diagrama de Base de datos Intermedia (SA)

Figura 40 Diagrama base de datos Intermedia

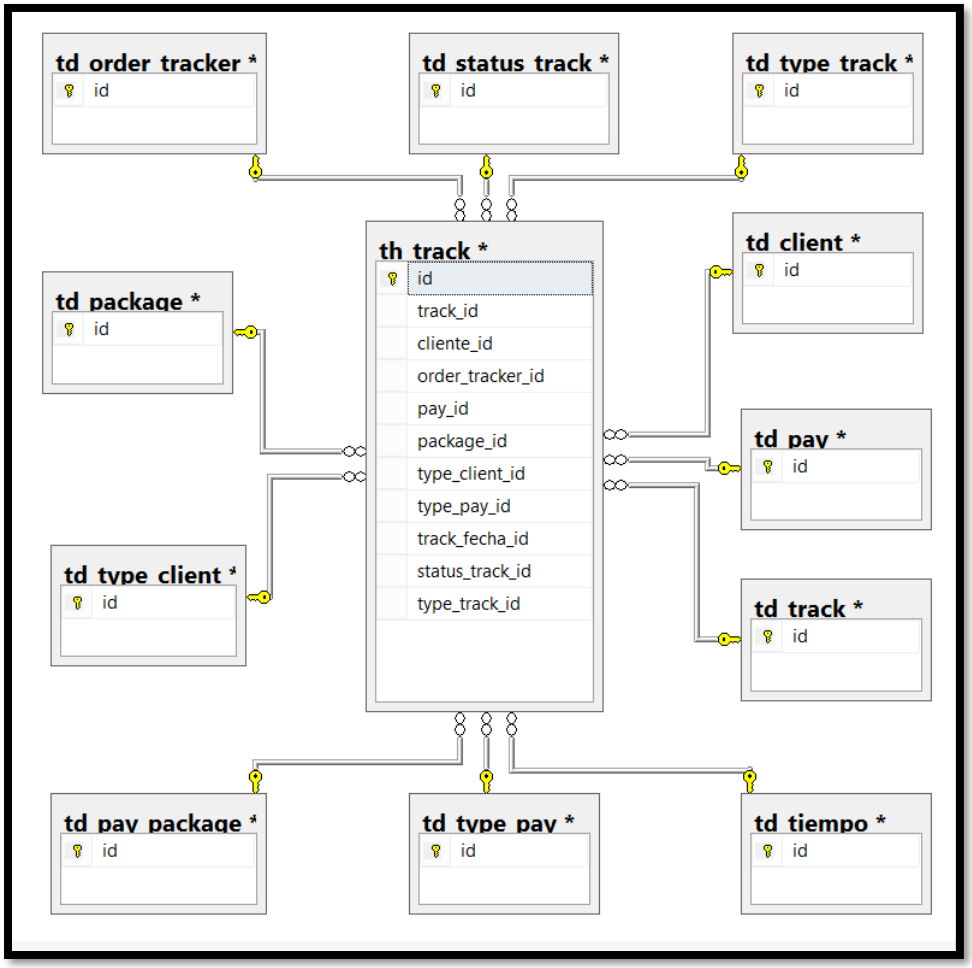


5.3.2.6 Base de datos Data Warehouse

De acuerdo con las necesidades del negocio se realiza un modelo esquema de estrella con la finalidad de que las consultas sean lo más optimizadas posibles. Este modelado se realizó tomando en cuenta las necesidades de información que se investigaron del negocio.

En primera instancia se tiene en la parte central de la tabla de hechos **th_track** en la cual están las principales métricas del trackeo de paquetes, así como las ventas. Estas métricas se pueden analizar por una jerarquía de tiempo, por cliente, por paquetes, entre otras.

Figura 41 Diagrama depósito de datos



5.3.3 Datos

En esta etapa se debe de centrar en los datos que se requieren para realizar el descubrimiento de conocimiento por medio de la ciencia de los datos.

5.3.3.1 Análisis de datos

Dentro de la exploración de los datos se debe analizar las ventas y traqueo de paquetes debido a que se cuenta con los registros que son tomados de la base transaccional, en donde queda registrada todas las operaciones de ventas.

Los datos relacionados con las ventas, el cliente relacionado y la venta total. Así mismo estos datos se pueden relacionar con los paquetes que son transportados para determinar si el servicio fue proporcionado de forma correcta, así mismo se puede identificar a los clientes que nos han generado un mayor número de ventas, así como sus hábitos de consumo para poder determinar ofertas especiales para estos clientes

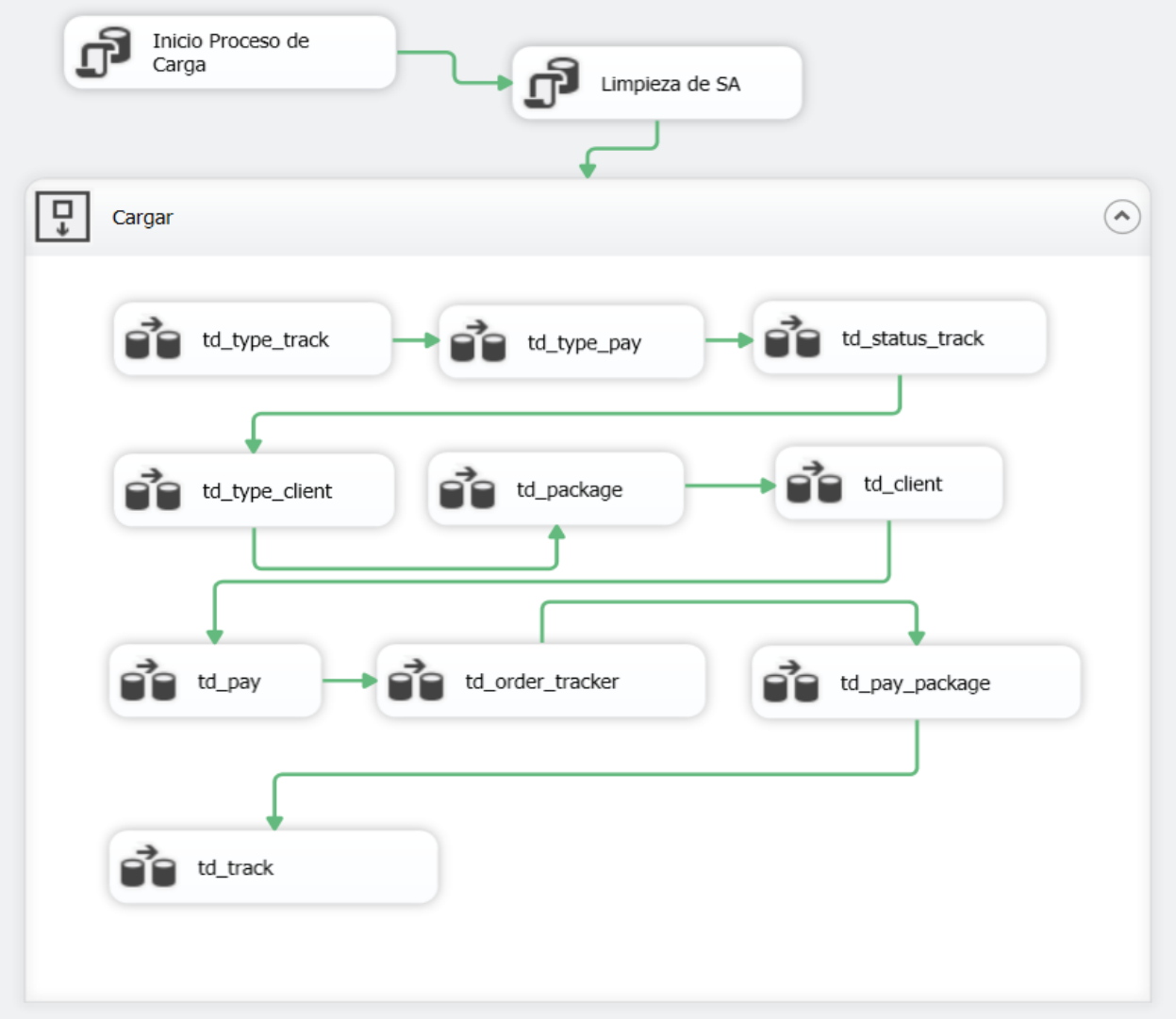
5.3.3.2 Desarrollo procesos de ETL

Los ETL se usan para extraer, transformar y cargar los datos en las bases de datos, en el proceso de la ciencia de los datos los ETL se usa para pasar los datos a las diferentes etapas del proceso como son las bases de datos Stagin Area y Data Warehouse.

5.3.3.2.1 ETL para carga del Stagin Área SA

El ETL hace la carga inicial de Staging Área este solo sirve para la primera carga, los datos de carga son tomados de la base de datos transaccional y se insertan en base de datos Staging Área, luego este se modifica después de la primera carga y se le agrega el filtro que corresponde a los de la fecha del día anterior para ser configurado que se ejecute diariamente.

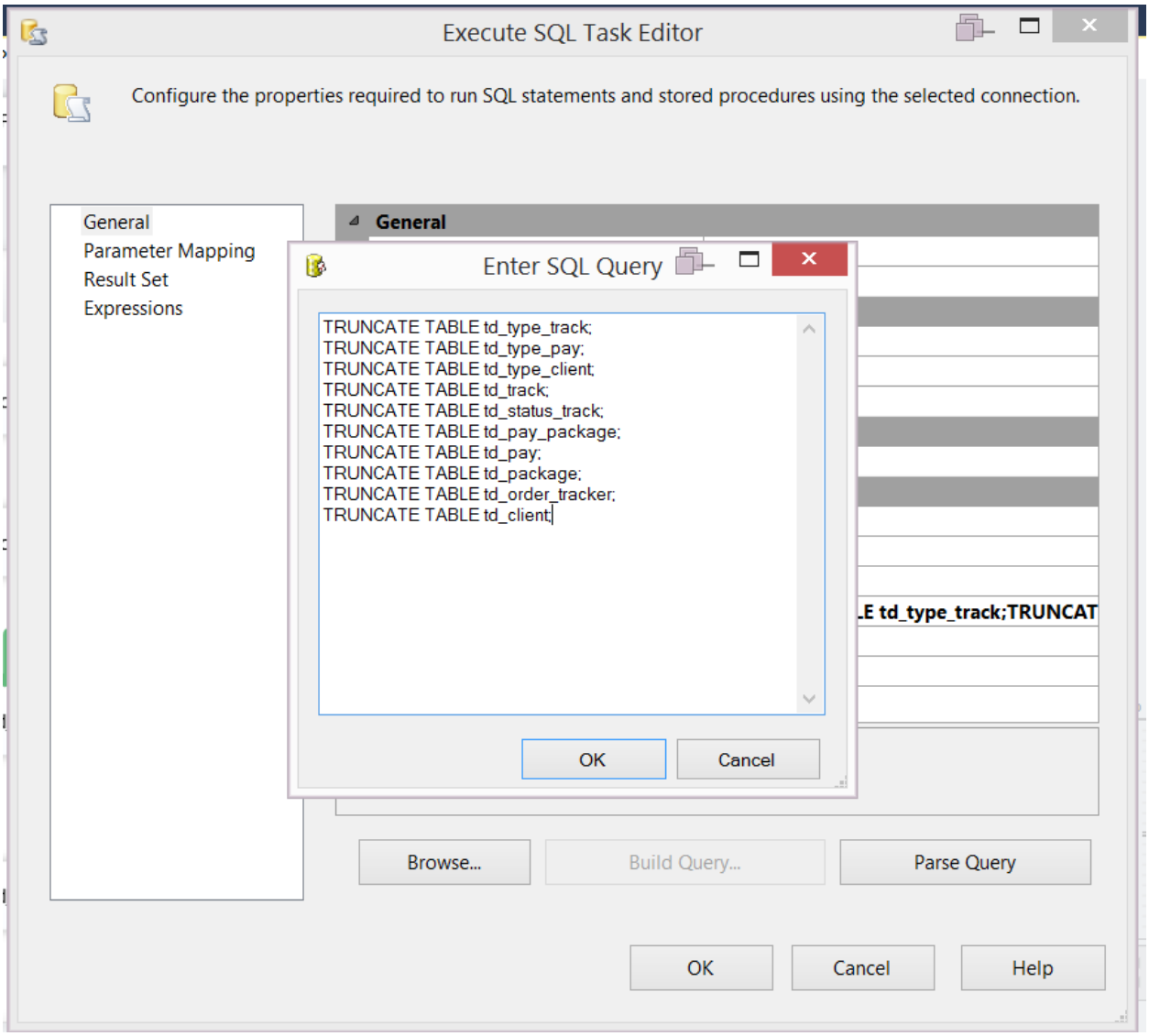
Figura 42 Diagrama ETL para Stagin Area



5.3.3.2.1.1 Borrado inicial

Se realiza el borrado de las tablas del Staging Área cada vez que se va a realizar una carga.

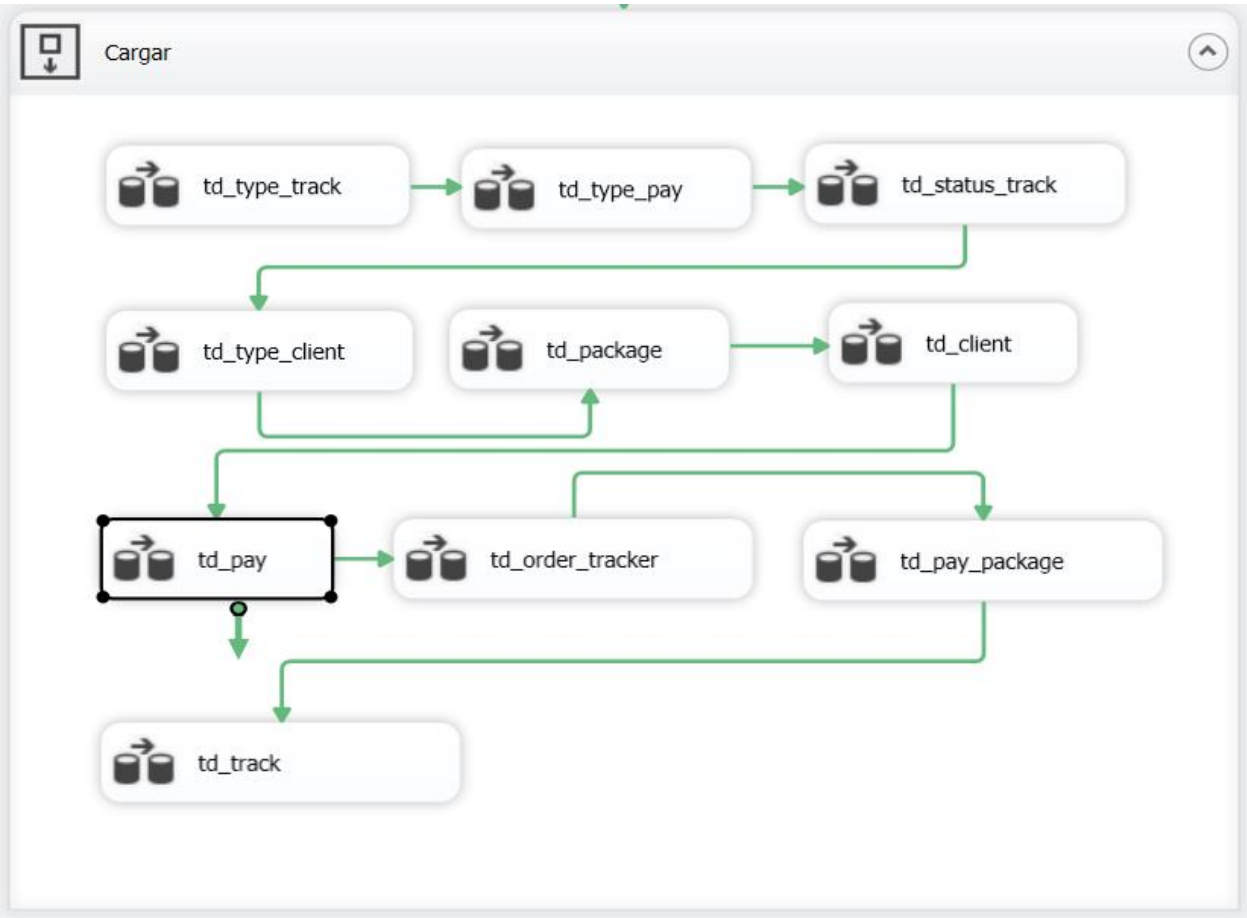
Figura 43 Tarea limpieza de datos



5.3.3.2.1.2 Carga de las tablas

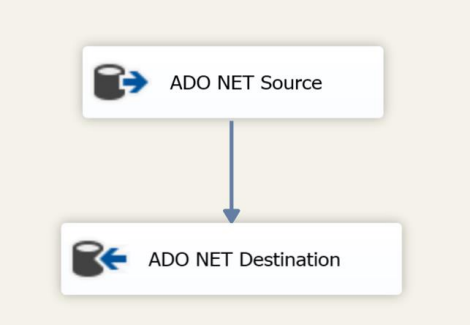
Se hace una obtención de los datos de la base de datos transaccional y se carga en la bases de datos Staging Área, esta carga los datos se debe realizar por cada una de las tablas usando esta estructura.

