UNIVERSIDAD DE CHILE

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

DISEÑO Y CONSTRUCCIÓN DE UN DATA WEBHOUSE DE APOYO A LA INDUSTRIA DEL TURISMO DE LA X REGIÓN.

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN.

MANUEL ALEJANDRO CASTRO ASTETE

PROFESOR GUÍA
JUAN D. VELASQUEZ SILVA

MIEMBROS DE LA COMISIÓN
BÁRBARA POBLETE LABRA.
DIONISIO GONZÁLEZ GONZÁLEZ.

SANTIAGO DE CHILE
2014

Financiado por el proyecto FONDEF D10|1198 – WHALE

RESUMEN.

El desafío propuesto en este trabajo de memoria consistió en el desarrollo y construcción de un Data Webhouse que albergue distintos componentes asociados a un sitio web que presenta un catálogo de productos en internet, como lo son los módulos de recomendador de productos (AWS), de análisis de opinión en internet (WOM), y de registro de acceso al sitio web (Web Log), aplicados a un proyecto real relativo al turismo de la X Región.

Apoyado en los avances y desarrollo de memoristas anteriores en temas relativos al web Mining, análisis de log, y sitios adaptativos es que se presenta el proyecto W.H.A.L.E. que alberga a la industria del turismo de la X Región, bajo a supervisión del Gobierno Regional de la X Región y el Sernatur, para buscar potenciar ese sector económico a través del desarrollo y estudio del sitio “patagonialoslagos” en donde se aplican estos avances por medios de módulos independientes.

El problema propuesto es justamente consolidar dichos avances en uno solo, a través de un modelo de datos que considere cada uno de los módulos mencionados y permita mejores puntos de vista a través de información agregada y combinada, bajo el supuesto que el conjunto es mayor que las individualidades.

El desarrollo de la presente memoria consistió en el desarrollo de un modelo de Data Webhouse que contuviera la información del sitio web Patagonia, y que agregara información pública de valor, generando un modelo en la medida de la disponibilidad de esa información, además se crearon las interfaces de carga para incorporación del registro de visitas del sitio web a través de su web log, al igual que las interfaces de carga de datos desde el módulo de análisis de opinión, sin embargo esto no se logró con el módulo de recomendación de productos ya que su funcionamiento está codificado dentro de la aplicación tomando por sí mismo los datos de la base de datos de la página web. Con el modelo resultante se generó una base de reportes que combina los datos de los productos visitados y las características relacionadas.

Como conclusiones más importantes está la prueba empírica de integración de datos entre los módulos, que se logra en parte, y que sugiere que en una próxima versión del mismo, el diseño de los módulos deben desarrollarse en conjunto con el sitio web para asegurar la compatibilidad esperada.

A mi madre Carmen,

esta alegría que me faltó entregarle en vida.

AGRADECIMIENTOS.

A mi familia, a mi esposa Consuelo por su persistente apoyo, a mis hijos Tomás, Trinidad, Magdalena y Leonor, que son nuestras alegrías diarias, y a la Sra. Lucy por cuidarlos con cariño, lo que nos permite abocarnos a nuestros horizontes con tranquilidad.

A mi madre y a mi padre, Carmen y Manolo, por su diaria dedicación, quienes desde su humildad ejemplificaron el camino de valores a seguir. A mis tías Alicia y Cristina, que nos guiaron a mi hermano y a mí, en todo momento de nuestro crecimiento educacional; y también a mi hermano Toño, que mostró el camino de esfuerzo para obtener logros.

A mis amigos Yerko, por traerme del mundo laboral a este proyecto, y Julio Salas, quien ha sido un verdadero guía en todo este camino de la computación y un cable a tierra desde los tiempos de Liceo, pasando por la FCFM, luego DCC, y finalmente en lo laboral.

Al profesor Juan Velásquez S. por darme esta tremenda oportunidad, dado el tiempo que llevaba alejado de la Universidad, y al Departamento de Ingeniería Industrial por albergar el desarrollo de las tecnologías de Inteligencia de Negocios que es a lo que me dedico laboralmente hace varios años.

Finalmente, a todos quienes contribuyeron al desarrollo de este trabajo, y a quienes ayudaron a sacar adelante mi vida universitaria, ya lejana en el tiempo.

Manuel Alejandro Castro Astete

TABLA DE CONTENIDO

Capítulo 1.- INTRODUCCIÓN

Si consideramos un sitio web que presente una variedad de productos y/o servicios para la venta, de una determinada compañía o empresa, en el cual los visitantes se identifican y compran en línea, entonces lo más probable es que se esté estudiando el comportamiento de compra de los usuarios a partir de esa información, buscando con ello ampliar la oferta de productos y/o servicios al máximo posible. A esto lo podemos identificar como el análisis basado en el consumo, que tiene como particular característica que es un estudio posterior a la compra, y que por cierto se puede combinar con otros análisis internos del negocio.

Para llevar adelante estos sitios web de buena manera, las compañías o empresas deben incurrir en costos más bien elevados en infraestructura, asegurar alta disponibilidad, y seguridad digital; es por ello que para instituciones de menor envergadura no es posible disponer de estos canales de venta, pero sí es considerado importante para la supervivencia del negocio, al menos, publicar su catálogo de productos en internet.

Para ambos casos descritos, el sitio web resulta ser una vitrina virtual, en donde las empresas exponen sus productos y/o servicios, y en donde los usuarios, como clientes potenciales, comparan y eligen el producto más aproximado a sus intereses (los que no son conocidos realmente), para finalmente decidir la compra y posteriormente concretarla de forma digital o presencial según sea el caso. El análisis de este comportamiento, anterior a la compra, resulta igualmente interesante desde el punto de vista del negocio para empresas de cualquier envergadura, ya que se busca como foco central descubrir o deducir los puntos de interés que mueven al potencial cliente a decidir la compra de un determinado producto, pretendiendo a partir de ello mejorar el despliegue de la oferta de productos.

En particular, lo que se describe en el presente trabajo, es el caso de una industria específica, de una región de Chile, en donde las empresas oferentes de productos y/o servicios son muy distintas entre sí, de tamaño muy diverso, y que participan de la

muestra de sus productos en un sitio web determinado, el que además cuenta con un interés de las autoridades regionales que busca potenciar dicha industria y regularla positivamente para promover su desarrollo.

Otro elemento que se consideró en el presente trabajo, es la interacción que se produce con un módulo de recomendación de productos en un sitio web; si bien existen algoritmos incluidos en los módulos de recomendación, era interesante determinar si con información más precisa se podía mejorar esa función y con ello perfeccionar la oferta del sitio web.

Parte de los esfuerzos en investigación que se realizan actualmente para conocer los elementos relevantes para la decisión de compra, de un producto que se muestra en sitios web antes descritos, es considerar y estudiar las opiniones que vierten sobre ellos en la web. Esta tarea no es sencilla, ya que la información que entrega un usuario de internet no tiene la estructura lógica que uno desearía para poder estudiarla, evaluarla, y aplicarla en la industria respectiva, sino que requiere de varios pasos más complejos; uno de ellos es la extracción remota de un sitio web que considere esas opiniones (eventualmente se puede contar con un módulo local de opiniones de clientes), luego de ello debe procesarse con algunos algoritmos de evaluación que distinguen si el comentario es positivo o negativo en relación al producto y/o servicio, para finalmente ponderarlo con un valor escalar según las condiciones del comentario.

Considerando todos los aspectos ya descritos, se planteó el desafío de averiguar si se podrían potenciar dichos aspectos al relacionarlos entre sí. Para ello se requirió de un proyecto que conjugara dichos componentes; fue así que se llegó al proyecto W.H.A.L.E., el cual dispone de un sitio web para la oferta de productos y/o servicios del rubro del turismo, específicamente de la X Región de Los Lagos; dicho proyecto contó con el rol patrocinador del SERNATUR y del Gobierno Regional de la X Región.

Finalmente, la propuesta llevada a cabo en el presente trabajo trata de la elaboración de un modelo de datos, y creación de la respectiva base de datos, que albergue y combine correctamente, a través de los datos, los distintos aspectos de

estudio de un sitio web; esto implicó establecer un flujo de información por la vía del transporte de datos entre las distintas bases de datos comprometidas en los componentes antes descritos. Dicha combinación de información se plasmó en la construcción de reportes orientados al análisis de los expertos en el rubro, y autoridades regionales del turismo, que permitieran corregir la oferta publicada de productos en el sitio web por los empresarios de turismo de la X Región.

1.1 Motivación.

Dado el amplio desarrollo de las técnicas de inteligencia de negocio, y del desarrollo de los Data Warehouse como repositorios centrales de datos, y dado que existen trabajos anteriores tanto del análisis de comportamiento de los usuarios web en un sitio determinado como de trabajos en investigación de opiniones en la web, además del desempeño de un recomendador de productos, era razonable preguntarse cómo la sinergia de todas las iniciativas en lo global podría entregar un conocimiento mayor a los resultados esperados de cada una por separado.

El presente desarrollo se enmarcó en la iniciativa W.H.A.L.E., que pretende apoyar a la industria de turismo de la X Región, a través de la construcción de una plataforma tecnológica web ampliada. Esta oportunidad nació de la necesidad de las autoridades y empresarios del turismo de esa región, de conocer aún mejor las características de la demanda de su industria, para así poder gestionar de manera más efectiva la oferta de productos o servicios de la región vía web.

En particular, el proyecto W.H.A.L.E. contiene trabajos de memoristas anteriores que desarrollaron componentes que aportan al conocimiento de la demanda de productos y servicios de la industria del turismo de la X Región. Lo interesante, es que este modelo propuesto para el caso específico, es una aplicación empírica y acotada de un concepto aplicable a cualquier otra industria que oferte por vía web sus productos.

1.2 Objetivos.

1.2.1 Objetivo General.

- Diseñar, construir e implementar un repositorio central de la información, centrado en la actividad de los usuarios de un sitio web determinado, y complementarlo con opiniones vertidas en la web e información oficial relativa a la industria del turismo de la X Región.

1.2.2 Objetivos Específicos.

- A partir de la información generada a través del cruce de datos de navegación en el sitio web, más los indicadores propios de la gestión del negocios del turismo, los tomadores de decisión podrán elegir más y mejores estrategias para el desarrollo de la actividad en la X Región.

1.3 Contribución de la memoria.

El aporte que se desprende del presente trabajo es el desarrollo empírico de la interacción que se alcanza a lograr entre los diferentes componentes involucrados: productos y/o servicios ofrecidos en el sitio web, el análisis del log de visitas, y el análisis de opiniones en internet respecto del mismo tema, no así con el módulo de recomendación de productos; a través de interfaces de traspaso de datos y un modelo central que aglomera los distintos componentes mencionados, con un set de reportes asociados.