Figura 44 Carga de datos de SA



Para cada una de los dataflow se debe realizar la estructura básica de crear un ado net source para obtener los datos de la base transaccional y un ado net destination para insertar los datos en el SA.

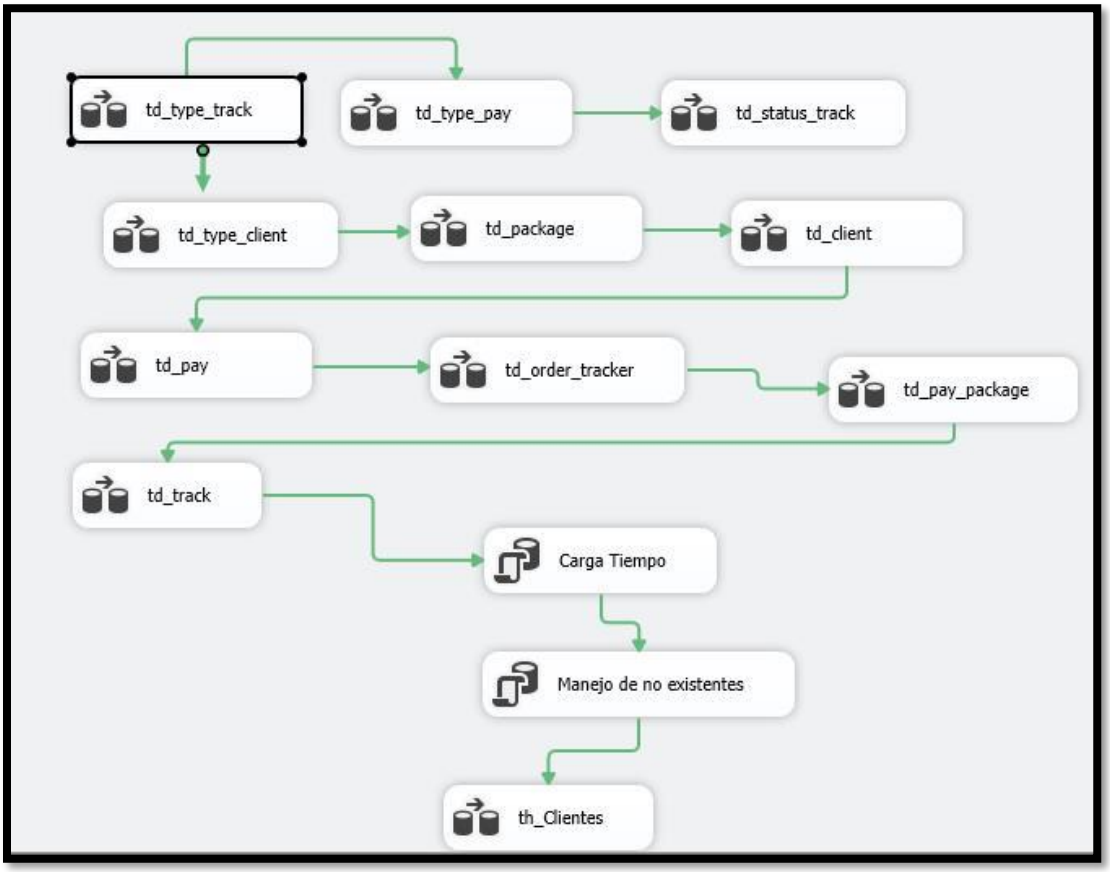
Figura 45 Componente dentro de cada carga



5.3.3.2.2 ETL para la carga del Data Warehouse

El ETL es el encargado de cargar la información al Data Warehouse, tiene transformaciones y cálculos para que se inserten. Este obtiene la información de la base de datos Staging Área y la carga al Data Warehouse.

Figura 46 ETL para carga de depósito

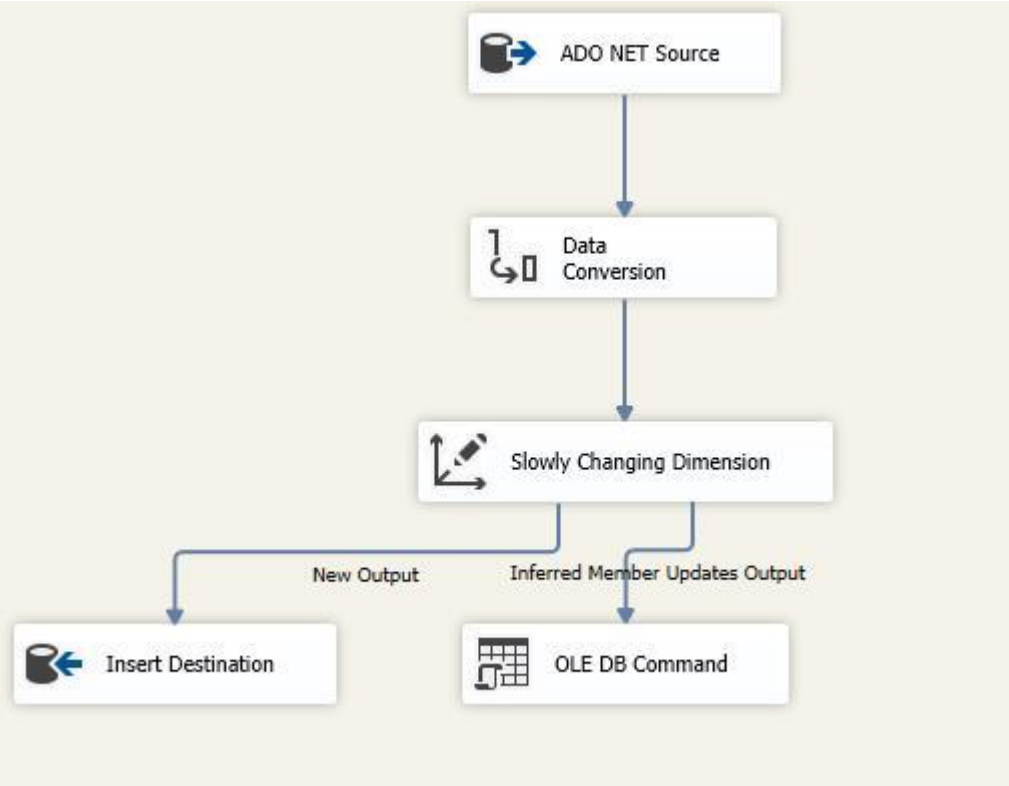


Fuente: Diseño propio

5.3.3.2.2.1 Carga de dimensiones

ETL que hace la carga de cada una de las dimensiones. El proceso que hace es que obtiene los datos del Staging Área(SA), luego hace una conversión de datos y se procede a realizar una validación en donde obtiene los datos y se asignan a las columnas para cargarlos al Data Warehouse.

Figura 47 Composición de componentes para carga de dimensión

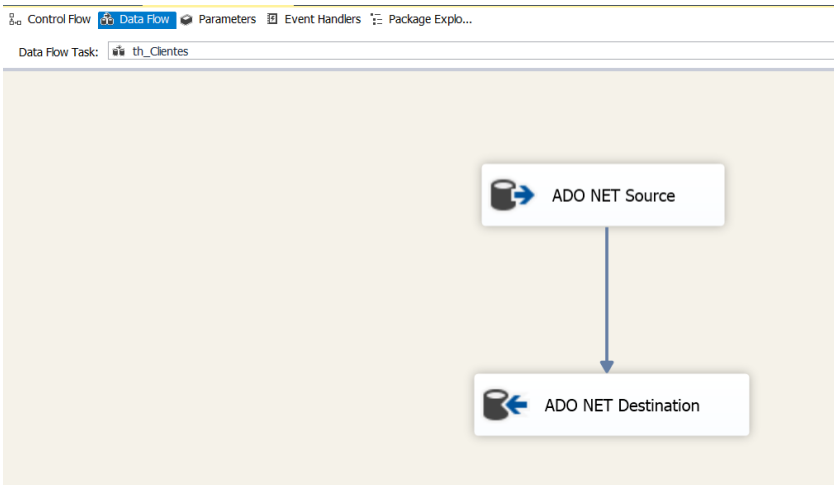


Fuente: Diseño propio.

5.3.3.2.2.2 Carga de tabla de hechos

Después de haber realizado la carga de las dimensiones es necesario hacer la inserción de los datos en la tabla de hechos en este caso se realiza con un ADO net source que permite obtener todos los datos de las dimensiones e ir a insertarlos con el ADO net destination en la tabla de hechos.

Figura 48 Componentes para carga de tabla de hechos



Fuente: Diseño propio.

5.3.3.2.3 Pruebas

En esta etapa se realiza las pruebas de carga de los datos, se verifica la integridad y si las conversiones de datos en casos donde son requeridas se están realizando de forma correcta.

Este proceso se debe realizar en un ambiente de pruebas que sea muy similar al de producción para así si hay algún inconveniente para poder detectarlo.

Para la carga de datos, se verifica cada vez que se ejecuta el ETL que se limpien todos los datos de las tablas y se toma el periodo de carga determinado, ya sea la carga de histórico o si es una carga periódica, esto con el fin de poder determinar qué tiempo se esperará en el proceso de producción, la prueba de la carga del histórico será un proceso que requiere de mucho tiempo por lo que lo ideal es que cuando se realice la prueba se haga con una cantidad de datos semejante a la de producción para así saber en qué momento es ideal realizar esta carga. También la carga que se

realiza periódicamente debe de probarse, asumiendo una cantidad de datos considerable y también cuál va a ser la periodicidad de las cargas.

5.3.3.3 Población histórica DWH y Calendarización ETL

Las bases de datos OLTP usualmente retienen datos en periodos de tiempo cortos, después son protegidos por los administradores de base de datos en almacenamientos secundarios fuera de línea (por ejemplo, en disco a nivel de backup). De igual forma es común que contengan sólo valores corrientes, por ejemplo, el actual balance de cuentas para clientes y no valores históricos, incluso esta puede no incluir el tiempo como un elemento clave, sólo el balance corriente de cuentas es almacenado, por lo tanto, no tiene sentido guardar el tiempo como parte de la clave de los datos.

Las bases de datos OLAP, a diferencia de los OLTP, almacenan datos históricos tanto como se considere necesario para el análisis del negocio, normalmente de dos a cinco años de datos históricos. Retienen valores para cada período (el atributo más atómico de la dimensión tiempo) en la base de datos. Almacenan una serie de “fotos” instantáneas de los datos operacionales y la frecuencia con la cual se define el nivel de detalle que se debe indicar en hoja correspondiente de la dimensión tiempo. Toda esta cantidad y tipo de historia ayuda a generar reportes de comparación de tendencias y períodos de tiempo.

Asimismo, las bases de datos orientadas al análisis siempre contienen el tiempo como dato clave, dado que una de las principales razones para la construcción del Data Warehouse es el almacenamiento de datos históricos y el análisis a lo largo del tiempo.

Debido a ello es necesario que se determine en la manipulación de los datos que se deben cargar al Data Warehouse, el periodo de datos de la base OLTP que se van a almacenar y cada cuanto tiempo se estará realizando la operación de carga por medio de los ETL, esto con el fin de incrementar y tener información, lo más actualizada posible, para que a la hora de generar reportes y tomar decisiones que se realice con certeza de lo que está sucediendo. Se recomienda que estos procesos de

carga de datos se realicen en horas de la madrugada para así no tener que influir en el rendimiento de la base OLTP mientras está en uso.

5.3.4 Minería

La minería de datos se aplicará según los problemas presentados y aplica en cada una de las etapas de desarrollo para obtener un modelo que ayude a la empresa Go-Labs a clasificar los clientes y órdenes.

5.3.4.1 Clasificación de clientes

5.3.4.1.1 Análisis del problema

La compañía Go-Labs quiere mejorar la relación con sus clientes por lo cual ha decidido, que los próximos clientes en ingresa y utilizar el sistema clasificarlos como clientes buenos y malos bajo con el fin de mejorar el departamento de ventas. Los criterios en los cuales se basa la compañía para decidir si un cliente es bueno: Si la última conexión fue hace menos de un mes, si el monto pagado supera los 30 dólares.

5.3.4.1.2 Entendimiento del problema

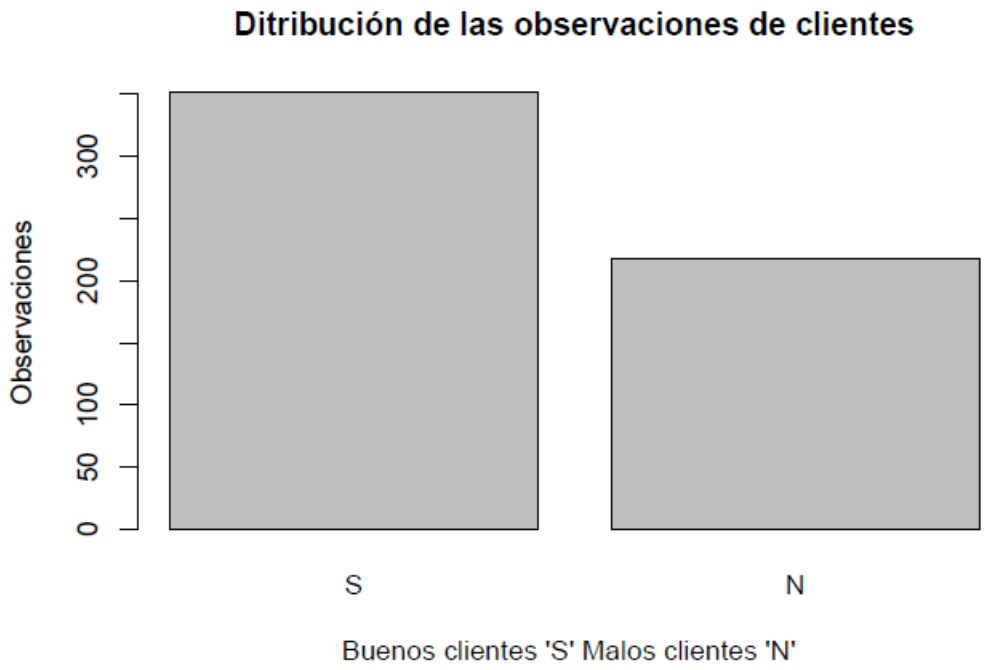
Como parte del estudio, se recolectaron los siguientes datos:

- id: número de muestra.
- type_client_id: identificador del cliente.
- track_available: cantidad de tracks disponibles.
- create_date: fecha creación del cliente.
- update_date: fecha actualización del cliente.
- last_connection: última actualización del cliente.
- clasificacion: clasificación del cliente

5.3.4.1.3 Exploración de datos

El este grafico se observar como hay una distribución de 218 observaciones de malos clientes y 352 buenos clientes.

Figura 49 Gráfico de Distribución de las Observaciones de los clientes



Fuente: Diseño propio.

5.3.4.1.4 Modelo de Minería de datos

Se construye un modelo el cual consideramos que es el más apto para satisfacer las necesidades propuestas en el problema.

```
#Modelo de bosque aleatorio
td_client.bosque=randomForest(clasificacion~ .,data=td_client.entrenamiento,ntree=11)
#Ver detalle del modelo
summary(td_client.bosque)
## Length Class Mode
## call 4 -none- call
## type 1 -none- character
## predicted 570 factor numeric
## err.rate 33 -none- numeric
## confusion 6 -none- numeric
## votes 1140 matrix numeric
## oob.times 570 -none- numeric
## classes 2 -none- character
## importance 6 -none- numeric
## importanceSD 0 -none- NULL
## localImportance 0 -none- NULL
## proximity 0 -none- NULL
## ntree 1 -none- numeric
## mtry 1 -none- numeric
```

```
## forest 14 -none- list
## y 570 factor numeric
## test 0 -none- NULL
## inbag 0 -none- NULL
## terms 3 terms call
```

5.3.4.1.5 Evaluación

Vamos a evaluar un modelo de árboles aleatorios para poder clasificar qué clientes estarán dentro del grupo de buenos y no tan buenos clientes.

```
#Creamos la predicción para el modelo del bosque aleatorio
predicciones.bosque=predict(td_client.bosque,newdata = td_client.prueba,type="response")
head(predicciones.bosque)
## 1 2 3 4 5 8
## N N S N S N
## Levels: N S
#Ver detalle de la predicción
summary(predicciones.bosque)
## N S
## 98 147
```

El nivel de exactitud del modelo de bosque aleatorio es del 96.33% lo cual es bastante alto y se recomienda su uso por la buena clasificación.

```
table(td_client.prueba$clasificacion,predicciones.bosque)
## predicciones.bosque
## N S
## N 92 2

## S 6 145
round((91+145 )/nrow(td_client.prueba),4)*100
## [1] 96.33
```

5.3.4.1.6 Resultados

En concusión se podría decir que el modelo generado es bastante confiables debido a que el modelo de bosque aleatorio dio una exactitud del 96%, es recomendable la utilización de estos dos, ya que las variables pueden predecir el tipo de cliente.

5.3.4.2 Clasificación de órdenes

5.3.4.2.1 Análisis del problema

La compañía Go-Labs quiere mejorar la experiencia en la compra de órdenes de tracking por lo cual ha decido ver la posibilidad de incrementar la cantidad de paquetes disponibles según los gustos de los clientes en los distintos grupos de compras

5.3.4.2.2 Entendimiento del problema

Como parte del estudio, se recolectaron los siguientes datos:

• id: número de muestra.

• type_client_id: identificador del cliente.

• td_package.quantity,

• td_package.amount AS precio,

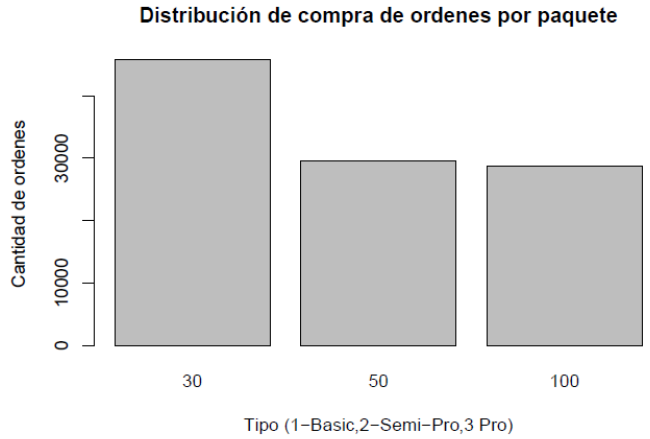
• td_pay.amount AS pago,

• th_track.cliente_id.

5.3.4.2.3 Exploración de datos

La distribución de compra de órdenes según los paquetes comprados por los clientes de track it.

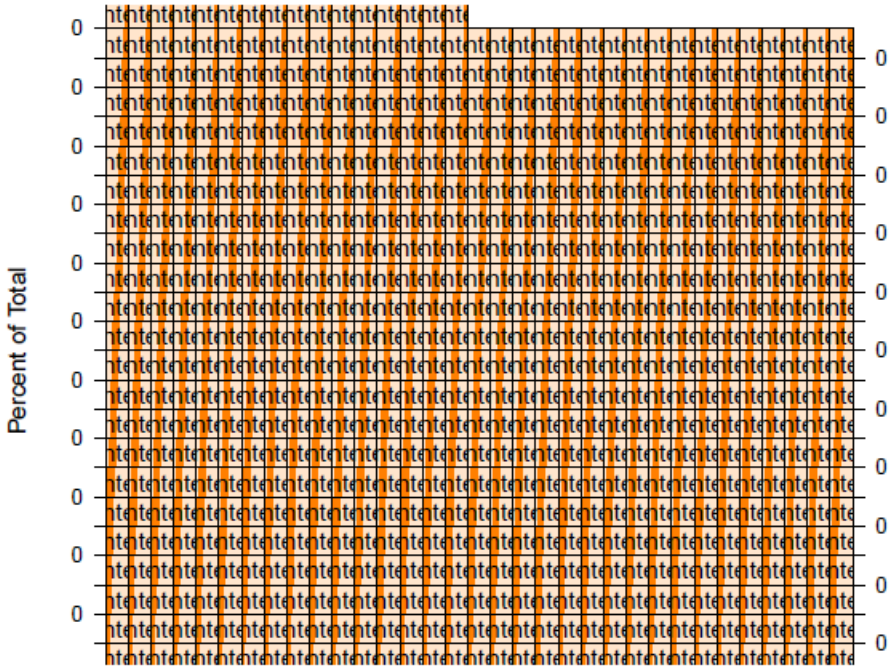
Figura 50 Gráfico de Distribución de compra de órdenes por paquete



Fuente: Diseño propio.

La distribución del top 10 de clientes con los montos más altos por los clientes, con esto se podrá analizar que cliente gastan más dinero en el uso de la herramienta.

Figura 51 Gráfico de agrupación



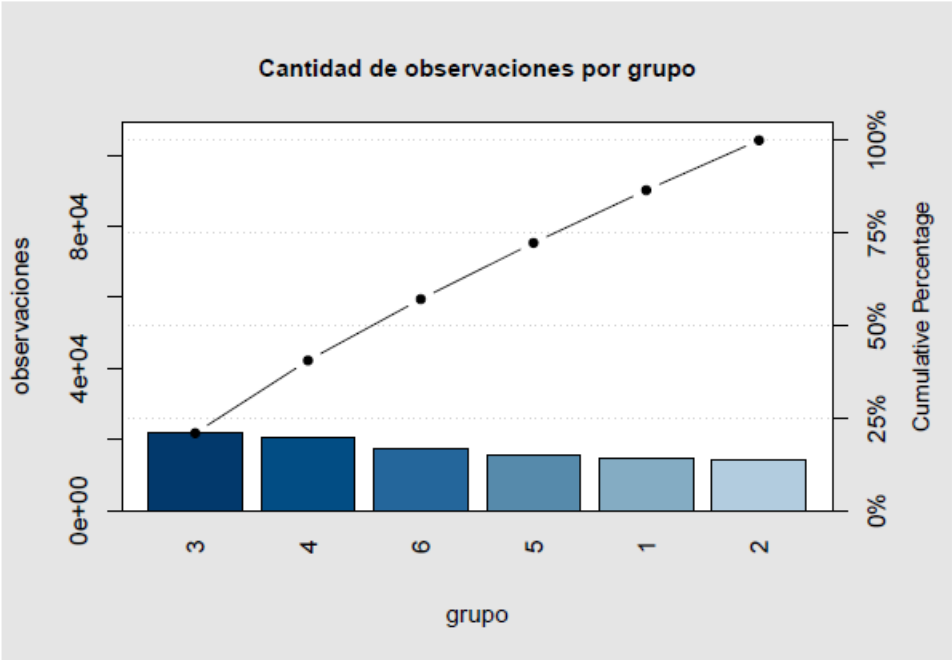
Fuente: Diseño propio.

5.3.4.2.4 Modelo de minería de datos

Se construye el modelo de agrupación para lograr el problema presentado.

```
modelo.agrupacion <- kmeans(th_track, centers = 6)
th_track$cluster <- modelo.agrupacion$cluster
th_track$cluster <- factor(th_track$cluster)
```

Figura 52 Gráfico de cantidad de observaciones por grupo



Fuente: Diseño propio.

```
## th_track$cluster: 1
## quantity precio pago cliente_id
## Min. : 30.00 Min. : 0.000 Min. : 0.000 Min. : 1.00
## 1st Qu.: 30.00 1st Qu.: 5.000 1st Qu.: 0.000 1st Qu.: 33.00
## Median : 50.00 Median : 7.000 Median : 0.000 Median : 70.00
## Mean : 57.92 Mean : 7.115 Mean : 7.393 Mean : 73.59
## 3rd Qu.:100.00 3rd Qu.:12.000 3rd Qu.:12.000 3rd Qu.:116.00
## Max. :100.00 Max. :12.000 Max. :60.000 Max. :145.00
## cluster
## 1:14827
## 2: 0
## 3: 0
## 4: 0
## 5: 0
## 6: 0

## th_track$cluster: 2
## quantity precio pago cliente_id
## Min. : 30.00 Min. : 0.00 Min. : 0.000 Min. :146.0
## 1st Qu.: 30.00 1st Qu.: 5.00 1st Qu.: 0.000 1st Qu.:181.0
## Median : 50.00 Median : 7.00 Median : 5.000 Median :217.0
## Mean : 50.33 Mean : 6.55 Mean : 9.077 Mean :215.2
## 3rd Qu.: 50.00 3rd Qu.: 7.00 3rd Qu.:12.000 3rd Qu.:249.2
## Max. :100.00 Max. :12.00 Max. :57.000 Max. :283.0
## cluster
## 1: 0
```

```
## 2:14032
## 3: 0
## 4: 0
## 5: 0
## 6: 0
## ------------------------------------------------------------
## th_track$cluster: 3
## quantity precio pago cliente_id cluster
## Min. : 30 Min. : 0.000 Min. : 0.000 Min. :459.0 1: 0
## 1st Qu.: 30 1st Qu.: 5.000 1st Qu.: 0.000 1st Qu.:483.0 2: 0
## Median : 50 Median : 7.000 Median : 0.000 Median :512.0 3:21802
## Mean : 61 Mean : 7.693 Mean : 8.549 Mean :513.3 4: 0
## 3rd Qu.:100 3rd Qu.:12.000 3rd Qu.:12.000 3rd Qu.:543.0 5: 0
## Max. :100 Max. :12.000 Max. :60.000 Max. :570.0 6: 0
## ------------------------------------------------------------
## th_track$cluster: 4
## quantity precio pago cliente_id
## Min. : 30.00 Min. : 0.000 Min. : 0.000 Min. :571.0
## 1st Qu.: 30.00 1st Qu.: 5.000 1st Qu.: 0.000 1st Qu.:599.0
## Median : 50.00 Median : 7.000 Median : 0.000 Median :625.0
## Mean : 54.11 Mean : 7.131 Mean : 8.687 Mean :626.4
## 3rd Qu.:100.00 3rd Qu.:12.000 3rd Qu.:12.000 3rd Qu.:653.0
## Max. :100.00 Max. :12.000 Max. :74.000 Max. :678.0
## cluster
## 1: 0
## 2: 0
## 3: 0
## 4:20427
## 5: 0
## 6: 0
## ------------------------------------------------------------
## th_track$cluster: 5
## quantity precio pago cliente_id
## Min. : 30.00 Min. : 0.000 Min. : 0.000 Min. :284.0
## 1st Qu.: 30.00 1st Qu.: 5.000 1st Qu.: 0.000 1st Qu.:316.0
## Median : 50.00 Median : 7.000 Median : 0.000 Median :349.0
## Mean : 52.78 Mean : 6.135 Mean : 6.809 Mean :350.1
## 3rd Qu.:100.00 3rd Qu.:12.000 3rd Qu.:12.000 3rd Qu.:384.0
## Max. :100.00 Max. :12.000 Max. :56.000 Max. :416.0
## cluster
## 1: 0
## 2: 0
## 3: 0

## 4: 0

## 5:15828
## 6: 0

## ------------------------------------------------------------
## th_track$cluster: 6
## quantity precio pago cliente_id
```

```
## Min. : 30.00 Min. : 0.000 Min. : 0.000 Min. :679.0
## 1st Qu.: 30.00 1st Qu.: 5.000 1st Qu.: 0.000 1st Qu.:702.0
## Median : 30.00 Median : 5.000 Median : 5.000 Median :720.0
## Mean : 51.87 Mean : 6.953 Mean : 7.635 Mean :731.3
## 3rd Qu.: 50.00 3rd Qu.: 7.000 3rd Qu.:12.000 3rd Qu.:771.0
## Max. :100.00 Max. :12.000 Max. :48.000 Max. :815.0
## cluster
## 1: 0
## 2: 0
## 3: 0
## 4: 0
## 5: 0
## 6:17264
```

5.3.4.2.5 Evaluación

Al hacer un análisis de Pareto sobre los grupos creados, se puede apreciar que los grupos #2 y #5 son los que tienen mayor cantidad de clientes asignados: en ellos se agrupa el 50 % Estos grupos grandes pueden representar al cliente promedio, aquellos que no muestran patrones significativamente diferentes a los otros clientes.

5.3.4.2.6 Resultados

El algoritmo de minería de datos utilizado pudo encontrar características muy específicas para dividir a los clientes, y formar grupos interesantes para el negocio.

-En el grupo 1 el gasto promedio es de 55 dólares.

-En el grupo 2 el gasto promedio es de 59 dólares.

-En el grupo 3 el gasto promedio es de 45 dólares.

-En el grupo 4 el gasto promedio es de 62 dólares.

-En el grupo 5 el gasto promedio es de 51 dólares.

-En el grupo 6 el gasto promedio es de 57 dólares.

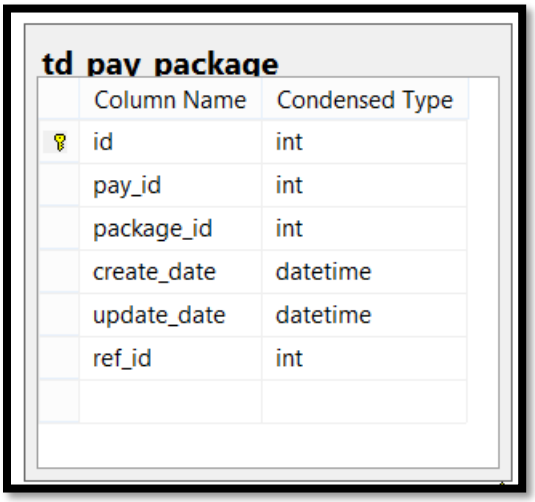
5.3.5 Cubo MOLAP

5.3.5.1 Dimensiones y jerarquías

Las dimensiones son perspectivas o factores pro lo que se hace un análisis de un área de negocio y determinan el cómo analizar a las dimensiones. Son pequeñas y usualmente no están normalizadas. Vamos a mencionar las dimensiones según el proceso de negocios:

1. Dimension de paquetes de pago, esta permitirá realizar filtros por los distintos paquetes de pago

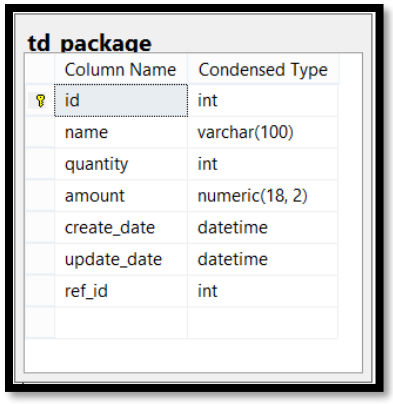
Figura 53 Dimensión pay_package



Fuente: Diseño propio.