Mirado desde el punto de vista genérico, se comprueba la compatibilidad que se puede llegar a lograr entre los módulos antes descritos construidos para

una industria específica, pero aplicable a cualquier otro sitio con características similares, sin embargo la experiencia muestra las consideraciones que debe tenerse al incorporar éstos módulos a un proyecto de este tipo para que puedan complementarse efectivamente y no quedar aislados entre sí.

1.4 Estructura de la memoria.

Este documento se distribuye de manera de presentar en forma gradual los temas tratados, comenzando con los conceptos más globales que se utilizan, para pasar a otros puntos más específicos y de interés puntual para la presente memoria, luego se pasa a describir el proyecto que posibilita este desarrollo y finalmente ver la aplicación empírica de conceptos tratados. Así la distribución de capítulos se presenta de la siguiente manera:

- En el Capítulo 2 se presentan los conceptos de Data Warehouse, Web Logs, ETL, Data Mining y Web Mining, y la utilización progresiva de estas técnicas por parte de las empresas o industrias.
- El Capítulo 3 muestra el proyecto que acoge el presente trabajo y permite la realización de esta memoria, el proyecto W.H.A.L.E., las características con las que cuenta y las particularidades que acotan el desarrollo. Además, se presentan los trabajos ya realizados por otros memoristas basados en este mismo proyecto y que forman parte de los módulos ya concluidos al inicio de este desarrollo. (WOM y AWS).
- Con el Capítulo 4 se comprende lo que se estableció como requerimientos y los alcances del desarrollo, tanto para el proyecto como sitio web del turismo como para la presente memoria como investigación.

- En el Capítulo 5 se muestra la implementación propiamente tal, desde los modelos propuesto, los modelos finalmente construidos, y lo que no se pudo incorporar, el análisis y estructuración de las interfaces de traspaso de datos (ETLs), y las interfaces de salida o visualización de resultados.
- Finalmente el Capítulo 6 presenta las conclusiones del presente trabajo, tras comprobar las interacciones esperadas y las que no pudieron realizarse.

1.5 Metodología Utilizada.

Considerando la complejidad en la implementación de las interfaces que relacionan las distintas fuentes de datos, la metodología estuvo determinada por:

- Levantamiento del Estado del Arte respecto al análisis de Logs en sitios WEB, las técnicas utilizadas, y evaluación del impacto que aporte al objetivo de análisis.
- Elaboración, construcción y puesta en marcha del Data Warehouse del proyecto WHALE.
- Levantamiento de información pública disponible para indicadores de apoyo a la industria del turismo de la X Región. Conclusiones de viabilidad.
- Modelamiento de Data Webhouse genérico unitario para el proyecto WHALE, basado en el suministro de información para la generación de los indicadores de apoyo a la industria.
- Levantamiento del modelo de datos utilizados por los sistemas WOM y AWS.
- Estudio de compatibilidad de información entre los sistemas WOW, AWS y el Data Warehouse de proyecto WHALE.
- Elaboración, construcción y puesta en marcha de procesos de traspaso de datos desde y hacia el Data Warehouse contra los sistemas WOW y AWS, generando un modelo aumentado.
- Elaboración y construcción de reportes para el análisis de la información conglomerada, que apunten al mejoramiento de las estrategias de publicación en el sitio web.
- Elaboración y construcción de reportes para estudio de comportamiento web de usuarios-visitantes del sitio con información aumentada.
- Análisis de la información conglomerada, y elaboración de propuestas para el mejoramiento de las estrategias de publicación en el sitio web.
- Elaboración de reportes orientados al perfil analista global de la industria.
- Observación del proceso conjunto y conclusiones.

Capítulo 2.- MARCO CONCEPTUAL.

2.1 Importancia de la Información Valiosa.

Es sabido, que la información obtenidas por las empresas respecto del mercado, puede darles ventajas comparativas que les permitan aumentar las ventas; sin embargo, se debe distinguir la información valiosa de la que no lo es, para ello estas empresas se apoyan en analistas expertos que conocen el rubro en cuestión y se espera que reconozcan los resultados importantes que arrojen los análisis de los datos.

Dentro de esta búsqueda por conocer los intereses del mercado existe gran diferencia en el estudio de los datos internos, de los externos; los primeros son más certeros y se obtienen del propio proceso de venta de la empresa, pero la información obtenida de ellos es más bien limitada, por lo que para poder inferir información más completa los datos se adjuntan con información de las ventas de otras líneas de productos, de manera de inferir los intereses de los clientes y poder establecer algoritmos predictivos respecto de futuros procesos de venta con valor estratégico.

Por otra parte, los datos externos no tienen una estructura determinada y el resultado de su análisis depende de la correcta interpretación de los analistas expertos, pero tienen una relación más directa con los intereses de los potenciales clientes; de hecho en la actualidad esa información se obtiene a través de encuestas que estructuran previamente la información que van a obtener y luego la cuantifica. Sin embargo, el alto costo, y el acotado alcance de estos métodos están empujando el desarrollo de nuevas líneas de investigación, una de ellas, extraer información de las opiniones vertidas en la Web.

Entre ambos grupos, datos externos e internos, se ubican los relativos al comportamiento del usuario dentro del sitio web, ya que antes de que se produzca la compra el usuario web observa la variedad de productos y según sus intereses (no conocidos) decide; ésto se registra en el archivo de movimientos del sitio web (WEB LOG), y es un comportamiento externo con datos internos estructurados.

2.2 Repositorios de Información y Data Warehouse.

Los repositorios de información nacieron como sistemas que aprovechaban los datos de distintos sistemas operacionales con el objetivo de ayudar, con información confiable y oportuna, al proceso de toma de decisiones, estos datos de origen interno cuentan con estructuras establecidas por la propia de la compañía, por lo que la captura, traspaso y transformación de esos datos cuenta con el absoluto control de las áreas de desarrollo y las planas directivas.

Esta necesidad de almacenamiento conjunto, que comienza en la década de los 90', enfrentó a los analistas internos a múltiples problemas derivados de las propias bases de datos operacionales, entre estos problemas se encontraban, que los datos presentaban gran heterogeneidad desde las distintas fuentes de procedencia, que las bases de datos se encontraban diseñadas para consultas cortas y predecibles, y por último, que resultaba costoso agregar datos desde distintas tablas relacionales[1]. Ante dichos problemas, se originó la arquitectura Data Warehousing, entendida como un conjunto de tecnologías que tienen por finalidad permitir al analista tomar mejores decisiones en un lapso de tiempo menor, algunas de las ventajas de esta arquitectura se relacionan con mejora en el rendimiento, una mejor calidad de los datos y la posibilidad de consolidar y resumir datos desde distintos sistemas[2].

Una de las principales funcionalidades de un Data Warehouse se identifica con actuar como fuente de datos, tanto para el Analytical Online Processing (OLAP), como también para aplicar técnicas de Data Mining[3]. Así, la consolidación de datos operacionales de interés permite a las empresas obtener indicadores sobre predeterminadas consultas, como también encontrar patrones sobre los datos.

Dentro de sus características principales de los modelos relacionales, se encuentran sus variadas formas de evitar la redundancia entre las relaciones y la posibilidad de asegurar la consistencia en la base de datos. Asimismo, permite responder, prácticamente, cualquier tipo de consulta; sin embargo, con el paso del tiempo y con el incremento de la potencia del Hardware han surgido alternativas al modelo relacional,

que simplifican en varios aspectos la escritura y lectura de datos, pero que no fueron considerados para el presente trabajo, no descartando su uso para futuras variaciones del mismo.

En las organizaciones es común que en distintas áreas ligadas a la toma de decisiones se encuentren problemas relacionados con el manejo de grandes volúmenes de datos, puesto que la generación de cualquier reporte sustentado en bases de datos operacionales consume una gran cantidad de tiempo, y los reportes muchas veces cuentan con datos o información que no resulta útil para la toma de decisiones[1].

Considerando los problemas universales de las organizaciones derivados de obtener información a partir de datos operacionales, en Spiliopoulou y otros se propone que un Data Warehouse:

- Debe hacer fácilmente accesible la información de la organización (debe ser entendible por todos dependiendo de su relación con el sistema)
- Debe presentar información consistente de la organización.
- Debe ser adaptativo y resistente al cambio, pues los requerimientos van cambiando con el tiempo.
- Debe proteger de forma segura la información.
- Debe servir como fundamento en la toma de decisiones.
- Debe ser aceptado por la comunidad para que sea exitoso, ya que es una herramienta adaptativa.

2.2.1 Modelamiento Relacional de Datos.

Dentro de las posibilidades de modelamiento relacional que se ocupa en el DataWareHouse existe el modelamiento multidimensional, que privilegia su orientación hacia las consultas de rápida respuesta. Se desarrollan estructuras para consultar dichos datos y extraer información valiosa de ella, una de las alternativas es utilizar un modelo de datos llamado “modelo estrella”, que cuenta con información distribuida en “dimensiones” y una tabla de “hechos”, la que contiene los datos numéricos denominados “medida” que se van a sumarizar, y agrupar, según sea el análisis.

El Modelo Multidimensional de Datos (MDM) provee una vía para realizar consultas que agreguen datos desde distintas dimensiones de manera más eficiente que el modelo relacional. En concreto, es una técnica que conceptualiza los modelos de datos como un conjunto de medidas que describen los aspectos comunes de la empresa[1]. El MDM es útil a la hora de resumir y ordenar los datos para apoyar el análisis de información. Para lograr esto, se focaliza en datos numéricos (como el valor de un elemento) o en funciones numéricas (como contar elementos, el peso porcentual, la suma de los valores, entre otros).

Esta información debe ser preparada para responder un cierto ámbito de preguntas y generalmente se almacena en una base de datos separada, llamada DATAMART, eventualmente se pueden preparar más de un conjunto de datos estructurados de manera similar.

Dicha estructura posibilita la generación de los llamados cubos OLAP, que apuntan a responder esas consultas de forma casi instantánea; para ello requieren un soporte tecnológico extra por lo que generalmente se disponibilizan en un servidor especial para ese fin.

El modelo estrella y de los cubos OLAP responden un set de consultas que pueden ser hechas directamente, o a través de herramientas de software intermediarias que le dan un valor agregado al resultado, y que generalmente cuentan con una importante

capacidad gráfica que permite visualizar de mejor forma los resultados y potenciar el análisis de los expertos.

En particular, los cubos corresponden a una implementación del MDM en bases de datos multidimensionales (MDBMS), creadas para este enfoque. Su nombre se debe a que su diseño se puede interpretar como un cubo de n-dimensiones, lo que permite hacer consultas eficientes a través de sus caras[1], comparando por ejemplo sucursales, productos y años de las ventas, que es la medida a tomar como valores de cantidad. En este nivel del cubo, se pueden tomar indicadores como la cantidad de ventas totales de todos los productos en cierta sucursal, ventas totales de la empresa entre dos años distintos, etc.

2.3 Data Mining.

Para encontrar información valiosa entre los datos disponibles se utilizan técnicas y softwares especializados en dicha búsqueda, estos procesos que llamamos Data Mining requieren siempre del apoyo experto para distinguir la información valiosa de la que no lo es. Generalmente se aplican a los datos internos almacenados en los Data Warehouse, eventualmente ya procesados en un primer nivel, y separados en una estructura de almacenamiento dirigida hacia un estudio en particular llamada Data Mart.

Parte del análisis de los datos se lleva a cabo con herramientas de minería de datos, cuyo fin es encontrar información que no es evidente al observar directamente los datos, con ello pretende relacionar comportamientos pasados con posibles escenarios futuros y eventualmente modificarlos para mejorar las oportunidades competitivas.

2.4 Procesos ETL.

Al diseñarse un sistema de información que reúne datos desde distintas fuentes, es necesario utilizar técnicas que extraigan los datos, los transformen a formatos específicos y útiles, y los carguen en la base de datos. Este conjunto de técnicas es conocido como el proceso ETL (Extract, Transform and Load)[1].

Para ello existen una gama bastante amplia de tecnologías y un modelo de procesamiento de la información muy bien madurado al interior de las organizaciones y por las empresas especializadas en llevarlos a cabo; éstas consideran distintos orígenes de la información, con periodicidades distintas, y para los cuales se utilizan procesos de carga específicos, individuales e independientes; además se provee de un ambiente o Base de Datos especial para el ingreso de esos datos, llamada Staging Area.

A partir de estos datos se alimenta el Data Warehouse corporativo, que tiene un diseño de modelo de datos especial, abierto a todas las características relevantes del negocio, pero que descarta detalles que no aporten al objetivo de estudio o análisis posterior.

2.4.1 Extracción.

Consiste en extraer datos operacionales desde distintas fuentes de la empresa, como de sus sistemas informáticos, archivos de datos, o bases de datos relacionales, entre otros. Durante su diseño, es necesario examinar cómo se realizará en el tiempo la extracción de datos, considerando escenarios cambiantes, como pueden ser nuevas fuentes o formatos.

2.4.2 Transformación.

Esta etapa consiste en transformar los datos extraídos a valores y formatos específicos. Para esto, primero se debe realizar su estandarización, que consiste en llevarlos a la misma unidad de medida. Luego, se deben generar los valores que serán almacenados en el repositorio, que representarán las distintas jerarquías dentro de cada dimensión. Estos dos complejos pasos conllevan a que esta etapa del proceso ETL sea la que tarda más tiempo en desarrollarse.

Para llevar a cabo esta etapa existen 3 alternativas:

- Programar algoritmos en los lenguajes de cada fuente de datos.
- Utilizar herramientas de pago.
- Utilizar una base de datos transitoria para posteriormente utilizar un único lenguaje para la transformación.

Debido a la complejidad de la primera alternativa, y el costo de la segunda, la base de datos transitoria resulta ser la más utilizada. Este repositorio es la DSA (Data Staging Area), que sirve como prueba de un DW completo, de manera de entender el nivel de complejidad futuro y el estado actual de los datos a transformar.

Dentro de este punto, se realiza un proceso de “limpieza” de datos, el que apunta a descartar los valores incorrectos provocados por anomalías en la toma del dato, también se establecen valores “por defecto” en el caso de ausencia de información; lo anterior pasa (como todo el proceso de ETL) por la voz del experto en el negocio respectivo, quien es el llamado a calificar un dato como reemplazable o descartable, teniendo conciencia de la alteración en la calidad de los datos, y del posterior resultado de los análisis.

2.4.3 Carga de Datos.

Consiste en cargar los datos de la DSA en el repositorio final, utilizando el lenguaje SQL de base de datos; dependiendo de los ambientes en los que se encuentran los datos se hace necesario un apoyo de herramientas especializadas, que encapsulan las instrucciones SQL creadas para la carga de datos.

2.5 Información del Negocio con Datos Externos.

Actualmente, los esfuerzos en investigación se han centrado en el análisis de la información desplegada en internet, información que no tiene una estructura definida y que depende exclusivamente de cada fuente de información. La gran apuesta para estos estudios, es que esta información dé valor nuevo a lo que ya se conoce del mercado, de manera de ponerse un paso al frente en el conocimiento del comportamiento del mercado, y de los potenciales clientes, así como de la propia industria a la que pertenece la empresa.

2.6 Web Mining.

Así como se considera relevante el estudio de los datos externos, es también como se han ido generando estrategias de búsqueda de información, extracción de la misma, recopilación de esos datos y el estudio finalmente de ellos, para reconocer información valiosa para el negocio lógicamente. Así, se da paso a la disciplina del Web Mining, que corresponde a la adaptación del Data Mining sobre los datos esparcidos en la Web, con variaciones con respecto al proceso en sí debido a su falta de estructura.

2.7 Web Logs.

El análisis de un sitio web también puede ser realizado con la finalidad de extraer conocimiento desde los datos emanados de la navegación de los usuarios y el contenido de las páginas, que permiten extraer patrones acerca del comportamiento en la navegación del usuario. Dicho análisis comienza con el registro de la actividad en el sitio web, que tiene un formato definido, y que más tarde se procesa para llevarlo a una estructura preparada para su análisis, y que se almacena en alguna base de datos especial o dentro del Data Warehouse como un sistema operacional más.

La generación del Log, en estricto rigor es al Access Log o Registro de Acceso, se configura por parte del administrador del sitio web y consta de un archivo plano de texto que contiene en cada línea la actividad solicitada hacia el sitio web de parte del visitante, y este archivo se dispone internamente a diario para la carga al modelo de manera automática. La estructura del archivo consta de la siguiente información:

- IP del cliente (host remoto) que hizo la petición al servidor, aunque si existe un proxy entre el usuario final y el servidor, se registrará la dirección del proxy.
- "-", un guión por defecto, corresponde a información de identidad que se usa para controles estrictos en redes internas.
- Usuario, es el nombre de usuario con el que se solicita el documento, si no está protegido con contraseña entonces se mostrará un guión, "-".
- Hora completa de la petición. El formato es:

  [día/mes/año:hora_minuto_segundo zona_horaria].
- Línea de petición, se muestra entre comillas dobles, y contiene el método utilizado, el recurso solicitado, y el protocolo usado.

  "GET /apache_pb.gif HTTP/1.0"
- Código de estado, el cual envía el servidor al cliente. Ej: 200, éxito.

- Tamaño del objeto retornado.
- La ruta completa desde donde se origina la petición, o la página que contiene el enlace.
- Navegador del cliente, versión, sistema operativo. Es la información que identifica al navegador y la información extra que incluye de sí mismo.

2.8 Web Opinion Mining.

Por otra parte, existen otras dos líneas de investigación en pleno desarrollo a nivel mundial y que actualmente concitan el mayor interés: Web Opinion Mining (W.O.M.) y Sentiment Analysis (S.A.) [4], las que permiten extraer información valiosa desde comentarios y opiniones que consignan diversos usuarios de blogs, wikis y páginas personales.

En particular el WOM tiene como objetivo extraer, resumir y descubrir varios aspectos subjetivos insertos en documentos y textos existentes en la web, lo que permite apreciar puntos de vista y opiniones relevantes con diversos grados de profundidad [5].

Por otro lado el SA utiliza la aplicación de técnicas específicas de Procesamiento de Lenguaje Natural (ó NLP, del inglés Natural Languaje Processing) enfocadas en detectar ciertos adjetivos o sustantivos en frases, textos o documentos, los cuales indican qué opinión o qué sentimiento está expresando el autor sobre algún tópico específico. Cabe mencionar que la determinación de qué características se están analizando en general, se realiza automáticamente a partir del documento, siempre que éste se encuentre contextualizado.[6]

El Web Opinion Mining (WOM) es una técnica de minería de datos basada en la extracción de conocimiento desde opiniones vertidas en la web. Según Bing Liu [8], la metodología principal para obtener información de numerosas opiniones se relaciona, primero, con reconocer las entidades involucradas, ordenarlas, reconocer los aspectos

a tomar en cuenta (tanto explícitos como implícitos), para luego asociarlos a las entidades. El paso siguiente es efectuar un “Sentiment Analysis” que se basa en cuantificar el grado de ánimo de las opiniones. Hay numerosos algoritmos que cumplen esta labor, basados, en su mayoría, en el concepto de Machine Learning (algoritmos adaptativos de aprendizaje). Asimismo, se realiza una sumarización basada en la visualización de datos o una caracterización mediante texto. Por último, cabe señalar que se le debe dar un enfoque, entre los cuales se establece la orientación, la entidad, el aspecto o el tiempo.

2.9 Adaptive Web Site.

Un AWS corresponde a la nueva generación de portales que se están investigando con el objetivo de hacer frente a los exigentes requerimientos de los usuarios, los cuales son variables en el tiempo. A modo de definición, un Adaptive Web Site (AWS) es un sitio capaz de adaptar su contenido en línea basándose en el comportamiento de navegación y las preferencias de los usuarios, lo cual es capaz de realizar gracias a la construcción de un módulo de “recomendación” de alta complejidad, el cual es lo más complicado de realizar dentro del AWS.[2]

2.10 Data WareHouse y WEB LOGs, Data WebHouse.

Actualmente el Data Warehouse es la forma más difundida en la que las compañías almacenan su información corporativa, allí se almacenan grandes volúmenes de información provenientes de los distintos sistemas operacionales de la propia compañía con el fin de descubrir en ella información valiosa, que se utilice como apoyo para la toma de decisiones para el mediano y largo plazo.

Si consideramos la actividad de navegación de los usuarios web en el sitio en cuestión, tenemos una gran cantidad de registros cuyos datos podrían indicar alguna secuencia que lleve al usuario a hacer su elección, por lo que se cree que si podemos determinar la mejor secuencia, se puede inducir la compra más eficientemente. Toda la información de la actividad en el sitio web se registra en el LOG de la página web, cuyos datos son incorporados al DataWarehouse que contiene información propia de la compañía, por lo que este concepto es más amplio y se le llamó DataWebHouse.

El paquete de procedimientos, base metodológica, herramientas de software, es lo que llamamos Business Intelligence, y opera en el formato descrito evolutivamente desde fines de los 80´s; sin embargo el desarrollo de la propia industria del Business Intelligence se ha visto favorecido por el aumento de potencia del hardware disponible, que permite actualmente un mayor volumen de procesamiento en un menor tiempo, lo que implica un avance que podría influir positivamente en el proyecto.