2. Dimensión de paquetes, esta permite realizar filtros sobre los paquetes por nombre, cantidad, etc.

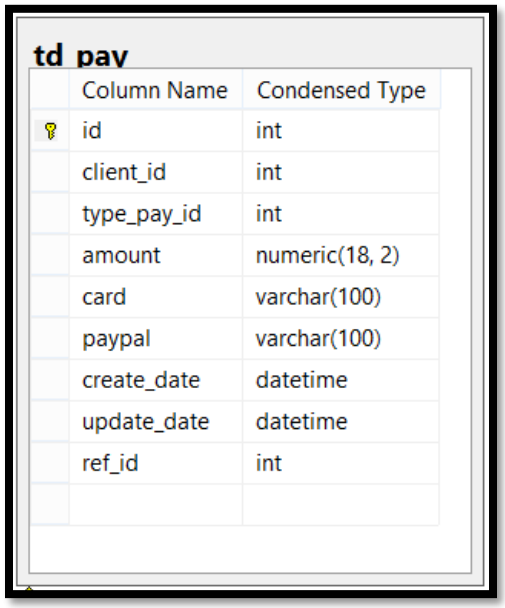
Figura 54 Dimensión td_package



Fuente: Diseño propio.

3. Dimensión de pagos, esta permitirá realizar filtros sobre las formas de pago utilizadas por los clientes

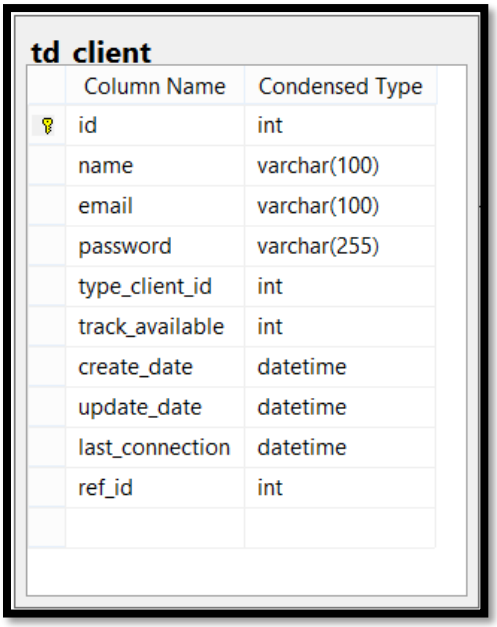
Figura 55 Dimensión td_pay



Fuente: Diseño propio.

4. Dimensión de clientes, esta permite realizar filtros por las propiedades de los clientes.

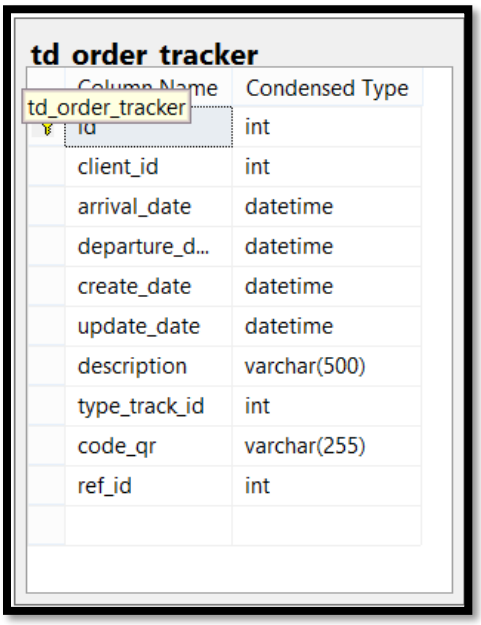
Figura 56 Dimensión td_client



Fuente: Diseño propio.

5. Dimensión de seguimiento de órdenes, esta permite filtrar sobre las propiedades de los seguimientos de órdenes.

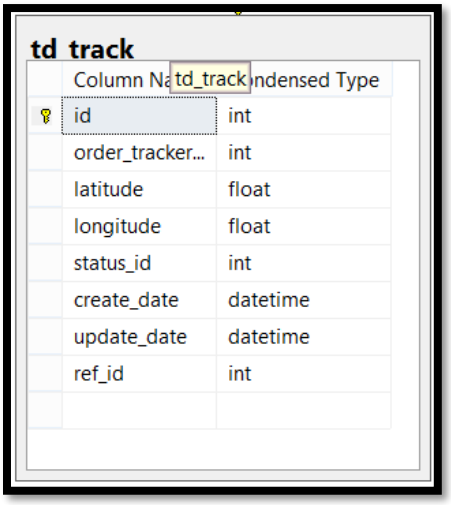
Figura 57 Dimensión td_order_tracker



Fuente: Diseño propio.

6. Dimensión de trackeo, esta permite realizar filtros sobre las propiedades del trackeo de los paquetes.

Figura 58 Dimensión td_track



Fuente: Diseño propio.

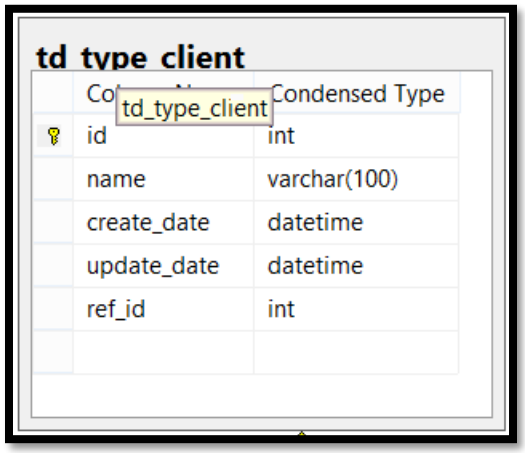
7. Dimensión de tiempo, esta permitirá poder usar la granularidad de cada campo de la dimensión de tiempo.

Figura 59 Dimensión td_tiempo

Fuente: Diseño propio.

8. Dimensión de tipos de clientes, esta permite poder filtrar por los tipos de clientes que se encuentren registrados.

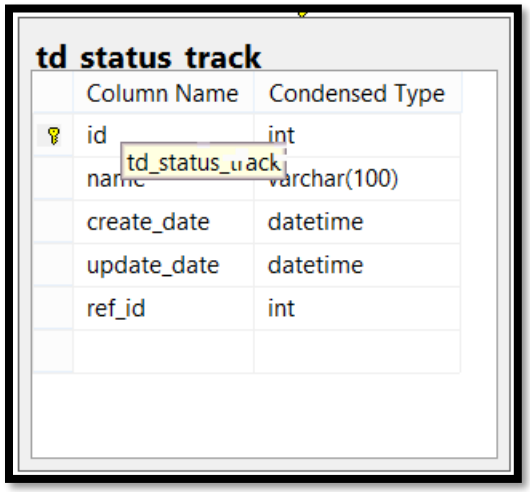
Figura 60 Dimensión td_type_client



Fuente: Diseño propio.

9. Dimensión de estado del seguimiento, permite filtrar por los diferentes estados de un seguimiento y sus propiedades.

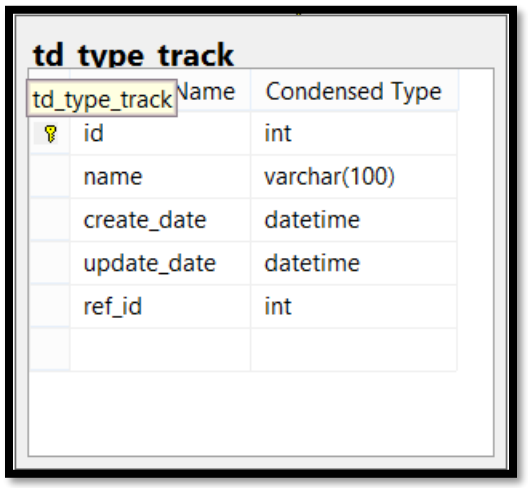
Figura 61 Dimensión td_status_track



Fuente: Diseño propio.

10. Dimensión de tipos de seguimientos, esta permite filtrar por los diferentes tipos de seguimientos que se les dan a los paquetes.

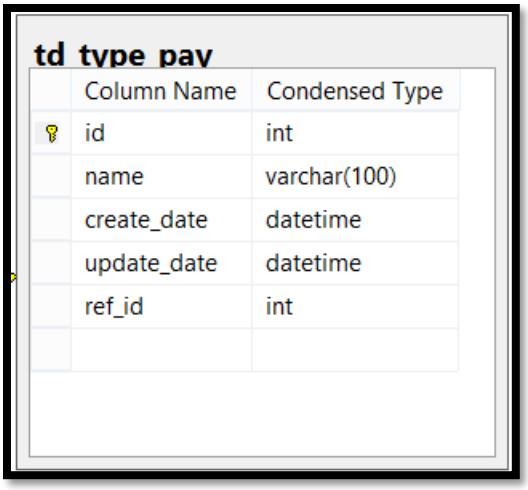
Figura 62 Dimensión td_type_track



Fuente: Diseño propio.

11. Dimensión de los tipos de pagos, esta permite realizar filtraciones por los tipos de pagos que se realizaron.

Figura 63 Dimension td_type_pay



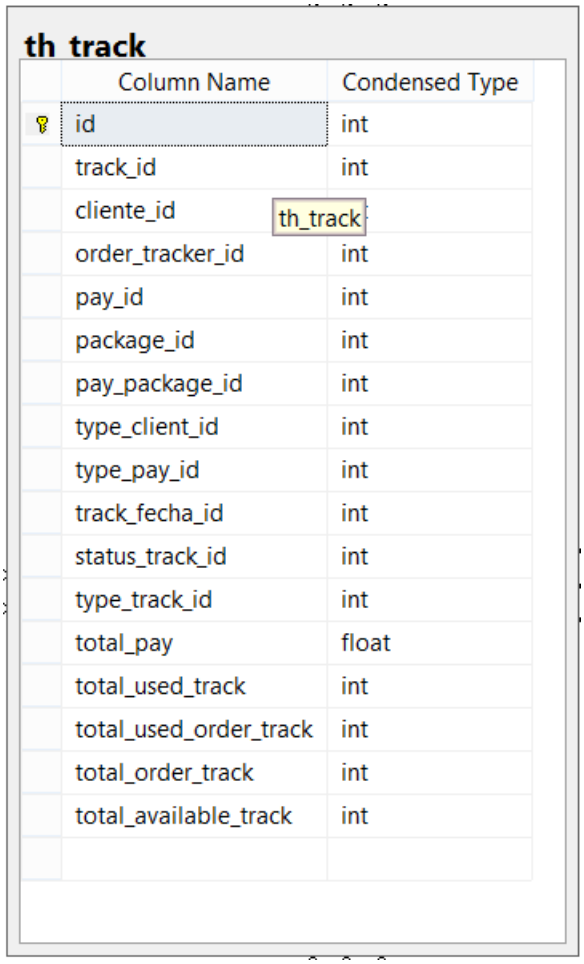
Fuente: Diseño propio.

5.3.5.2 Indicadores

La tabla de hechos está compuesta por cada uno de los id de cada dimensión y las métricas que se evaluarán. Las Métricas (también llamados KPIs, Indicadores, Valores, etc.) Son los valores numéricos generados en una o varias operaciones o transacciones de negocio. Por ejemplo: total ventas, total costes, total pagos. Ayudan a responder las preguntas referidas a cantidades o importes.

Combinando adecuadamente las dimensiones y métricas, según las necesidades de información que tenga el usuario, se podrá responder a sus interrogantes. Por ejemplo, se podrían generar los siguientes informes: ventas por cliente y meses, productos más vendidos, facturación año actual versus año anterior, etc.

Figura 64 Tabla de hechos th_track

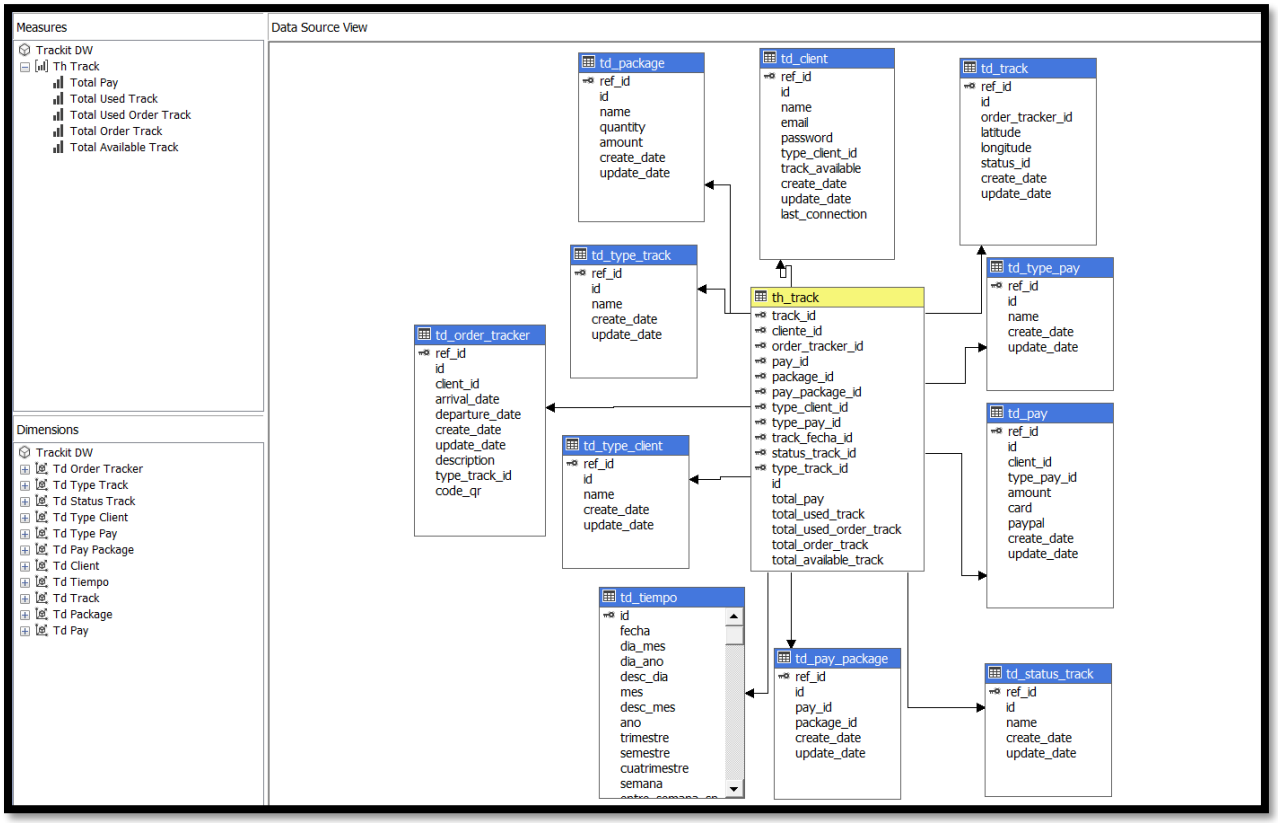


Fuente: Diseño propio.

5.3.5.3 Cubos

El modelo multidimensional para la solución de BI se generó con la herramienta Microsoft SQL Server 2013 Analysis Services. Esta herramienta permite la creación de elementos base a partir de las tablas construidas en el proceso ETL.

Figura 65 Diagrama modelo multidimensional

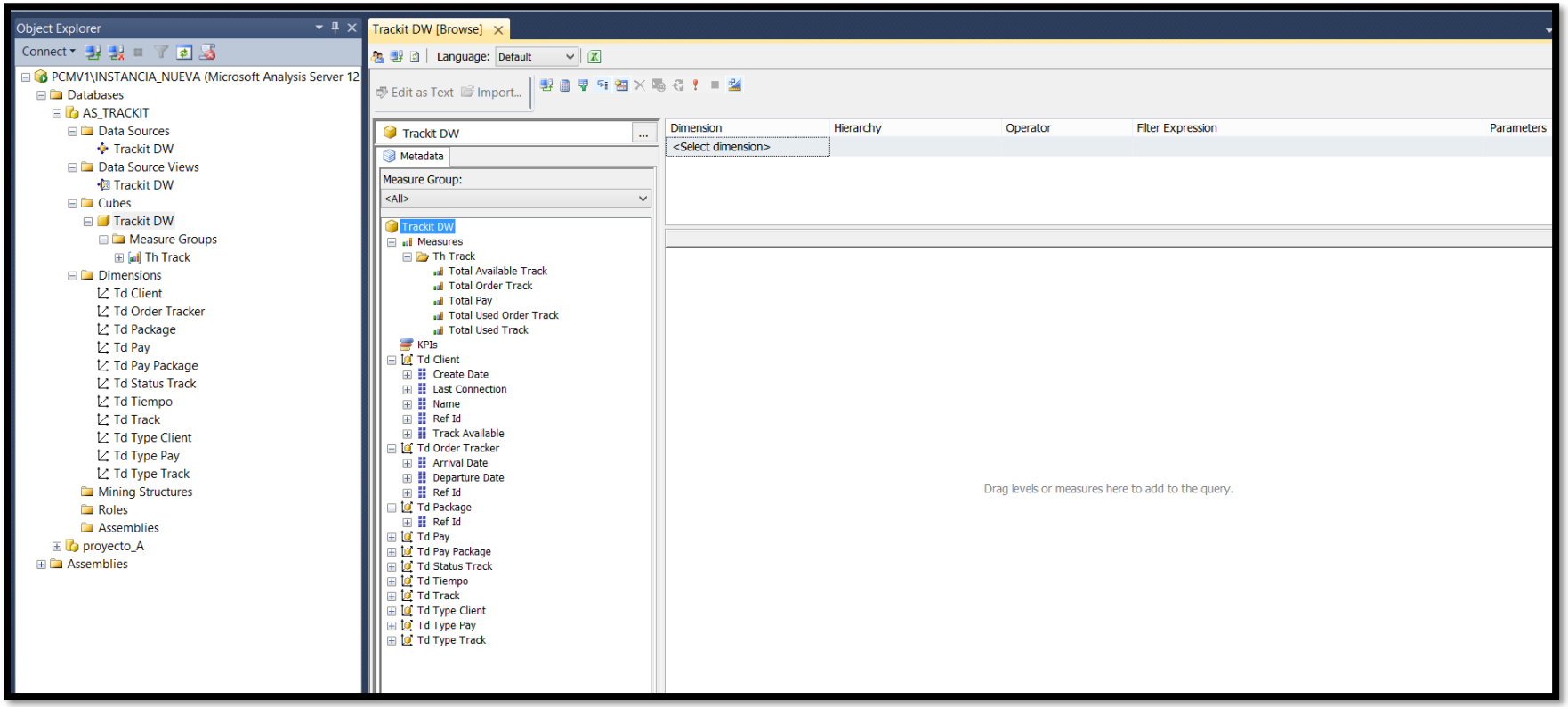


Fuente: Diseño propio.

Para la creación del cubo, utilizamos la tabla de hechos, th_track, y todas las dimensiones en el data warehouse. Definimos como métricas el pago *total , trackeo total usado*, total de órdenes usadas, total de órdenes de trackeo, total de trackeo disponibles.

La Figura siguiente muestra lo que vería un usuario utilizando SQL Management Studio:

Figura 66 Exploración de modelo multidimensional en SQL server



Fuente: Diseño propio.

5.3.6 Presentación

Debido a que ya está construido el Data Warehouse y posee un modelo multidimensional, se logra obtener el insumo necesario para explorar los datos y así obtener diferentes formas de presentación de la información, de acuerdo con las distintas necesidades que la organización presente, para así poder tomar decisiones que apoyen el funcionamiento y estrategia de la organización.

La Figura siguiente muestra las coordenadas de donde se presenta el traqueo de productos, permitiendo así saber cuáles áreas geográficas o en qué lugares hay una mayor concentración del uso del producto.

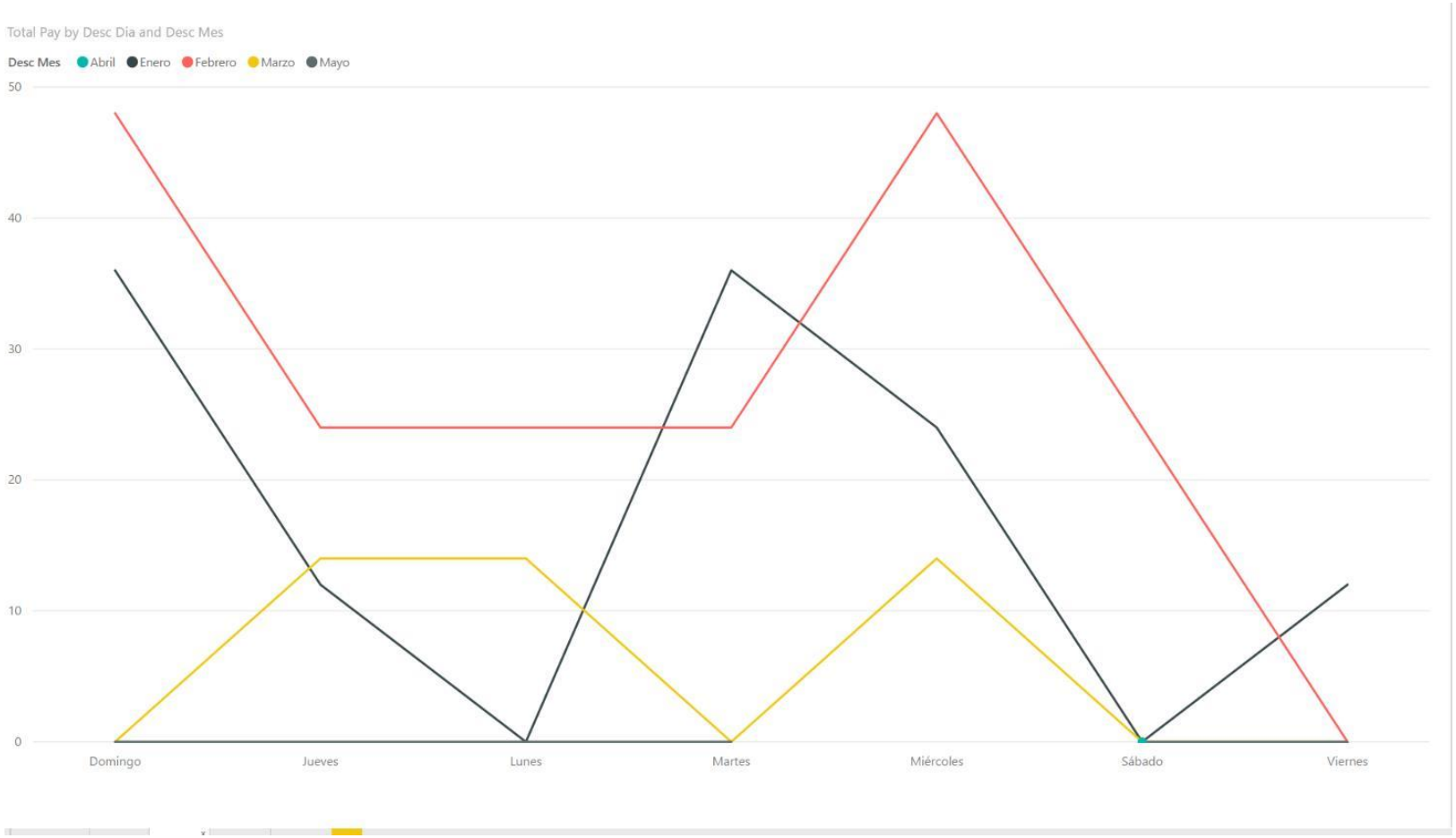
Figura 67 Presentación GPS por región

Fuente: Diseño propio.

5.3.6.1 Reportes

Los siguientes son reportes informativos de lo que está sucediendo en el sistema de Trackit.

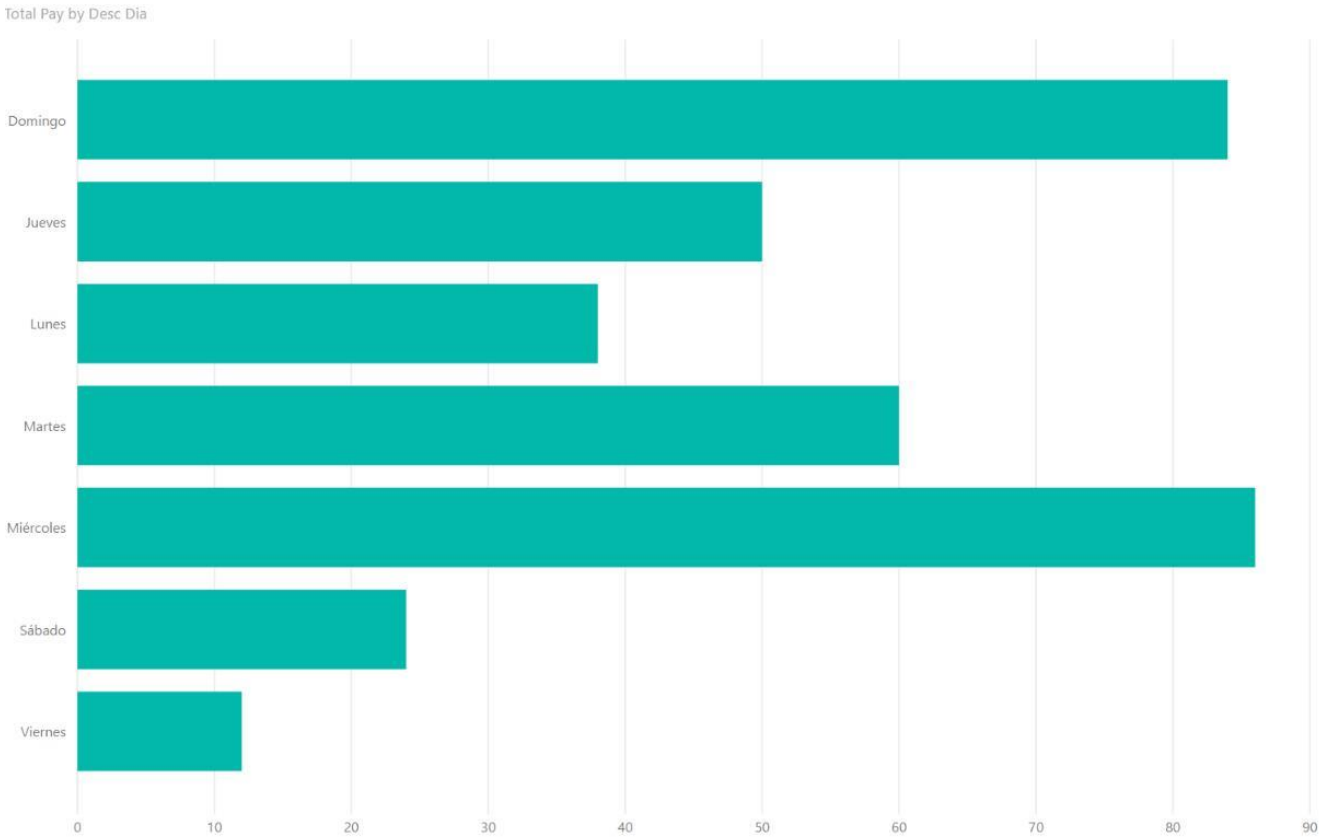
Figura 68 Reporte lineal del total de pago por mes y día de la semana



Fuente: Diseño propio.

El reporte de la figura anterior se muestra el comportamiento del total de pago según los días y el mes, poder identificar qué días se vende más y en qué mes, esto permite identificar los días en que se vende menos o como en cada mes las preferencias de consumo cambian.

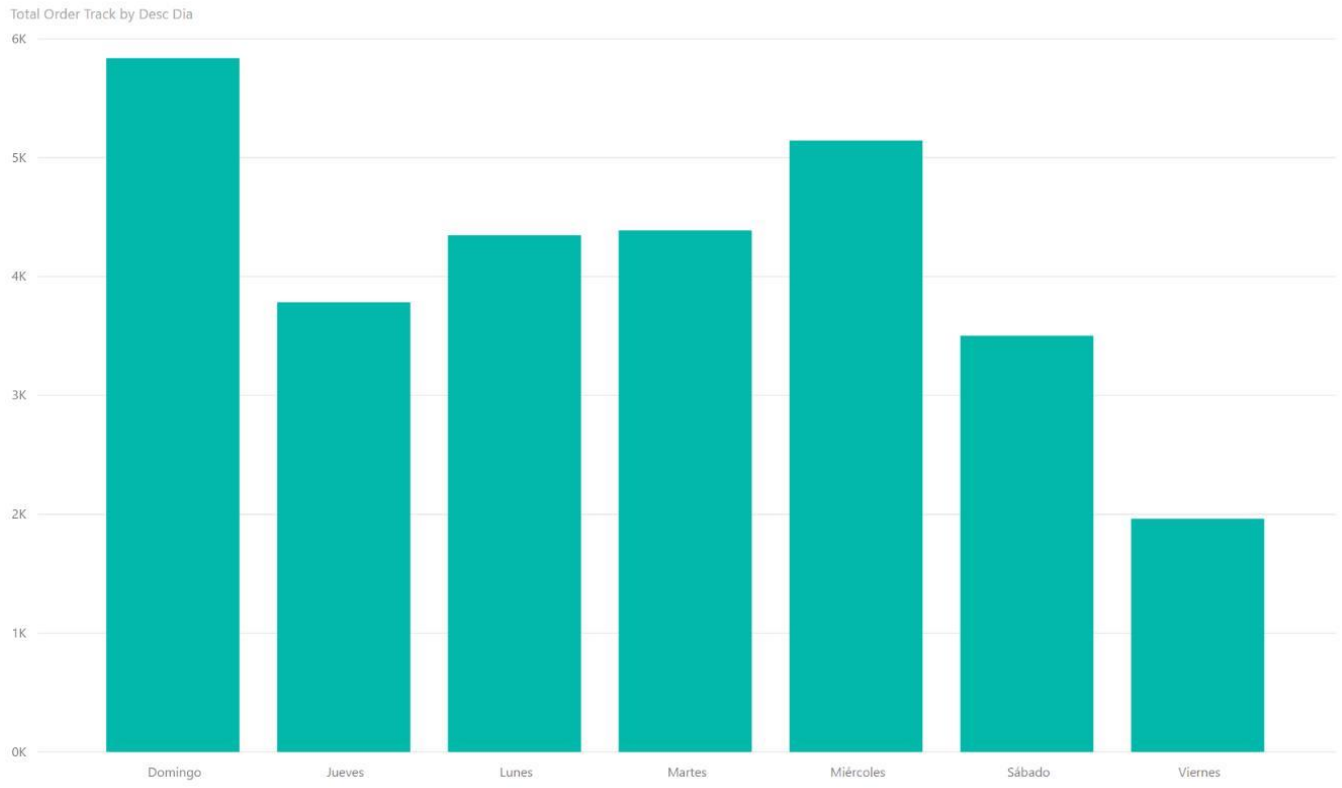
Figura 69 Reporte de total de pago por día de la semana



Fuente: Diseño propio.