A partir de estos conceptos, se hicieron avances al aplicarlos para el estudio del comportamiento de los visitantes de un sitio web determinado, dado el gran volumen de información que se registra de ellos cuando navegan el sitio web. A ésto se le llamó WEB INTELLIGENCE, concepto que fue ampliándose más tarde al incorporar otras posibilidades de estudio del comportamiento de las personas en la web.[7]

Cuando un usuario realiza acciones en un sitio web, se puede registrar su secuencia de navegación en archivos de texto conocidos como “web logs” [8, 9, 10], con los

cuales se puede realizar un análisis posterior para mejorar la estructura y el contenido del sitio [11].

De modo general, los web data (ó datos originados en la WEB) entregan información valiosa sobre un sitio en particular y los usuarios que lo visitan [9, 10, 12, 13, 14, 15]. Al aplicar técnicas de web mining sobre los web data, se puede conocer cuál es el contenido preferido por los usuarios de un sitio, cómo acceden a la información que les interesa y cuánto les ayuda la estructura del sitio a encontrar lo que buscan.[16]

El análisis de un sitio web permite predecir las necesidades de los usuarios, evaluar el impacto del portal, y guiar las mejoras, tanto en su contenido, diseño y estructura. Esto suele ser desarrollado a través del análisis de los clickstreams, los que son transformados en una gran cantidad de datos almacenados en los archivos weblogs, y que permiten conocer qué es lo que visitó el usuario en cada sesión, información que es utilizada por técnicas ligadas al Data Warehousing y al Web Mining, con la finalidad de encontrar nuevo conocimiento valioso sobre el comportamiento de navegación del usuario.[3]

Se espera que la secuencia de clicks permita identificar sesiones exitosas o fallidas en el sitio estudiado, determinar visitantes exitosos y buenos prospectos, y mostrar las partes del sitio web que son efectivos en atraer y retener visitantes. [1]

Podemos llevar flujo de clicks como fuente de datos hacia una estructura de datamart para analizar tal como alguna otra fuente de datos interna de nuestro negocio. Por supuesto, al tiempo que levantamos este datamart debemos tener cuidado de enganchar las dimensiones y hechos compatibles con los datos del resto de la empresa. Hecho esto, el flujo de click significará una aporte en el Webhouse distribuido global de la empresa. [1]

2.11 Reportes e Indicadores.

Finalmente, y luego del procesamiento de los datos, se requiere de herramientas para mostrar los resultados con la información más representativa; es así que se utilizan herramientas de reportabilidad que ayudan en la visualización de la información, y permiten tener una mirada más simple de los resultados.

Para ello, existen softwares que se acoplan con los recipientes de información, y sobre los cuales se dejan establecidas las consultas a visualizar, estas herramientas tienen un formato web para mayor simpleza, y tienen medidas de seguridad incorporadas para evitar la filtración de información estratégica.

Los indicadores que se establezcan a partir de los requerimientos de los expertos caen en la categoría de la reportabilidad, y su definición de cálculo es responsabilidad de los involucrados en el proyecto.

2.12 Supervisión de Expertos y Stakeholders.

Los conceptos anteriores requieren una fuerte participación de los expertos del negocio y de los involucrados en las decisiones de más alto nivel o estratégicas; sin esta participación fundamental el resultado será incierto.

A menudo se compromete un esfuerzo de la plana más alta, sin embargo esa asistencia es necesaria durante todo el desarrollo del proyecto.

2.13 Oportunidad de Aplicación.

La caracterización de la demanda, la configuración y promoción de la oferta, y la generación de indicadores orientados a mejorar la gestión y la toma de decisiones, son focos potenciales de desarrollo en la industria del turismo debido al bajo grado de

implementación tecnológica que presentan los agentes turísticos, especialmente en la Región de los Lagos. [17]

Considerando el nivel de ingresos (no permite investigación de mercado tradicional), y la baja asociatividad de quienes componen el sector, se visualiza la enorme oportunidad de explotar otros medios de captura de información fidedigna que permitan a los agentes turísticos de la región de Los Lagos (extensible) caracterizar la demanda de turistas (principalmente extranjeros) y construir una oferta de productos y servicios con mayor valor agregado que el actual, permitiendo atraer un mayor número de turistas que busquen destinos con características similares al sur de Chile y aumentar el gasto por turista una vez que se encuentren en nuestro país.[6].

Ante este escenario, el desarrollo de técnicas avanzadas de WEB Intelligence [7], y su integración a una plataforma de acceso rápido y seguro, surge como una potente opción que permitiría aprovechar esta oportunidad.

En la actualidad, la web se ha transformado en una base de información donde confluyen diversos contenidos y temáticas que influyen en un sinfín de industrias, siendo la del turismo una de las más beneficiadas a nivel mundial. (En efecto, según el trabajo [18],, la WEB representa una de las principales fuentes de información al momento de evaluar la oferta turística mundial).

En contraste con lo anterior, a nivel nacional aún existe un bajo uso de la web por parte de las empresas de turismo, incluso para tareas básicas como la publicación de información relacionada con su oferta turística. Sin embargo, algunas localidades del país que han construido interesantes portales web (un ejemplo es el sitio Pisco Elqui: www.piscoelqui.cl) demuestran cómo el desarrollo de las TIC´s puede permitir a pequeños y medianos empresarios turísticos ofrecer sus productos en el mercado digital obteniendo resultados exitosos. Por lo tanto, si se considera que la mayoría de los sitios desarrollados para el turismo chileno aún no capturan los beneficios que presenta la WEB 2.0 y otras tendencias actuales, se presenta como una posibilidad interesante el explorar metodologías de punta, como la nueva generación de portales

adaptativos, capaces de personalizar la oferta según el comportamiento de navegación y preferencias de los usuarios[2].

Todos estos avances, análisis de navegación web, portales adaptativos, análisis de opiniones en la web, sumado al propio desarrollo de un modelo de datos que represente los intereses y las reglas del negocio, deben combinarse adecuadamente, y potencialmente mejorar el conocimiento que se tenga de los usuarios o visitantes de un sitio web en el que se ofrecen productos, de tal manera que se les haga una “mejor oferta”, en forma más precisa, con mayor claridad, y con gran influencia en el mejoramiento de las estrategias de venta.

FIG 1.- Módulos y compatibilidad

(Elaboración propia)

2.14 Desarrollos o Aplicaciones Similares.

Dado los componentes que se conjugan en el presente desarrollo, es que existen diversos avances en cada una de las líneas de investigación, de hecho se cuenta con herramientas comerciales y grandes compañías que prestan el servicio de análisis de opinión en internet como por ejemplo IBM con COGNOS Consumer Insight, apuntando más bien al posicionamiento de las marcas en los distintos mercados y a la apreciación de los consumidores de las mismas. También se pueden encontrar empresas de menor

envergadura que ofrecen realizar esos análisis, así como también el mismo sitio TRIPADVISOR cuenta con herramientas de generación de campañas de marketing por correo para oferentes del turismo, y con evaluación de las opiniones vertidas en su sitio con un panel de resumen. (www.tripadvisor.es/ReviewExpress).

Para el análisis del Log de visitas a un sitio web, también se pueden encontrar en internet distintas soluciones, incluso gratuitas, sin embargo hay variedades, ya que algunas apuntan a rastrear el comportamiento previo a la venta de los productos que se realiza a través del mismo sitio, y otros apuntan a la mejora de la estructura de la página web, incluso midiendo la efectividad de campañas de marketing presentadas por esa misma vía.

Tal como se mencionó anteriormente, la utilización de estas componentes supone un desarrollo informático interno de las compañías, o instituciones, es así como COGNOS ofrece la integración de dichos módulos, pero en un “alto nivel”, esto es a través de COGNOS BI que muestra la información cuando los datos ya fueron procesados por las herramientas de Business Intelligence (eventualmente permite integrar el análisis del web log).

Todo lo anterior apunta a que aunque existan herramientas individuales que cubran ciertos aspectos del análisis de las preferencias de los clientes, hay diversas líneas de desarrollo y un amplio conjunto de empresas muy interesado en potenciarse frente a la competencia; sin embargo, no se aprecia un desarrollo con las mismas características del presente trabajo, ya que el sitio Patagonia aglutina a toda una industria de una región específica, más que a una compañía en particular, y con información sólo de oferta de productos más que de las ventas en sí mismas, además la integración se hace a nivel de datos a través de los procesos ETL.

.

Capítulo 3.- PROYECTO W.H.A.L.E.

Para dar forma a esta iniciativa se necesitaba un proyecto que utilice todos los avances mencionados en una sola temática transversal; es en este contexto que se sitúa el proyecto W.H.A.L.E., (Web Hypermedia Analysis Latent Enviroment) Data Webhouse [1].

Para comprobar la combinación de iniciativas señaladas, y también hacer un aporte a una industria en particular, se elaboró el proyecto que tiene como uno de sus objetivos principales mejorar las estrategias de oferta de productos específicamente de turismo de la X Región, a través de un sitio web para ese fin. El proyecto al momento de intervenirlo con el presente desarrollo ya contaba con el sitio web construido, un módulo de Adaptive Web Sites (AWS) que analiza la estructura y distribución de información en el sitio web, y un módulo que revisa las opiniones en la web (Web Opinion Mining – WOM).

Los tipos de actores que trabajan dentro del sistema del portal son turistas, empresarios turísticos, operador turístico, SERNATUR, el GORE X, analistas del rubro del turismo, y un administrador tecnológico.

3.1 Introducción del Mercado de Turismo.

El mercado turístico a nivel internacional se desarrolla actualmente en un contexto de creciente competencia entre los destinos. En este escenario la generación de información, para la toma de decisiones, resulta esencial al momento de detectar las principales tendencias de la industria y así segmentar adecuadamente las estrategias de acuerdo a las características del mercado y las preferencias de los turistas.

En respuesta a este desafío, el proyecto WHALE busca contribuir a disponer información relevante para la industria del turismo a través de una plataforma web con el fin de gestionar de mejor manera la oferta y comprender las principales características de la demanda de la industria del turismo en la región [1]. Para esto se realizó un levantamiento de información a partir de fuentes secundarias, entre los cuales se incluyen documentos de Organismos internacionales, como la Organización Mundial del Turismo, y documentos de países que se consideran benchmark para el caso chileno como son España, México, Nueva Zelanda, EE.UU., entre otros. Adicionalmente se realizó una recopilación de información a partir de una ronda de reuniones con empresarios de la industria del turismo en la región de Los Lagos, a quienes se les consultó por la información e indicadores que consideraban más relevantes para la toma de decisiones en sus negocios y reuniones con representantes del SERNATUR regional y nacional, con quienes se discutió acerca de las principales tendencias en la elaboración de indicadores y aquellos indicadores que debían ser incluidos como parte de la plataforma.

Como resultado de este trabajo, se elaboró una lista larga de indicadores que sirvió para centrar la discusión en un taller con empresarios y representantes del sector público, con quienes se establecieron las prioridades.

En base a esta definición, el cálculo de indicadores tiene como objetivo analizar el estado del turismo en un periodo de tiempo. Para esto, los países deben disponer de la información necesaria, para lo cual se utilizan bases de datos difundidas por organismos nacionales o internacionales.

En la industria del turismo, la información para la construcción de indicadores que habitualmente se utilizan, se obtiene de las cuentas nacionales y de balances de pagos, instrumentos que con distinto grado de desarrollo y aplicación se encuentran generalmente disponibles en todos los países [3]. Como ejemplo de variables, la demanda es determinante para el nivel de producción y el empleo en dicho sector, por lo que un análisis detallado de su estructura y evolución es muy importante [3].

Una de las características más singulares y valiosas del proyecto WHALE es que los mismos empresarios participantes de este conglomerado eligieron, en un ejercicio progresivo, una lista de 42 indicadores, los que permitirían “ver” cómo se desarrolla la industria y mostrar la relación con otros factores e iniciativas en la región. Estos indicadores debieran ser construidos a partir de consultas directas al Data Webhouse.[6]

Un segmento importante en el proyecto es el estamento regulador del mercado, las autoridades participantes en esta iniciativa: GORE X Región y SERNATUR X Región, tienen un rol muy relevante, ya que a través de los análisis posteriores, que sean realizados por expertos conocedores del mercado del turismo en esa región, tengan elementos importantes para vislumbrar el desarrollo de la industria en esa región, y poder así elaborar mejoras en políticas públicas que afecten positivamente a dicha industria.

3.2 Sitio WEB del Proyecto W.H.A.L.E.

El sitio web del proyecto W.H.A.L.E. es ***www.patagonialoslagos.cl***. En esta página, los usuarios deben crear una cuenta, que le otorgará un rol dentro del portal. Las áreas más importantes del modelo son:

1) Usuarios: información básica del usuario dentro del portal. Entre sus atributos se encuentra un ID de usuario, información de *login* y *pass*. Los usuarios pueden crear comunidades dentro del mismo sitio.

2) Seguimiento al Usuario: a cada usuario cuentan con un historial de los links que tomó para cruzar eso con sus preferencias, dependiendo del seguimiento dado al usuario se le muestran distintas ofertas de productos que están totalmente personalizadas según sus preferencias.

3) Productos: Cada producto cuenta con la información completa acerca de su proveedor, ubicación y disponibilidad según época del año y clima. Los productos serán divididos por categorías, por lo que si se busca la categoría de turismo aventura, entonces se ofrecerán todos los productos correspondientes a ello. Cada producto tiene un área de comentarios donde los usuarios pueden escribir sus experiencias.

4) Entidades: los usuarios pueden ser de distinto tipo. Entre las entidades encontradas que pueden entrar al sitio se encuentran los turistas, los empresarios turísticos, los operadores turísticos, instituciones y analistas. Cada uno tiene distintos permisos dentro del portal, pues sus objetivos con él son diferentes. Los operadores y empresarios tienen la posibilidad de actualizar la información con respecto a los productos, los analistas pueden ver indicadores de acceso restringido respecto del sitio.

3.3 Indicadores del Turismo.

Indicadores para Sumar Valor a la Sustentabilidad de la Industria del Turismo de la X Región. Como parte del proyecto podemos contribuir a articular información disponible y ponerla a disposición de todos los actores del sector. En base a este trabajo de búsqueda y recopilación de información es que se articuló la siguiente lista de Indicadores.

Se puede concluir en base a la información recabada los siguientes puntos:

- Chile cuenta con los principales indicadores propuestos por la OMT para caracterizar la oferta y demanda de manera descriptiva.

- Los datos disponibles se construyen a partir de fuentes fuera del SERNATUR, lo que genera algunos problemas porque es necesario ajustar los resultados que originalmente buscan medir otras cosas para caracterizar la industria.
- Existe un desafío importante en la generación de indicadores a nivel regional, (no existen fuentes de información o bases de datos fácilmente generables, es necesario identificar nuevas fuentes que podrían servir).

El plan regional de turismo abre una oportunidad y genera una visión para la industria a partir del cual se pueden idear indicadores que podrían ser relevantes para la industria.

En Chile, el Servicio Nacional de Turismo y el Instituto Nacional de Estadísticas presentan en conjunto, cada año, un Informe Anual de Turismo. Este informe presenta una síntesis de la información anual sobre la composición básica de la oferta y distribución de los bienes y servicios, considerando variables como ingresos, compras, gastos y costos, empleo y uso de tecnología de la información en las empresas.

Las cifras turísticas mensuales de las cuales dispone el INE son: Número de llegada y pernoctación de pasajeros a establecimientos de alojamiento turístico, llegada en número de personas, pernoctación (N de noches) , pernoctación promedio.

Indicadores turísticos mensuales por tipo de establecimientos, (Hotel, Residencial, Motel, Apart Hotel, Camping: Sitios y Cabañas, Otros):

Variables de oferta: Capacidad de habitaciones y camas ofrecidas.

Variables de demanda: Noches ocupadas, llegadas y pernoctaciones de pasajeros, además de ingresos obtenidos por concepto de alojamiento turístico de establecimiento.

Las principales variables a encuestar periódicamente, en todos y cada uno de los establecimientos, corresponden a:

- Capacidad Ofrecida: Es la oferta normal de capacidad disponible por los establecimientos de alojamiento turístico.

- Personal Ocupado: Corresponde al número mensual de personas que trabajaron en el establecimiento, incluyendo una desagregación por género para cada categoría (hombre, mujer).

- Ocupación y Movimiento de Pasajeros: La ocupación corresponde a la suma de noches que se ocuparon habitaciones, incluyendo suites, departamentos y cabañas, por separado camping (sitios y cabañas), de acuerdo al tipo de establecimiento.

- El Movimiento de Pasajeros involucra dos variables: Llegadas y Pernoctaciones, en dónde las llegadas corresponden al número total de pasajeros (Chilenos y Extranjeros) llegados al establecimiento y que pernoctaron como mínimo una noche; por tanto, la pernoctación es el número total de noches “por cada una” de las personas que ocuparon las habitaciones.

A partir de lo anterior, se filtró lo más relevante en una lista corta de indicadores. Para su elaboración se programaron dos Talleres en la Región de Los Lagos: el primero, se realizó en Ancud el día 10 de Octubre de 2012, en el Hotel de Ancud (Panamericana Hoteles) el cual tuvo una asistencia de 15 participantes, el segundo Taller, se realizó en Puerto Varas el día 11 de Octubre de 2012, en el Hotel Colonos del Sur donde alcanzó una asistencia de 9 personas. La asistencia total, sumando ambos talleres fue de 24 participantes.

En esta etapa el equipo de levantamiento solamente llevó a unidades cuantificables la percepción, sin considerar otros aspectos tales como disponibilidad de la información o facilidad en el acceso de los datos, ya que ese análisis corresponde a instancias posteriores.

Tabla 4: Propuesta de Lista Corta de Indicadores

Tema/Categoría	*Cultural*	
Ámbitos	**N°**	**Nombre del Indicador**
Patrimonio Cultural	1	Número de destinos clasificados como patrimonio
Tema/Categoría	*Económico*	
Ámbitos	**N°**	**Nombre del Indicador**
Aporte regional/desempeño económico	2	Gasto medio de los turistas en la región
	3	Tasa de inversiones en el sector turismo respecto de inversión total
	4	Porcentaje del PIB, nacional, regional del sector
Caracterización de la demanda	5	Porcentaje de visitantes que regresan a la región
	6	Número de turistas que visitan la región por mes o trimestre (distribución durante el año).
	7	Proporción entre el número de turistas en periodos de baja afluencia respecto a los de afluencia máxima
	8	Llegada en número de personas (Total, De Chile y Extranjeros)
Caracterización de la oferta	9	Porcentaje de empresas certificadas con Sello de Calidad Turística
	10	Porcentaje de establecimientos abiertos todo el año.
	11	Tasa de ocupación para alojamientos oficiales por mes (temporada alta en relación con la temporada baja)
	12	Capacidad de habitaciones y camas ofrecidas
	13	Noches ocupadas
Oferta de variedad de experiencias	14	Índice de diversidad de la oferta de alojamiento turístico
	15	Calidad de la oferta de alojamiento turístico
	16	Oferta de actividades ecoturísticas
Patrimonio Cultural/natural	17	Número de rutas de acceso en buenas condiciones para el uso turístico de los espacios naturales
	18	Número de rutas turísticas nacionales que incluyen la X Región en su itinerario
	19	Número de guías expertos certificados
Productividad y	20	Porcentaje de empleados capacitados
Impuestos	21	Porcentaje pagado de impuestos en la región en relación al total nacional, con respecto a la industria del turismo
Satisfacción de los turistas	22	Satisfacción global del visitante
	23	Satisfacción del turista por la relación calidad-precio
	24	Duración de la estadía media
	25	Pernoctación (Nro. de noches)
	26	Pernoctación Promedio
	27	Fidelidad de la demanda
	28	Satisfacción de la visita de espacios naturales protegidos
	29	Satisfacción de la visita de los sitios culturales del destino

Transportes relacionados con el turismo	30	Dotación de vehículos de transporte de viajeros
	31	Tiempo de acceso al aeropuerto más cercano
Tema/Categoría	*Medio Ambiente*	
Ámbitos	**N°**	**Nombre del Indicador**
Protección de los ecosistemas	32	Superficie natural protegida
Gestión de residuos sólidos urbanos	33	Percepción de la limpieza del destino por parte del turista
Gestión del Agua	34	Ahorro de agua (% ahorrado, recuperado o reciclado).
Tema/Categoría	*Social*	
Ámbitos	**N°**	**Nombre del Indicador**
Infraestructura pública	35	Capacidad de servicios sanitarios
	36	Capacidad de servicios de transportes
	37	Mantención de luminarias
	38	Disponibilidad de señaléticas
	39	Mantención de aéreas verdes
Seguridad Pública	40	Valoración de la seguridad en el destino
	41	Disponibilidad de dispositivos de seguridad en las playas o lagos

Se puede inferir de los resultados que el tema Económico, es uno de los más seleccionados; al mismo tiempo cabe señalar que esta categoría es una de las que presenta mayor número de Indicadores propuestos. Dentro del tema Económico, es posible destacar la inclinación hacia ciertos ámbitos como: Patrimonio Cultural/Natural, Oferta de variedad de experiencias y Satisfacción de los Turistas que fueron escogidos casi en su totalidad.

Otra categoría que fue bien valorizada fue la Social, en donde sus dos ámbitos Infraestructura Pública y Seguridad Pública superaron promedios de rango de 2.