El reporte antepuesto consiste en identificar la cantidad de pagos que se realizan por semana y específicamente días, dando la posibilidad de poder analizar qué días se realizan los pagos y cómo se comporta cada una de las semanas del mes.

Figura 70 Reporte del total de órdenes de track por día de la semana



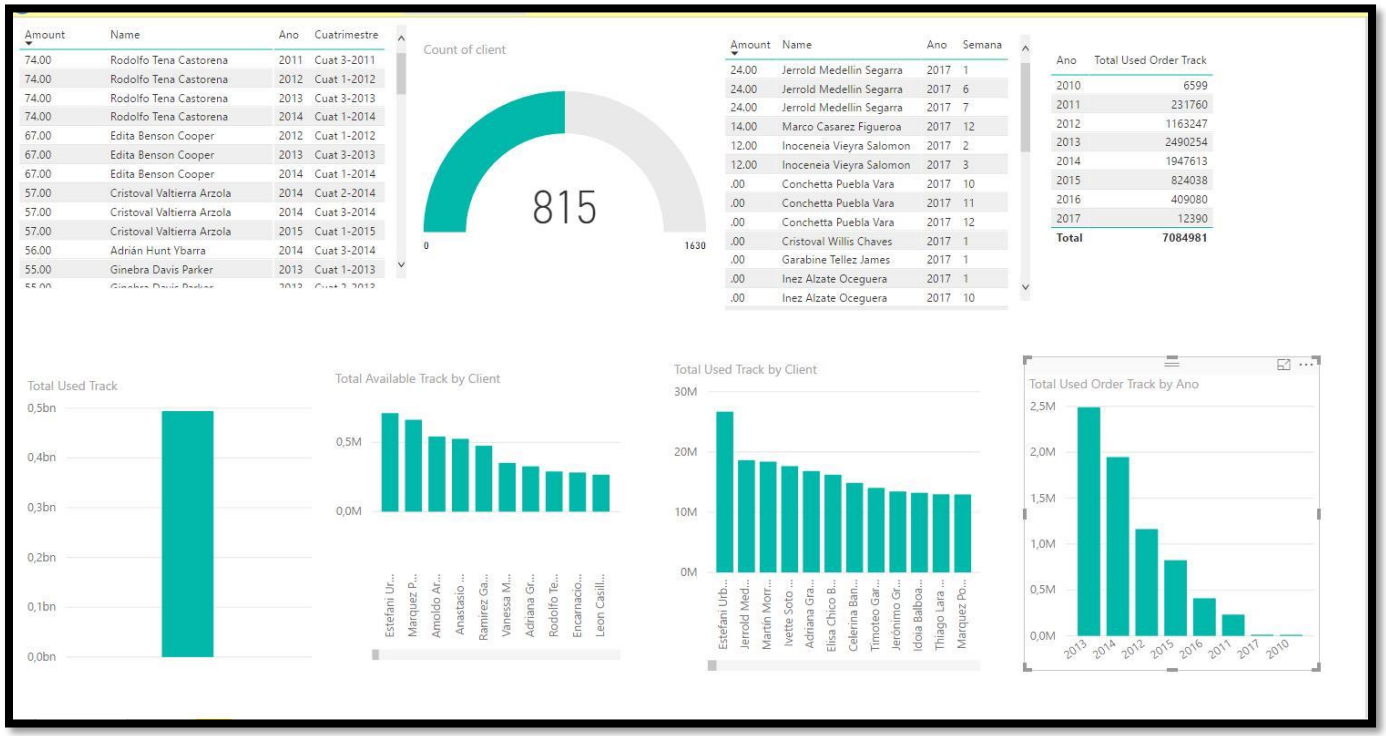
Fuente: Diseño propio.

El reporte de figura anterior muestra la cantidad de órdenes que se realizan por semana identificadas por cada día de la semana permitiendo a los tomadores de decisión impulsar las órdenes los días que hay menos movimiento.

5.3.6.2 Dashboards

La figura siguiente muestra el dashboards de la información principal del sistema trackit de la organización.

Figura 71 Dashboards



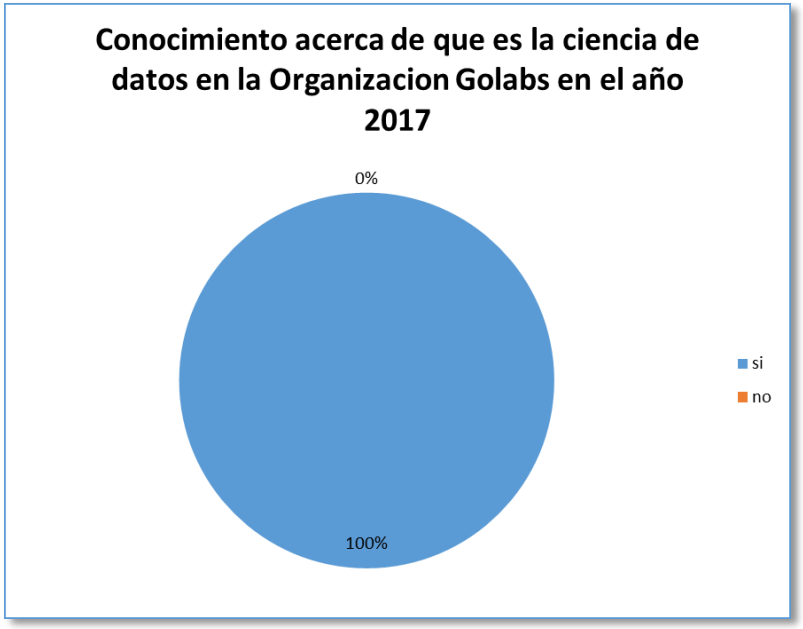
Los dashboards permiten una presentación visual de información muy relevante, la cual está consolidada y ordenada en una sola pantalla de manera de ser entendida y analizada fácilmente. El dashboard muestra los (KPI) más importantes de trackit en los que podemos encontrar el total de clientes, la cantidad de dólares gastados por clientes en el último cuatrimestre, el total de órdenes compradas por los clientes, además del uso de las órdenes de trackeo por año, y el monto de compras de los clientes en el último año. Es relevante conocer que los dashboard brindan la posibilidad de generar nuevos reportes, sin la necesidad de un experto, esto a raíz de la facilidad para agregar o eliminar nuevas columnas y desplegar nuevos gráficos a gusto del usuario.

5.4 Fase 4 Validación

La validación consiste en verificar que la información es correcta, además de ello consiste en determinar si el cliente está conforme con la solución presentada y la forma en la que se realiza y aplica la ciencia de los datos.

Para obtener una mejor perspectiva de la evaluación es necesario saber el conocimiento que tiene la organización de la ciencia de los datos, para así saber si el nivel de exigencia en cuanto a la evaluación del entregable es considerado. Por ello en esta etapa se realiza la tarea de evaluar a la organización por medio de una encuesta que permitirá observar qué tanto se conoce de ciencia de datos y la aplicabilidad de esta.

Figura 72 Gráfico circular sobre el conocimiento acerca de la ciencia de los datos en la Organización GoLabs en el año 2017

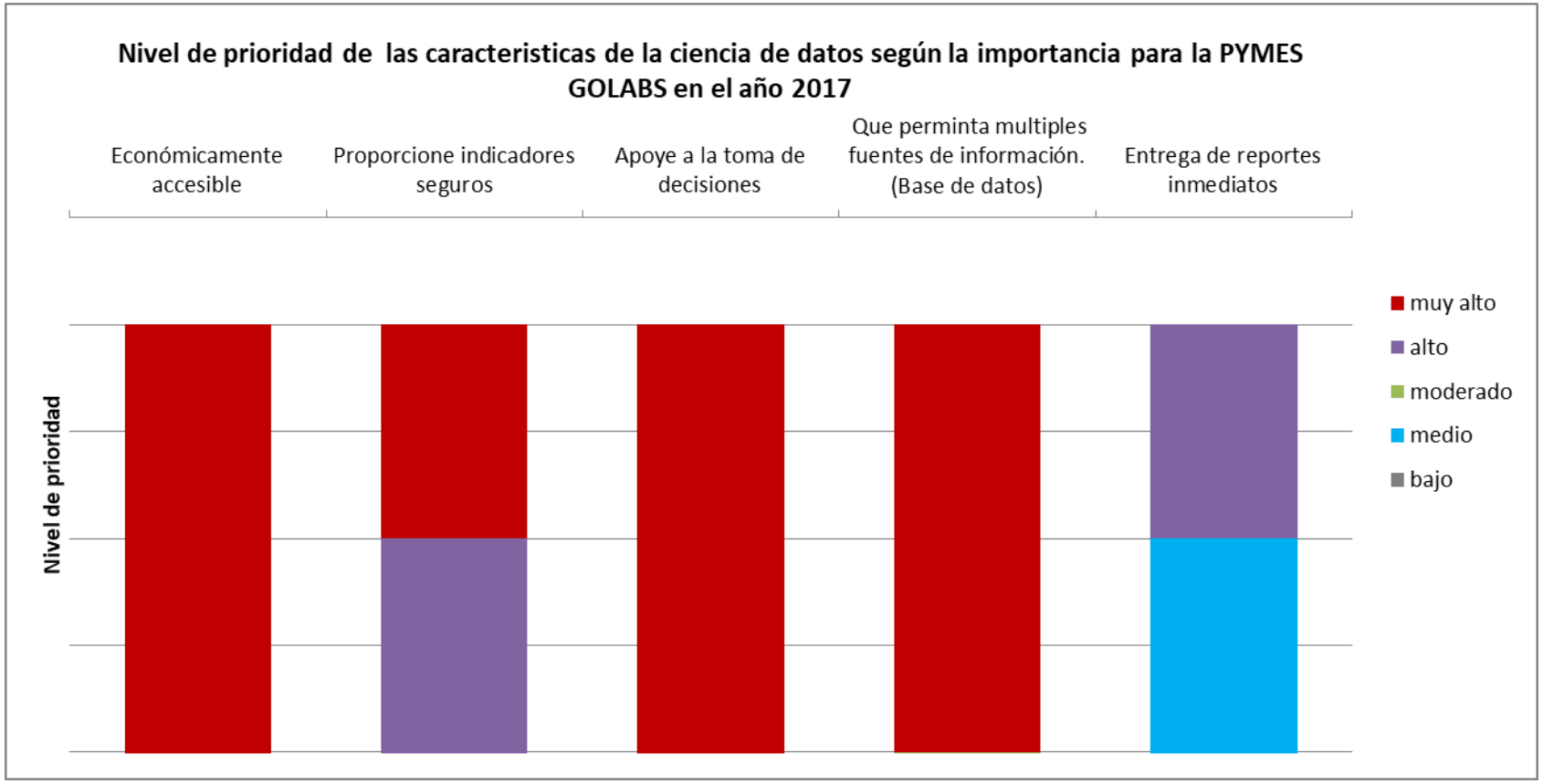


Fuente: Diseño propio.

En la figura anterior se observa que un 100% de los colaboradores conocen que es la ciencia de los datos por lo cual se deduce que habrá suficiente rigurosidad en la evaluación.

En la figura siguiente se muestra el nivel de prioridad según cada característica de la ciencia de datos, para la cual se observa que la mayoría de las características requieren que se tenga un nivel muy alto y la única característica que no posee un nivel medio y alto es el tiempo en la entrega del reporte, por lo que se deduce que se está dispuestos a esperar un momento por la información, mientras esta tenga las demás características.

Figura 73 Gráfico acumulado del nivel de prioridad de las características de la ciencia de datos según la importancia para la Pymes GoLabs en el año 2017

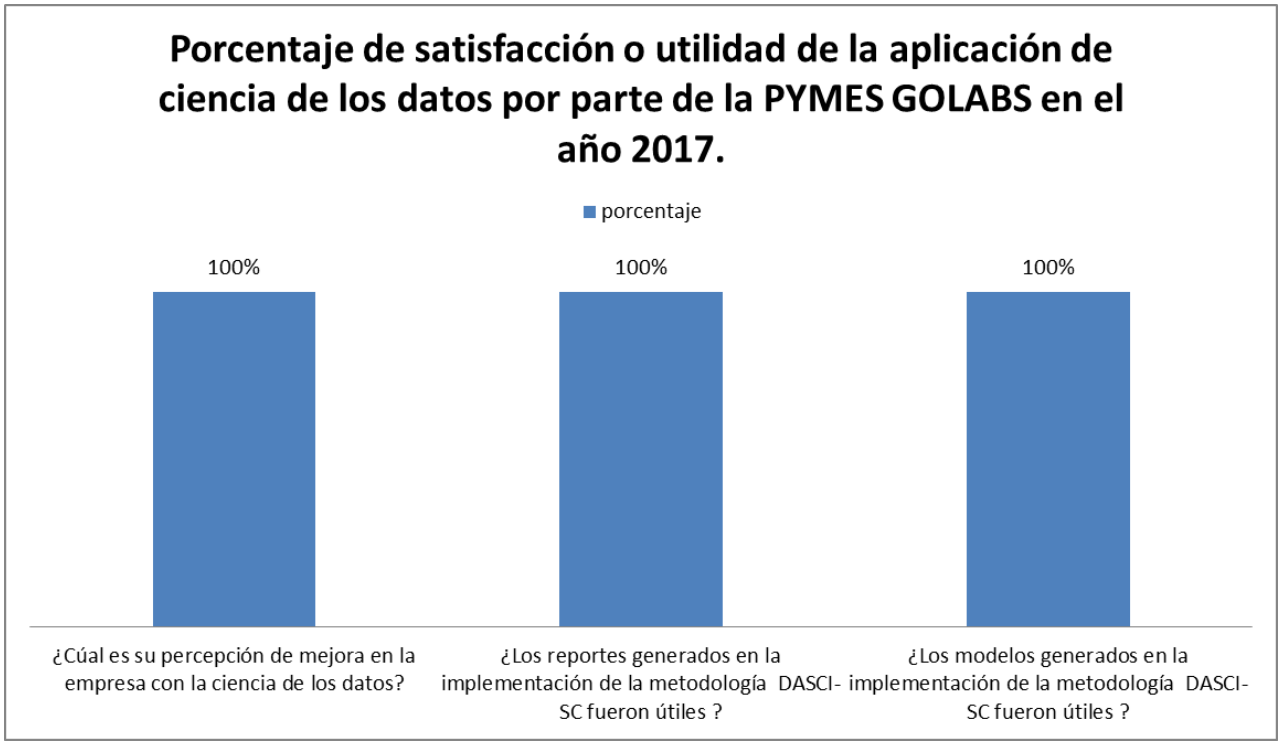


Fuente: Diseño propio.

Con la información obtenida de la encuesta se puede verificar que la evaluación por parte de la PYMES GO-LABS acerca de la solución presentada va a ser una evaluación rigurosa, siendo esto de gran importancia en este proyecto para así obtener una retroalimentación de los puntos altos y visualizar las mejoras que se pueden aplicar en futuros proyectos.

Se procede a presentar el resultado de la evaluación hecha a la Pymes GO-LABS sobre el resultado de la aplicación de la metodología.

Figura 74 Gráfico de barra sobre el porcentaje de satisfacción o utilidad de la aplicación de ciencia de los datos por parte de la Pymes GoLabs en el año 2017



Fuente: Diseño propio.

La figura anterior muestra cuál es el grado de satisfacción o utilidad que acerca de la aplicación de ciencia de los datos, el personal de Go-Labs tiene un 100% de certeza que la ciencia de los datos ha mejorado la empresa debido a que indican que el 100% de los reportes son de utilidad para la toma de mejores decisiones, así como también el 100% de los modelos, en los cuales descubrieron información para mejorar la cantidad de órdenes por paquetes.