Es importante señalar que los ámbitos de caracterización de demanda y de oferta también fueron preferidos por los participantes, de los cuales a lo menos la mitad de sus indicadores quedaron seleccionados en la Lista Corta.

En tanto el tema de Medio Ambiente, no fue un tema menor. Su promedio siempre estuvo en los rangos de Medianamente Importante e Importante, donde los ámbitos de Gestión del Agua, Gestión de residuos sólidos urbanos y Protección de los ecosistemas fueron los que consiguieron las más altas valorizaciones dentro de su Tema. Sin embargo, Gestión Energética estuvo muy cerca de ser seleccionada en la lista corta con un promedio de 2,2.

3.4 WOM, Módulo de Opiniones en la Web.

Uno de los desarrollos de memoristas anteriores incluídos en este proyecto, es el referente al Web Opinion Mining, que se basó en información relativa al turismo de la X Región rescatada del sitio TripAdvisor.

Dentro de la información que hay en el sitio de Chile, se cuentan cuatro categorías principales:

- Listado de los principales Hoteles.
- Listado de Hostales.
- Listado de Restaurantes.
- Listado de actividades o tours que se pueden realizar.

Cada uno de los ítems dentro de una de estas categorías (hoteles, hostales, restaurants, atracciones) se encuentra asociado a una sub zona geográfica. Estas zonas geográficas corresponden a destinos populares y son las siguientes:

- Distrito del Lago.
- Isla de Pascua.
- La Patagonia.
- Santiago.
- Valle Central.

Como puede apreciarse, esta segmentación no tiene una relación directa con la administración política del territorio de nuestro país, sin embargo dentro de cada sub zona existe un detalle por ciudades y/o localidades que permite identificar aquellas que pertenezcan a la X Región.

El aspecto que hace interesante este sitio como fuente de información para caracterizar la demanda de productos turísticos de la Región de los Lagos, es el sistema de asesoramiento en que los viajeros (como usuarios registrados de la página), hacen recomendaciones a los visitantes (futuros viajeros). Este sistema se presenta como una plataforma de opiniones sobre los distintos hoteles, hostales, actividades o restaurantes (categorías). Estas opiniones están a disposición de todos los visitantes de la página y pueden encontrarse mayoritariamente (para el caso de Chile) en español, pero también las hay en inglés, portugués, francés y holandés. El idioma dependerá netamente del origen del usuario que haya escrito el mensaje. Como el sitio está disponible en varios idiomas, está también integrado un servicio de traducción de las opiniones desde otros idiomas al español. Cabe mencionar además, que al contenido mismo de la opinión de cada usuario, se añade una calificación que el mismo usuario hace usando una escala de puntaje de 6 niveles (de 0 a 5) y además existe la posibilidad de que un encargado del hotel, hostal, restaurant o actividad, escriba un mensaje de respuesta.

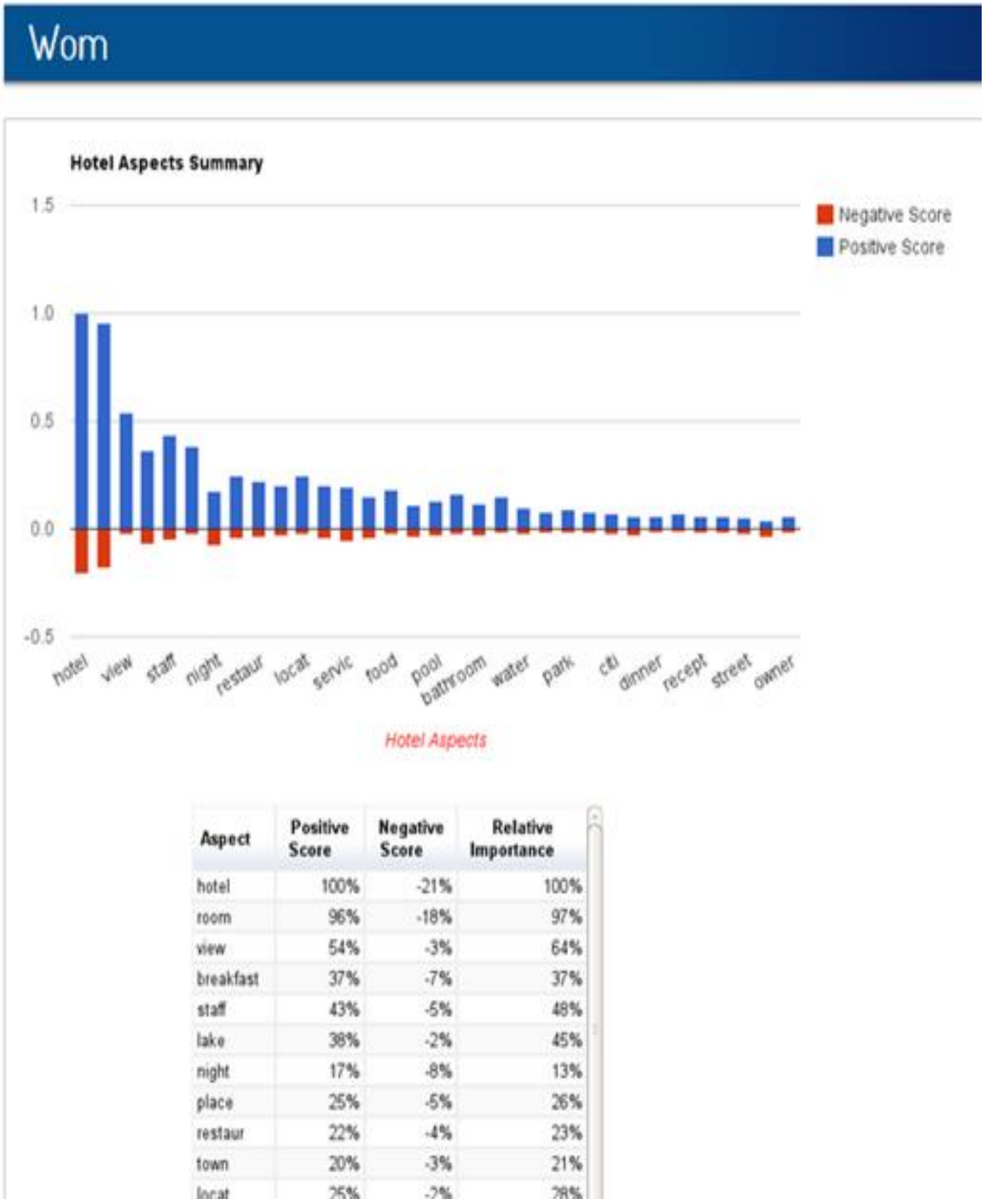
A continuación se presenta el detalle de la cantidad de información (ítems en cada categoría) que el sitio maneja para Chile:

Información	Número de Datos
Hoteles	668
Hostales	533
Atracciones (y Tours)	462
Restaurantes	1463

Tabla 5: Cantidad de datos al 14 de febrero de 2012.

Además de esta plataforma de opiniones, el sitio TripAdvisor ofrece a sus visitantes un foro en el cual los usuarios registrados pueden hacer preguntas o comentarios respecto a cualquier tema relacionado al turismo. Este foro se encuentra categorizado según el lugar geográfico o según algún tema turístico determinado y no hay restricciones de idioma para los mensajes. La dinámica es similar al sistema de opiniones salvo que en este caso, los tópicos no están definidos por la categoría y se puede opinar abiertamente.

Dicho desarrollo consideró apropiado el procesamiento de los datos nulo o mínimo previo al almacenamiento, esta base de datos se alimenta con datos crudos off-line, que evita la dependencia de una conexión rápida a Internet y los problemas de variabilidad de los datos on-line.



Aspect	Positive Score	Negative Score	Relative Importance
hotel	100%	-21%	100%
room	96%	-18%	97%
view	54%	-3%	64%
breakfast	37%	-7%	37%
staff	43%	-5%	48%
lake	38%	-2%	45%
night	17%	-8%	13%
place	25%	-5%	26%
restaur	22%	-4%	23%
town	20%	-3%	21%
locat	25%	-2%	28%

Ilustración 4: Vista de resultados módulo Web Opinion Mining (E. Marresse).

3.4.1 Estructura de Base de Datos W.O.M.

La estructura de almacenamiento de los datos está sustentada por el motor de bases de datos PostgreSQL en su versión 9.1.3.

Lo relevante para este trabajo es la porción del modelo de datos que refleja los resultados del proceso WOM, el que se almacena en 2 tablas estructuradas. Un punto importantísimo es que cada proceso de WOM “limpia” la información anterior, dicho de otro modo, no guarda historia de procesamiento.

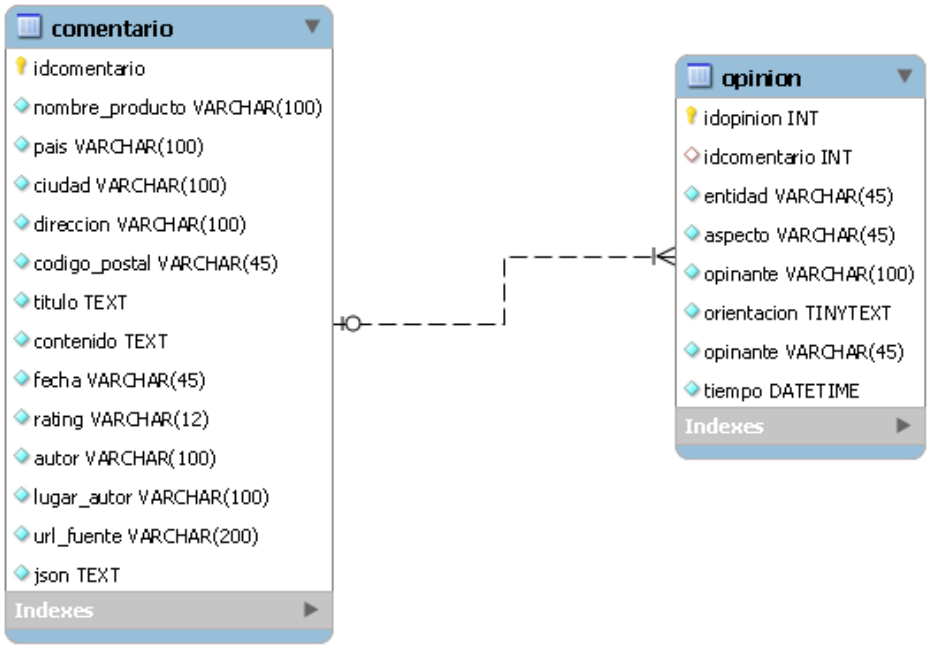


Ilustración 7: Modelo de resultados en WOM

El modelo toma la información de la página www.tripadvisor.com, elegida tanto por la gran cantidad de opiniones de turistas de todo el mundo como por el orden que mantiene, ya que divide sus áreas en lugares, productos y servicios.

Son dos tablas principales:

1) Comentario: Basado en los datos básicos de la opinión, pero externos a ella. Dentro de sus atributos se encuentra con qué producto se relaciona, su ubicación, quién la hizo, país de procedencia del autor y otros datos externos.

2) Opinión: Esta tabla se basa ya en características internas de la opinión. En ella. se reconoce su entidad, se detiene en qué aspecto de la opinión se hizo, qué características tenía la persona que escribió tal opinión (conexión con la tabla anterior) y se le da un enfoque con respecto a su orientación y el tiempo. Esta orientación se averigua por medio de un proceso de Sentiment Analysis.

3.5 AWS, Recomendador de Productos.

Este componente es otro desarrollo de memoristas anteriores, y se encuentra incorporado al sitio de WHALE, sin embargo no cuenta con base de datos, y opera por consulta directa sobre la base de datos del sitio web, haciendo sus recomendaciones directamente en la página web en un período desfasado, es decir no es en línea.

3.6 Data Warehousing en el Proyecto.

Los datos a considerados tanto de los turistas como de las empresas e instituciones se almacenan en un modelo central de datos (especificado en la Ilustración 2). Además, cuenta con modelos satélites alimentadores del modelo central; las fuentes iniciales van desde archivos de instituciones involucradas en el turismo de la X Región (SERNATUR, GORE X, INE) a información web de turistas (módulo WOM que oberva el sitio TripAdvisor). El esquema del proyecto en cuanto a información se especifica en la Ilustración 10.

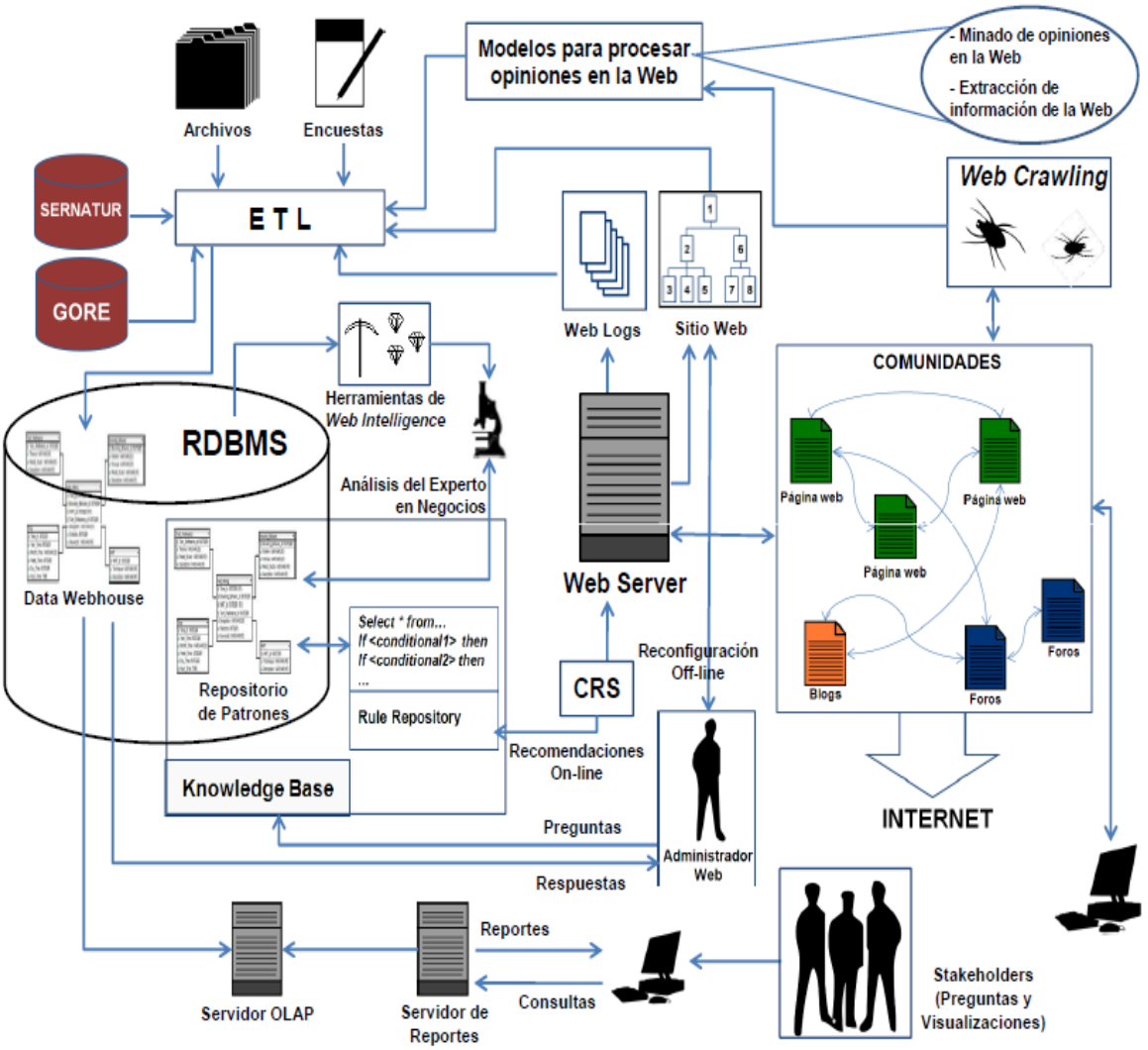


Ilustración 10: Esquema del proyecto WHALE

Capítulo 4.- LEVANTAMIENTO DE REQUERIMIENTOS.

Los requerimientos que se solicitaron para el presente trabajo fueron:

1. Diseñar y construir un modelo de datos, y los respectivos procesos automáticos de carga y transformación, para almacenar los datos registrados en el LOGs de acceso al sitio web “patgonialoslagos” del proyecto W.H.A.L.E.
2. Diseñar y construir una interfaz de datos, y los respectivos procesos automáticos de carga y transformación, para incorporar los datos con los que opera el recomendador de productos (AWS) del sitio web del proyecto W.H.A.L.E.
3. Diseñar y construir una interfaz de datos, y los respectivos procesos automáticos de carga y transformación, para incorporar los datos que obtiene el módulo de análisis de opiniones web (WOM) del proyecto W.H.A.L.E.
4. Diseñar y construir un modelo de datos, y los respectivos procesos automáticos de carga y transformación, que considere los datos centrales del negocio con los que opera el sitio web del proyecto W.H.A.L.E.
5. Diseñar y construir un modelo central de datos, y los respectivos procesos automáticos de carga y transformación, que permita la combinación de los módulos componentes del sitio web del proyecto W.H.A.L.E.
6. Diseñar y construir un modelo de datos, y los respectivos procesos automáticos de carga y transformación, que permita la eficiente generación de reportes para los distintos perfiles de acceso a la información.

ALCANCES:

1.- La información relativa al turismo la proporcionan los estamentos involucrados en el proyecto.

2.- Los desarrollos anteriores, Sitio WEB, WOM, y AWS, no se modifican y la interacción del presente proyecto se limita a la lectura de su base de datos y traspaso al DataWebHouse.

3.- La reportabilidad e indicadores, está sujeta a disponibilidad de los datos.

4.- La información utilizada y desplegada en el presente trabajo, debe cumplir las restricciones de la normativa legal vigente al respecto.

Capítulo 5.- IMPLEMENTACIÓN DEL DATA WEBHOUSE.

El desafío elemental del presente trabajo consistió en elaborar un modelo ampliado de datos que permita albergar un Data WebHouse central unitario con información de un negocio específico que publica sus productos y servicios en un sitio web. Para ello se debía elaborar un modelo central compatible con la información disponible al interior de la organización, que contuviera líneas comunes a todas las áreas involucradas; además se pedía considerar información pública relacionada con el negocio tratado.

Junto con un modelo de datos representativo, se debió elaborar los procedimientos de extracción de información y carga para esos datos, los cuales obedecen una lógica que permanezca en el tiempo y no sólo a una carga inicial. Estos procesos de carga, ETL, constan de etapas como: la extracción de los datos disponibles, procesos de limpieza de datos, asignación de valores a datos faltantes sin afectar la calidad de la información, combinación de datos, agregación y cálculo de cifras grupales, etc, y carga final al Data Webhouse.

Tal como se señaló anteriormente, existen trabajos ya realizados por otros memoristas que ya conducen el estudio de la secuencia de clicks en un sitio web, tratando de identificar comportamiento de los usuarios en la búsqueda de productos y servicios. Esta tarea se lleva a cabo con la construcción de un modelo estrella simple, que a través de procesos ETL, se nutre del LOG de navegación del sitio web.

Así mismo, respecto del WOM y AWS son trabajos también ya desarrollados por otros memoristas, aplicados al proyecto W.H.A.L.E., los cuales mostraron la compatibilidad con el Data Webhouse con distintos resultados para cada caso.

Uno de los elementos importantes que surgen a partir de los requerimientos del listado de indicadores que se solicitó consultar, es que se pretendió elaborar el modelo de Data Webhouse para que pudiera responder esas consultas prefijadas, sin embargo la provisión de esa información, o al menos su estructura, no fueron compatibles con procesos automáticos, ni los mecanismos de captura de la información, de manera que

no se pudo asegurar la respuesta aportativa con los datos desplegados en los reportes, ya que sólo se disponía de datos agregados por región que corresponde al contexto global del proyecto; con ello el modelo se tuvo que dirigir hacia las herramientas fidedignas disponibles para dar valor a los reportes, como lo es el análisis del LOG del sitio web.

5.1 Capas y Herramientas de Software.

La estructura del Proyecto WHALE se dividió en tres capas de funcionamiento, las que se muestran en el siguiente esquema, y que se relacionan en distintos niveles para proveer información y servicios:

1.- Ambiente WEB: es la capa más externa y con la cual se relacionan los usuarios del sistema. Existen distintos perfiles, que a su vez tienen funcionalidades específicas según su participación para con el sitio.

2.- Procesos de Business Intelligence (BI): Es el “cerebro” del sistema de análisis, sus funciones son exclusivamente internas en información, y administra las consultas que se hacen a los datos del sitio.

3.- DataWebHouse: Es el almacén de datos ampliado, en donde se guarda la información que luego será solicitada por los usuarios. Se relaciona exclusivamente con los procesos de BI y permanece alejado de todo acceso externo; por otra parte, almacena un gran volumen de información y la recibe desde distintos componentes de manera automática.

Esta estructura soporta los procesos fundamentales del presente trabajo: carga de datos y explotación de información.

Entre las alternativas de construcción del DataWebHouse del proyecto W.H.A.L.E. se eligió las opciones de Software Libre, por razones de costo y de apertura de sus códigos que permiten modificarlos a medida según lo requiere el desarrollo.