5.5Fase 5 Implementación

Se tienen dos aspectos a los cuales se les debe poner énfasis en la implementación de la metodología, por un lado, la construcción de los depósitos de datos y por otro, la presentación de los reportes y visualizaciones. Para lo primero, debe instruirse al usuario en tres aspectos claves: contenido del Data Warehouse, aplicación y herramientas de acceso. En cuanto a la metodología, el usuario final debe entender los límites entre las aplicaciones, los contenidos y las herramientas del Data Warehouse.

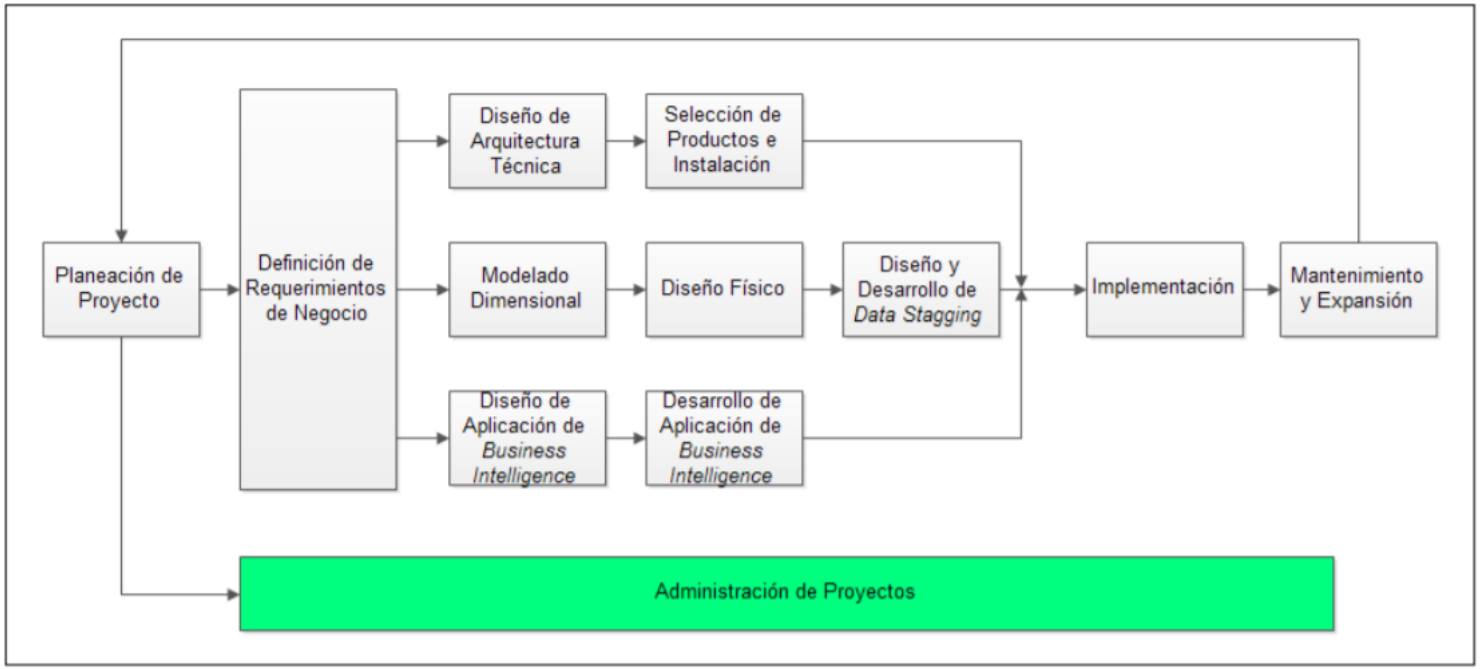
El Data Warehouse es un todo, no la suma de componentes discretos. Por lo tanto, la capacitación también debe reflejar la misma perspectiva. No debe interpretarse la capacitación de los usuarios finales como sólo el adiestramiento sobre las herramientas de acceso, pues esto resultaría inútil al menos que se complemente con la descripción del contenido en el Data Warehouse (cuáles datos están disponibles, qué significan, cómo se usan y para qué usarlos). El ensamble de un ambiente de capacitación (como subconjunto de datos de producción) es siempre recomendable, dado que los tiempos de respuesta en producción, si bien pueden ser apropiados para el análisis de la información, tal vez no sean los mejores para un curso de un día de duración. Se debe capacitar al usuario final sólo si el Data Warehouse está terminado (en tiempo y forma) y establecer la política de: “Sin capacitación no hay acceso”.

La administración mediante SCRUM el proyecto asegura que las actividades de la implementación de la metodología se lleven en forma y sincronización. Como lo indica el diagrama de la figura siguiente, la administración acompaña todo el ciclo de

vida. Entre sus actividades principales se encuentran: el monitoreo del estado del proyecto y

La comunicación entre los requerimientos del negocio, así como las restricciones de información para poder manejar correctamente las expectativas en ambos sentidos.

Figura 75 Diagrama de administración de proyecto



Fuente: (Universidad Autónoma de México, 2014)

5.5.1 Puntos problemáticos de la implementación de la metodología

Los principales puntos que pueden llegar a complicar un proyecto mediante la metodología propuesta se localizan en tres áreas:

• Rutinas de carga. Incluye programas de extracción y limpieza de datos. Surgen problemas en este punto dada la falta de integración y estructura consistente (alineada) entre los sistemas fuentes.

• Mantenimiento. Dados los diferentes períodos de almacenamiento para OLTP y OLAP y el hecho de que los Data Warehouse son sistemas secundarios de información, el problema surge para sincronizar los datos entre los sistemas operacionales fuentes y los Data Warehouse.

• Tuning. Dado los patrones de uso y los métodos de acceso de los sistemas OLAP, diseñadores y administradores deben realizar cambios significativos a los implementados en el tuning de sistemas OLTP.

• Visualización. Debido a que los reportes y sistemas de visualización se alimentan de datos provenientes del data warehouse, la veracidad de los datos podría contener datos erróneos no controlados desde la carga inicial. Por lo cual se puede llevar a malas decisiones por problemas previos.

La tabla 13 es el formulario utilizado para completar el cierre del proyecto y lecciones aprendidas.

Tabla 13 Formulario de cierre de proyecto

FORMULARIOS CP- CIERRE DE PROYECTO			
Nombre del Proyecto: Propuesta de una metodología para ofrecer el servicio de ciencia de datos en la PYMES GO-LABS ENTERPRISES		Código: 01	
Objetivo Estratégico correspondiente: Proponer una metodología, por medio de la identificación de componentes, herramientas y modelos, para el servicio de ciencia de datos de la PYMES GO-LABS ENTERPRISES			
Producto	Se realizó un análisis de las fuentes de datos y las necesidades del negocio.	No. Expediente:1	1

	Se realizó el diseño del DWH y del proceso ETL, así como el reporte por generar. Se creó el DWH y su proceso de ETL y se creará el reporte. Se realizaron las pruebas necesarias para verificar que el desarrollo funciona de manera adecuada. Se creó un dashboard para la presentación de principales reportes .		

FORMULARIO CP-1 (Proyecto concluido)				
Fecha de cierre: 15/11/2017	Se alcanzaron las metas:	Si___X__	No_____	

Presupuesto original	Presupuesto ejecutado
3200 dólares	3200 dólares

Recibo de Entregables a Satisfacción	Número de entregables:		
Nombre del entregable	Fecha de recibo	Fecha de pago	Referencia Documental
Identificación de componentes del modelo metodológico de ciencia de datos	21/5/2017		Objetivo específico 1
Describir los componentes del modelo en la metodología de ciencia de datos.	1/7/2017		*Objetivo específico 2*

Escoger las herramientas de software que se deben usar en cada componente del modelo de ciencia de datos	1/8/2017		*Objetivo específico 3*
Elaborar una guía de la implementación de los componentes del modelo de ciencia de datos	1/9/2017		Objetivo específico 4
Evidenciar cuáles son los resultados de la implementación de la metodología encontrados en el modelo de ciencia de datos para la PYMES GO-LABS ENTERPRISES.	1/11/2017		Objetivo específico 5

LECCIONES APRENDIDAS	RECOMENDACIONES
Con el transcurrir del tiempo aparecen nuevos requerimientos, nuevos indicadores que en su momento no existían o no se contemplaron, en ese sentido es importante que la aplicación de ciencia de datos este constantemente actualizándose, para que no pierda eficiencia y eficacia en cuanto a la información que brinda para el soporte de la toma de decisiones.	

ACEPTACIÓN FINAL DEL PROYECTO		
Beneficiario	Administrador del proyecto	Propietario
GO-LABS ENTREPRISES	EFREN JIMÉNEZ DELGADO	GO-LABS ENTREPRISES

Fuente: Diseño propio.

6 Capítulo VI

6.1Conclusiones

En la construcción de este trabajo se hizo una investigación en la que se logró integrar la información de diferentes fuentes con diversos enfoques, que se consideraron compatibles con este trabajo de tesis, la cual enriquece de esta forma el trabajo. La metodología mostrada en este trabajo toma en cuenta los contenidos de estas fuentes para tratar de satisfacer las necesidades de la Pymes Go-Labs y posteriormente poderlo aplicar a otras Pymes que posean necesidades semejantes.

Dentro de las conclusiones durante la implementación de la metodología se puede mencionar los siguientes puntos:

- Se logró identificar los procesos que permiten llevar a cabo la toma de decisiones, la cual permite que se realice un análisis de los requerimientos de la empresa.
- Se estudió el Departamento de ventas y producción de la empresa Go-Labs estableciendo de manera correcta, las medidas que se utilizan en la realización de reportes para la toma de decisiones.
- Se analizó y seleccionó los datos indispensables para el desarrollo de la solución de inteligencia de negocios, lo que garantizó la validez y calidad de los mismos
- Se construyó un modelo de datos OLAP, que permitió ejecutar las consultas, a partir de información previamente procesada, obteniendo como resultado la flexibilidad al usuario al realizar las diferentes consultas pree laboradas. Además, se efectuaron las pruebas para corregir los errores siguiendo la solución de inteligencia de negocios.
- Se diseñó las interfaces, mediante la herramienta Power BI para desplegar el cubo OLAP, lo cual acelera la generación de reportes tipo dashboard.

- Se diseñó los reportes de minería de datos, mediante el lenguaje R y su paquetería, lo que permitió una narrativa más limpia e entendible para los usuarios encargados de la toma de decisiones.
- Con la implementación de la metodología se incrementó la velocidad con la cual se pueden clasificar, agrupar clientes, además de conocer los posibles nuevos grupos de paquetes según los gustos generados por los mismos clientes.

La metodología está enfocada en, primordialmente, conocer el negocio para el cual se pretende desarrollar el conocimiento, esto permite que el líder del negocio establezca las prioridades y se realice un desarrollo progresivo que sea dé de acuerdo con el nivel de madurez. De tal forma que genere una perspectiva de los tiempos que implicará obtener la información y no tener una perspectiva errónea donde los tiempos de trabajo sean limitados y eso conlleve a no obtener los resultados esperados. Lo que se ha pretendido es que esta metodología sea un modelo para las PYMES, brindando el conocimiento para poder experimentar y explorar el mundo de la ciencia de datos en los negocios para que cada día puedan ser más competitivas.

La metodología abarca los principales elementos de la ciencia de los datos conocidos como lo es la inteligencia de negocios para ser aplicada, también aspectos del negocio, un poco de administración de proyectos y los tres principales elementos de la inteligencia de negocios que son la construcción de un Data Warehouse, la construcción de los procesos ETL que permiten poblar el Data Warehouse, los procesos de análisis como son cubos de información y minería de datos y la explotación por medio de herramientas que permiten presentar la información por medio de reportes, dashboard, etc.; los cuales permiten a los líderes de la organización tomar las decisiones necesarias para alcanzar los objetivos establecidos durante la planificación de las estrategias del negocio.

El desarrollo de soluciones de inteligencia de negocios con herramienta open source involucra un proceso de mayor complejidad con una curva de aprendizaje bastante más amplia, además de existir una menor cantidad de documentación con la que se gustaría contar y que fuera lo suficientemente clara para así poder resolver las

actividades o problemas de forma rápida. Es por eso que se considera que es muy importante el aprovechamiento de las herramientas propietarias que logren enfocar y tomar en cuenta a la Pyme como un mercado potencial, abriéndoles las puertas al mundo de la ciencia de datos ofreciendo versiones de sus herramientas adecuadas para ellas. De esta manera se puede aprovechar la estabilidad que ofrece una herramienta propietaria versus una herramienta libre.

No es necesario estar atado a un solo tipo de herramientas en el desarrollo de la ciencia de datos, ya que se puede trabajar en conjunto entre herramientas libres y herramientas propietarias aprovechando las facilidades que brindan y posibilidades de integración de cada una, aunado a ello la posibilidad de disminuir costos.

6.2 Recomendaciones

Como valor agregado al trabajo elaborado durante la implementación de la metodología, se presentarán a continuación una serie de recomendaciones con el fin de mejorar aspectos a la empresa Go-Labs:

- Es importante establecer una adecuada comunicación desde el principio con los departamentos en los que se realiza la implementación, esto mejorará la percepción del trabajo realizado por medio de la metodología, se podrá tener retroalimentación de los usuarios, se detectará posibles cuellos de botella o problemas de operación de las herramientas que evitarán la mala percepción de los servicios prestados.
- La principal estrategia de sistemas de ciencia de datos en resumen es el conocimiento a fondo la organización antes de comenzar a implementar los servicios en la compañía. Se deben tener claros aspectos tales como: visión, misión, valores, objetivos, estrategias, productos, clientes, etcétera.
- El principal factor de éxito en un proyecto de ciencia de datos es asegurar el patrocinio de los ejecutivos de la empresa, por ello es preciso vender correctamente la solución a los ejecutivos de alto nivel de la empresa y no solo al personal de Tecnologías de Información.
- Conocer las prioridades de los usuarios ejecutivos es muy importante para que las herramientas generen la información exacta que este tipo de usuarios necesita, sin sobrecargarlos de reportes, menús o pantallas.
- Es muy importante evitar los problemas con los usuarios, esto se puede lograr mediante un adecuado programa de capacitación que permita a los usuarios de todos los niveles y unidades funcionales de la organización conocer no solo el uso básico de las herramientas, sino todo el potencial que éstas pueden ofrecer.
- Por lo percibido en las entrevistas, los niveles de alta gerencia tienen acceso al conocimiento de las implicaciones de implementar ciencia de datos en la compañía. Es por esto que se recomienda estimular e incentivar en todo el personal de la compañía el uso de las herramientas, para lograr desarrollar un

lenguaje común, promover la responsabilidad y eficiencia además de estimular la curiosidad a todos los niveles de la organización.

- Sería de gran importancia que en proyectos de aplicación de ciencia de los datos se incluya un análisis sobre el nivel de madurez de las organizaciones y cómo este influye en el funcionamiento de estas y patrocinio de proyectos de ciencia de los datos.

7 Bibliografía

Krawatzeck, R., Barbara, D., & Duc Anh, P. (2015). *How to Make Business Intelligence Agile: The Agile BI Actions Catalog.*

Brock, J. D. (10 de 04 de 2015). *Being the DBA (database administrator): nifty assignment.* USA: Journal of Computing Sciences in Colleges. Obtenido de http://delivery.acm.org/10.1145/2840000/2831477/p275-brock.pdf?ip=181.193.125.10&id=2831477&acc=PUBLIC&key=842BC1E250410AEB%2EFDB887D4E02C11F2%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&CFID=922506782&CFTOKEN=64259032&__acm__=1491877647_dc219fc5e6388672a17fece

Datalabs. (2017). *Tableau Business Intelligence Dashboard Designer.* Obtenido de http://www.datalabsagency.com/tableau-business-intelligence-dashboard-designer/

Febles, J. R. (2005). *KDD y MD.* Obtenido de http://slideplayer.es/slide/3736333/

Guzmán Arenas, A., Martínez Luna, G., & Orantes Jiménez, S. (2015). *Ciencia de los datos.* Obtenido de http://www.comunidad.cic.ipn.mx/Red_computo/proyectos-insignia/ciencia-de-los-datos/

Huamantumba, R. (2007). *Datamart paso a paso.* Obtenido de http://www.raynerhd.com/wp-content/uploads/rayner-datamart.pdf

Jaramillo Parra, C. (2015). *Los Supuestos y Restricciones en proyectos.* Obtenido de https://sites.google.com/site/upcintroagerencia/los-supuestos-en-proyectos

kdnuggets. (2014). *kdnuggets*. Obtenido de http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Konta Asesores. (2014). *Cuadro de mando de ventas.* Obtenido de http://konta.es/cuadro-de-mando-de-ventas-konta-asesores-vitoria-gasteiz/

M Amala Jayanthi, R. L. (2016). *Research contemplate on educational data mining.* Advances in Computer Applications (ICACA), IEEE International Conference on.

Marsh, G. (2000). *Knowledge management: a cookbook for beginners.* New York: ACM New York.

Mazhar Hameed, U. Q. (2016). *Business intelligence: Self adapting and prioritizing database algorithm for providing big data insight in domain knowledge and processing of volume based instructions based on scheduled and contextual shifting of data.* Future Technologies Conference (FTC).

MEIC. (2016). *Estados situcion de las PYME en Costa Rica 2016.* http://reventazon.meic.go.cr/informacion/pyme/2017/informe.pdf.

Miciruel Ftichll, D. J. (1997). *Assistant for art Information Database.* New York: ACM New York.

Obiols, A. (2015). Obtenido de ¿Qué es un Data Scientist?: https://inlab.fib.upc.edu/es/blog/que-es-un-data-scientist

Tecnológico de Costa Rica. (2017). *Algoritmos en Ciencia de los Datos para la aplicación en pequeñas y medianas empresas .* Obtenido de http://www.compdes.org/compdes2017/docs/LibroCompdes2017.pdf

Universidad Autonóma de México. (2014). *ANÁLISIS DE INFORMACIÓN Y TOMA DE DECISIONES PARA ADMINISTRACION DE NEGOCIOS.* Obtenido de http://www.ptolomeo.unam.mx:8080/xmlui/bitstream/handle/132.248.52.100/3243/Tesis.pdf

Wagner, P. (10 de 04 de 2005). *Teaching Data Modeling: Process and Patterns.* Caparica, Portugal: ACM New York. Obtenido de http://delivery.acm.org/10.1145/1070000/1067493/p168-wagner.pdf?ip=181.193.125.10&id=1067493&acc=ACTIVE%20SERVICE&key=842BC1E250410AEB%2EFDB887D4E02C11F2%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&CFID=922506782&CFTOKEN=64259032&__acm__=1491877966_92f017dff16d

8 GLOSARIO

Pyme. Son empresas con un número reducido de ingresos y empleados.

Productividad. Es la relación que existe entre lo que genera una empresa y los recursos humanos con los que cuenta.

Competitividad. Es la capacidad de las empresas de hacer frente a la competencia que existe en el mercado.

Dato. Características aisladas de entidades.

Información. Son un conjunto de datos que al relacionarse tienen un significado.

Conocimiento. Es información que es almacenada y puesta a disposición de los interesados para que pueda realizar y/o mejorar sus actividades, permitiéndoles tener un aprendizaje.

OLAP. Procesamiento analítico en línea, permite utilizar estructuras multidimensionales que permiten agilizar las consultas.

OLTP. Procesamiento transaccional en línea, permite la administración de la información transaccional generada por aplicaciones operativas.

ROLAP. Tipo de procesamiento analítico que está basado en un esquema relacional.

MOLAP. Tipo de procesamiento analítico basado en un esquema multidimensional.

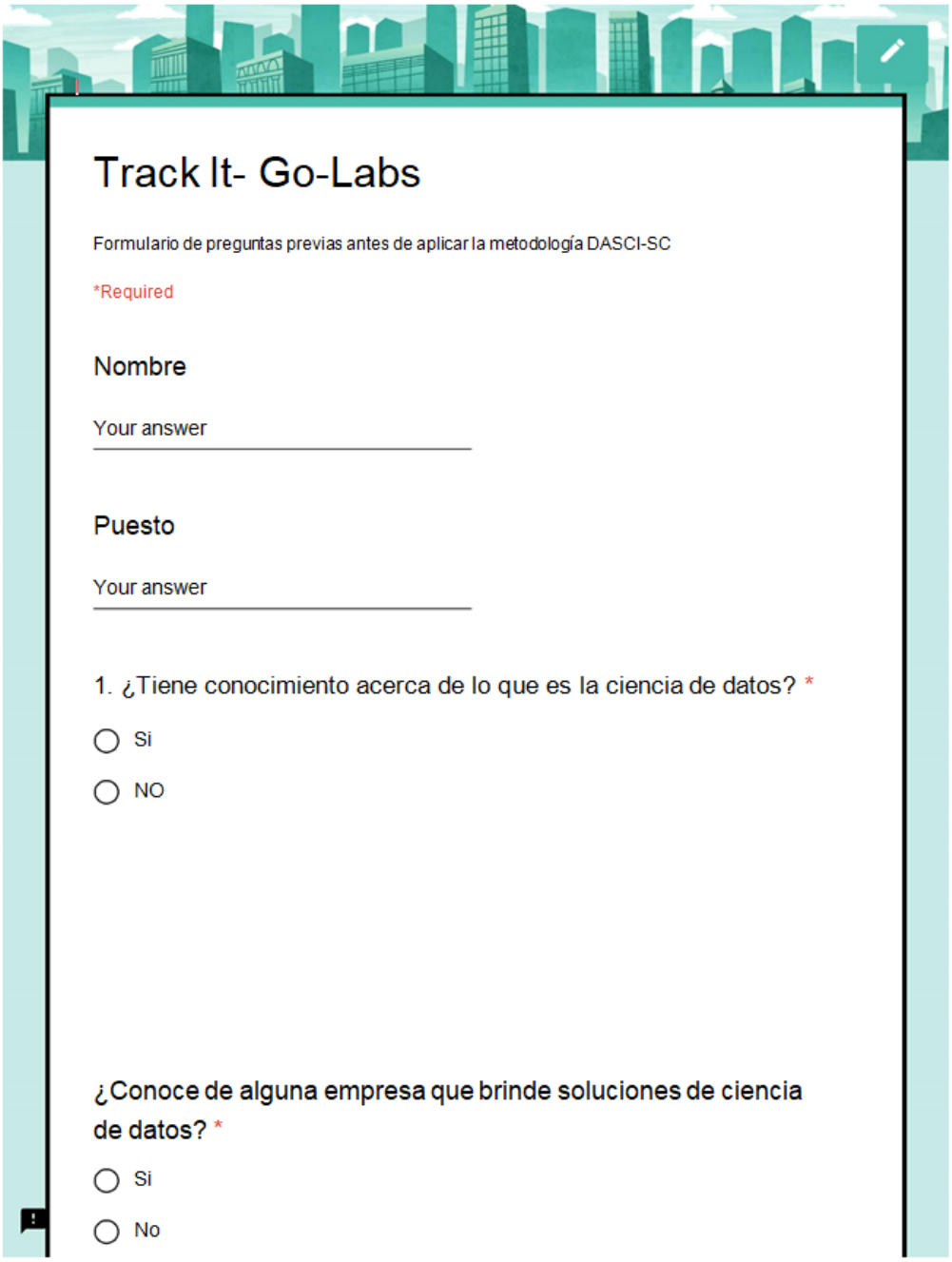
HOLAP. Tipo de procesamiento analítico que mezcla las características de un esquema relacional con un esquema multidimensional.

KDD. Acrónimo para Knowledge Discovery in Databases, que consiste en el proceso que se sigue para poder descubrir conocimiento de las bases de datos.

XML. Lenguaje de marcas extensible, es un lenguaje de etiquetas que permite almacenar información.

9 Anexo

9.1 Encuesta conocimiento de la ciencia de los datos



Track It- Go-Labs

Formulario de preguntas previas antes de aplicar la metodología DASCI-SC

*Required

Nombre

Your answer

Puesto

Your answer

1. ¿Tiene conocimiento acerca de lo que es la ciencia de datos? *

- Si
- NO

¿Conoce de alguna empresa que brinde soluciones de ciencia de datos? *

- Si
- No

¿Cuan importante considera aplicar soluciones de ciencia de datos para tomar decisiones? *

	1	2	3	4	5	6	7	8	9	10	
No es importante	○	○	○	○	○	○	○	○	○	○	Total importancia

De las siguientes características enumere del 1 al 5 según su prioridad, cuál son las características más importantes a considerar para adquirir una solución de ciencia de datos, el valor 1 es el mas importante. *

	1	2	3	4	5
Económicamente accesible	○	○	○	○	○
Proporcione indicadores seguros	○	○	○	○	○
Apoye a la toma de decisiones	○	○	○	○	○
Que perminta multiples fuentes de información. (Base de datos)	○	○	○	○	○
Entrega de reportes inmediatos	○	○	○	○	○

Marque las áreas que considere importantes para ser analizadas por una solución de Inteligencia de Negocio

- ☐ Control Pnanciero
- ☐ Rentabilidad de un producto concreto

- [ ] PlaniEcación de la producción
- [ ] Optimización de costos
- [ ] Análisis de perWles de clientes

¿Cuanto tiempo invierte en generar reportes gerenciales?

	1	2	3	4	5	6	7	8	
1 hora	○	○	○	○	○	○	○	○	8 horas

¿Cuando estaría dispuesto a pagar anualmente por el uso de la ciencia de datos?

Your answer _____

SUBMIT

Never submit passwords through Google Forms.

9.2Encuesta de evaluación de los resultados

Track It- Go-Labs

Formulario de preguntas previas después de aplicar la metodología DASCI-SC

Nombre

Tu respuesta

Puesto

Tu respuesta

Después de la metodología de ciencia de datos aplicada a Go-Labs. ¿Cual es su percepción de mejora en la empresa con la ciencia de los datos?

	1	2	3	4	5	
1- Nada	○	○	○	○	○	2- Excelente

¿Cual es su calificación después del trabajo realizado en la implementación de la metodología DASCI-SC ?

	1	2	3	4	5	
1- Malo	○	○	○	○	○	5- Excelente

https://docs.google.com/forms/d/e/1FAIpQLScPpnJSTIqJtXjzGXoAQJW4xg943RhZ_0svPQJD-IQ5w7jvIw/viewform?c=0&w=1 1/

Los reportes generados en la implementación de la metodología DASCI-SC fueron útiles ?

- ◯ SI
- ◯ No

Los modelos generados en la implementación de la metodología DASCI-SC fueron útiles ?

- ◯ SI
- ◯ NO

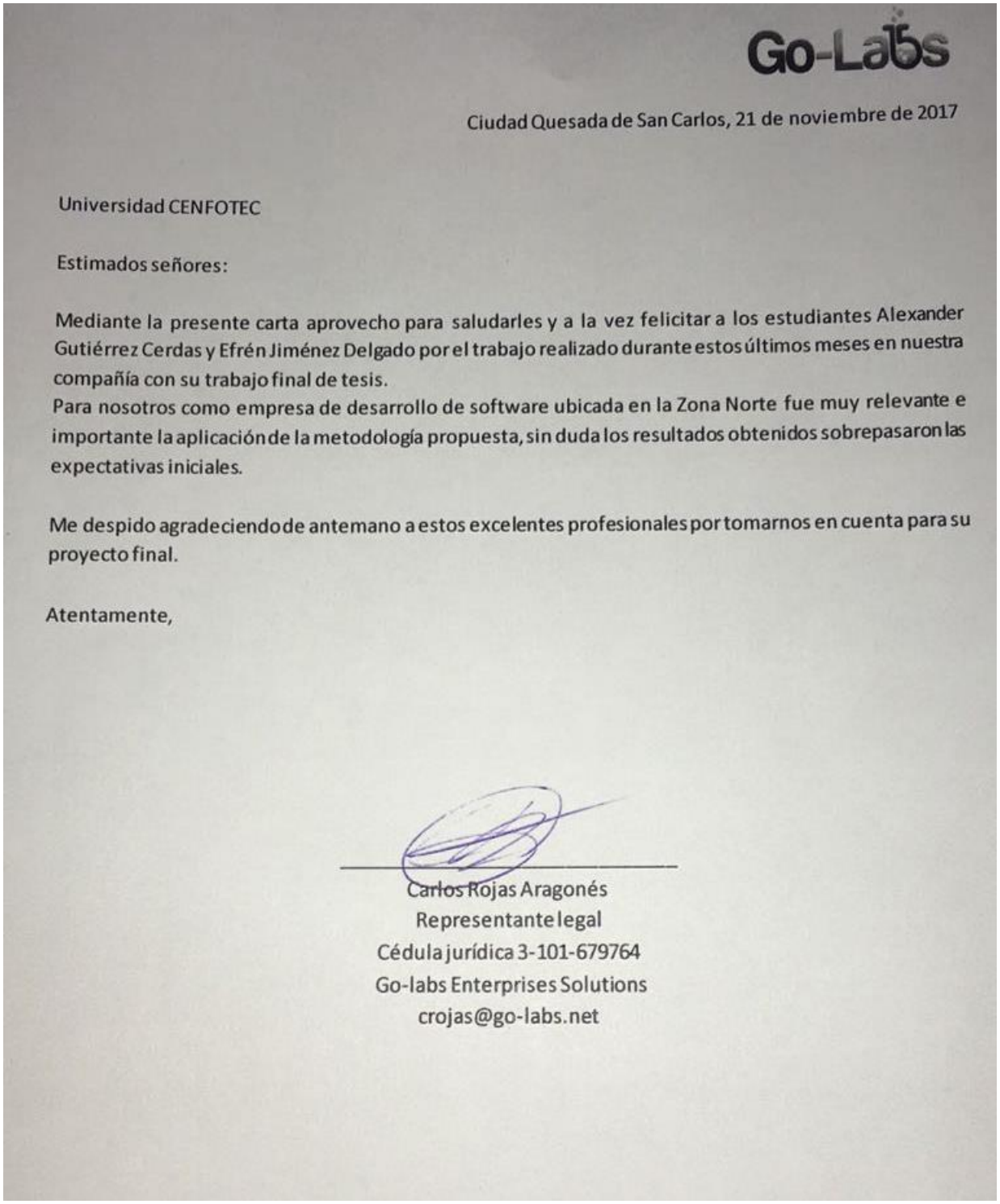
El nivel de satisfacción con la implementación de la metodología DASCI-SC en Go-Labs?

	1	2	3	4	5	
1 Nada	◯	◯	◯	◯	◯	5 Excelente

ENVIAR

Nunca envíes contraseñas a través de Formularios de Google.

9.3Carta de aprobación de proyecto final



Go-Labs

Ciudad Quesada de San Carlos, 21 de noviembre de 2017

Universidad CENFOTEC

Estimados señores:

Mediante la presente carta aprovecho para saludarles y a la vez felicitar a los estudiantes Alexander Gutiérrez Cerdas y Efrén Jiménez Delgado por el trabajo realizado durante estos últimos meses en nuestra compañía con su trabajo final de tesis.
Para nosotros como empresa de desarrollo de software ubicada en la Zona Norte fue muy relevante e importante la aplicación de la metodología propuesta, sin duda los resultados obtenidos sobrepasaron las expectativas iniciales.

Me despido agradeciendo de antemano a estos excelentes profesionales por tomarnos en cuenta para su proyecto final.

Atentamente,

Carlos Rojas Aragonés
Representante legal
Cédula jurídica 3-101-679764
Go-labs Enterprises Solutions
crojas@go-labs.net