Lo anterior determinó también la elección de la base de datos para el DataWebHouse, se eligió Mysql versión 5.0 por ser la última más estable a la fecha propuesta del proyecto, y la última libre de costo, además de tener un desempeño eficiente comprobado en desarrollos con software libre. Por otra parte, en ETL se eligió TALEND en su versión comunidad para las cargas de datos dado los mismos motivos anteriores, esta herramienta tiene una relación directa (en su versión comercial pagada, por lo que es razonable suponerlo para su versión comunidad) con el software JASPERSOFT, que fue elegido para el sitio del proyecto W.H.A.L.E. por su mejores características para la presentación de reportes en un sitio web.

5.2 Modelo de Solución.

A través del sitio WEB se relacionan distintos tipos de usuarios con este sistema, según sus intereses se pueden clasificar en empresarios (con facultad de publicar avisos y modificarlos) y no empresarios, éstos a su vez se dividen en usuarios anónimos y usuarios registrados los que sólo leen información del sitio.

En un nivel de uso inicial, se dispuso de un listado de empresarios mínimo oficial y a partir de allí se podrían agregar nuevos participantes con ese perfil. A partir de allí se contó con un set de productos bastante reducido, el cual se espera se incremente con el uso del sitio. Con esto se quiere hacer notar la cantidad reducida de registros de productos disponibles para la marcha blanca, pero que no debe afectar al modelo implementado ya que debiera resistir el aumento de volumen de datos con el tiempo.

Por los mismos motivos, la cantidad de usuarios registrados también es reducida en la marcha blanca, y la mayor cantidad de información se obtuvo de los visitantes anónimos al sitio web. Lo anterior, tuvo la complejidad para encontrar las estadísticas

de uso del sitio web en que sólo se podía distinguir el origen por la IP desde donde se hizo la solicitud web; por lo tanto se asumió (para efectos de cálculo) que un usuario se conectaba por una IP por hora.

5.3 Modelo para WEB LOG.

Para el procesamiento del LOG del sitio web se siguió lo propuesto y comprobado en trabajos anteriores, que consiste en un proceso diario que toma el LOG del sitio lo traspasa a una carpeta prefijada como INPUT y luego se ejecuta un proceso del Sistema Operativo que carga ese LOG completo a una Tabla en la Base de Datos "STAGING"; dicha tabla tiene un solo campo de tipo texto, y a partir de allí se gatillan dos procedimientos almacenados que separan cada línea del LOG y la trozan en sus distintos componentes (campos de ancho fijo) para llevarlos finalmente a una tabla que se queda con lo relevante, desde donde se carga al modelo estrella.

LOG --> CRON --> BD. STAGING --> Modelo estrella

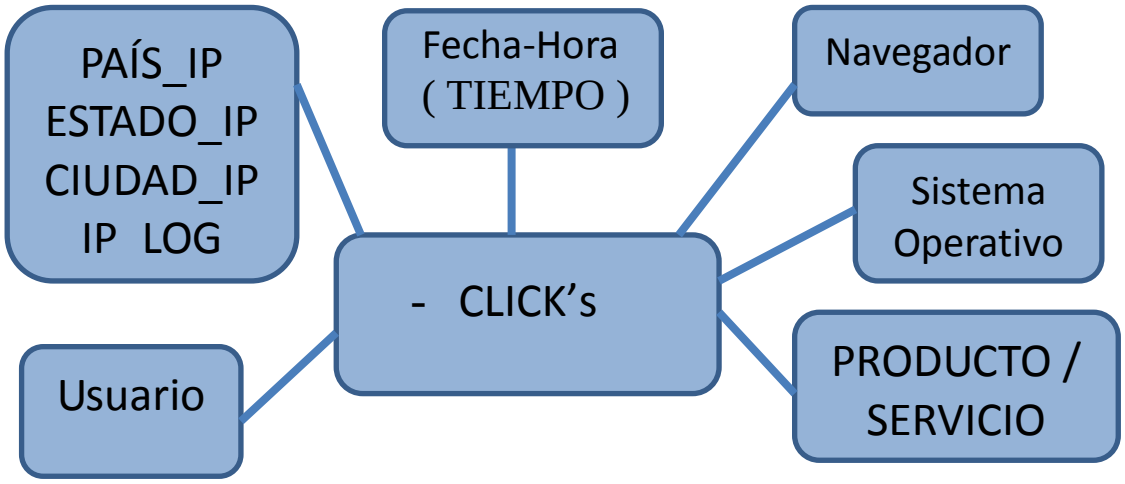


Fig .- Modelo de Datos Estrella para el WEB LOG.

5.4 Elaboración y Construcción del Data Webhouse.

Inicialmente se diseñó el DataWareHouse para contener la información del negocio, desprendible del sitio web, y de la información proporcionada por las autoridades involucradas, y con la información pública disponible relativa al turismo. Más tarde se le agregó la parte del Web LOG y WOM para concretar el DataWebHouse.

Este sector usualmente contiene los datos que se utilizan en todo el sistema, que se acumulan en gran cantidad, que se almacenan con una cierta estructura y orden según el objetivo de análisis, los cuales obedecen las reglas del negocio. Para almacenar estos datos se recopiló primero desde fuentes autorizadas, que para este proyecto son la información relativa al turismo de la X Región publicado por el INE, las que lamentablemente no contaban con información detallada, ni con periodicidad estable, ya que se publican estadísticas del turismo en tramos de 3 a 6 meses, y con formatos que no aseguran continuidad.

Un elemento que afectó fuertemente es el aspecto legal, ya que el INE no puede entregar datos que individualicen a los empresarios del turismo bajo ningún aspecto, por lo que si un empresario fuese el único que entregase un servicio dado en alguna comuna podría entonces individualizarse con esa consulta; el resultado: contamos sólo con estadísticas regionales y datos ya agrupados.

Lo anterior afectó el desarrollo del formato de los sistemas automáticos de carga y los reportes esperados. El aporte de las instituciones involucradas en el proyecto era fundamental como expertos del negocio en cuestión, SERNATUR de la X Región y Gobierno Regional de la X Región, tal como sucede en cada proyecto de Business Intelligence, pero esta colaboración (sólo en esta parte del proyecto) fue débil e implicó que esta Base de Datos “corporativa” fuera perdiendo peso frente al análisis de navegación WEB, la cual finalmente dominó el modelo de datos.

5.4.1 Modelo de Datos Inicial del Área de Negocio.

El modelo inicial propuesto para el área del negocio apuntaba a responder la lista de 41 indicadores, según la solicitud de los empresarios y autoridades involucrados en el proyecto. Como respuesta a este requerimiento, se elaboró un modelo de datos para el Data Warehouse, de manera tentativa inicial, que permitiera extraer los indicadores solicitados y que conservara las características de compatibilidad con los otros componentes del proyecto y sus bases de datos.

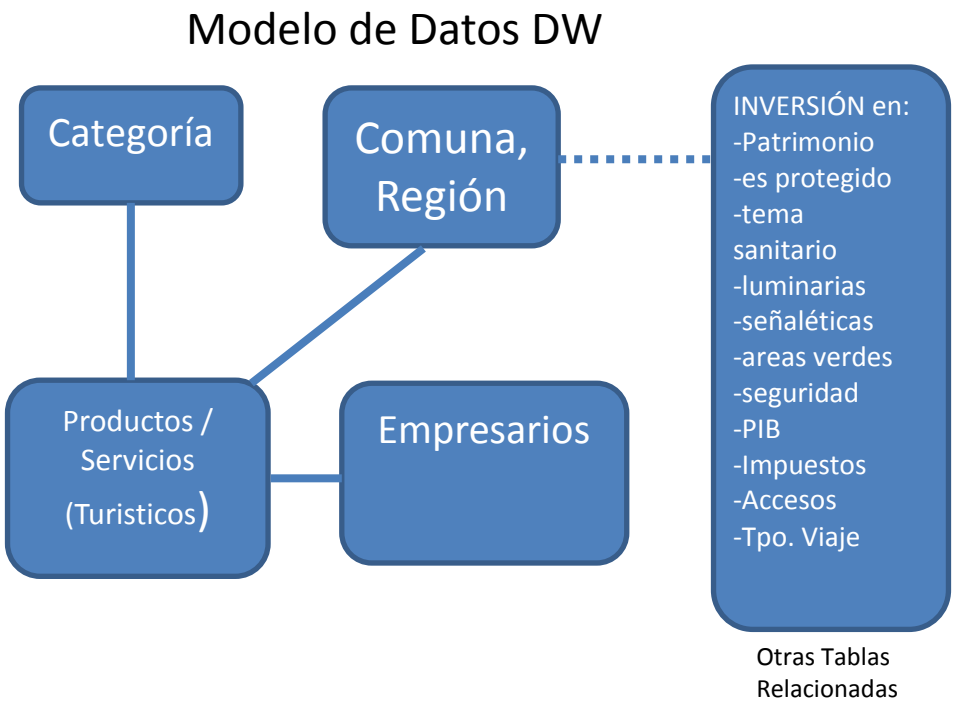
Dicho modelo consideraba cada uno de los elementos solicitados, estuvieran al alcance desde el principio o no, luego se vería la factibilidad de obtenerlos; sin embargo, se destacan indicadores de satisfacción que inicialmente requerían el uso de encuestas, las que fueron descartadas al definir el proyecto por los costos y la cobertura involucrados. Para emular esa información se confía que el avance en el desarrollo de los analizadores de opiniones en la web aporten con datos que permitan sacar conclusiones en esa línea, pero ya no como registros detallados. Lo paradójico es que la información regional con la que contamos por parte de las autoridades involucradas, se obtiene a partir de encuestas que se les hace a los empresarios del turismo de la X Región, pero no se comprueba la fidelidad de esos datos.

Como análisis posterior al modelo presentado inicialmente, se puede notar que en los modelos corporativos de Data Warehouse el centro lo ocupa la demanda propiamente tal, cuya tabla contiene los montos de cada operación de compra de los productos o servicios y su fecha, además de características geográficas y eventualmente características del cliente involucrado; acá se pretendía llegar a un concepto similar pero a través de la captura de información, posiblemente a través de la colaboración de los mismos empresarios, o por medio de una breve encuesta en el mismo sitio, lo que fue descartado como ya se señaló.

Modelo de Datos Inicial DW

5.4.2 Modelo de Datos Final del Área de Negocio.

El modelo final correspondiente al área del negocio, del turismo de la X Región, fue recogido principalmente del sitio web, esto se tradujo en que el centro de estudio del movimiento web es el click sobre los diferentes productos (o servicios), por lo que la información central más estructurada para estudiar es justamente las entidades “productos” y “empresarios” provenientes de la base de datos del sitio web.



5.4.3 Modelo de Datos Ampliado, Data Webhouse.

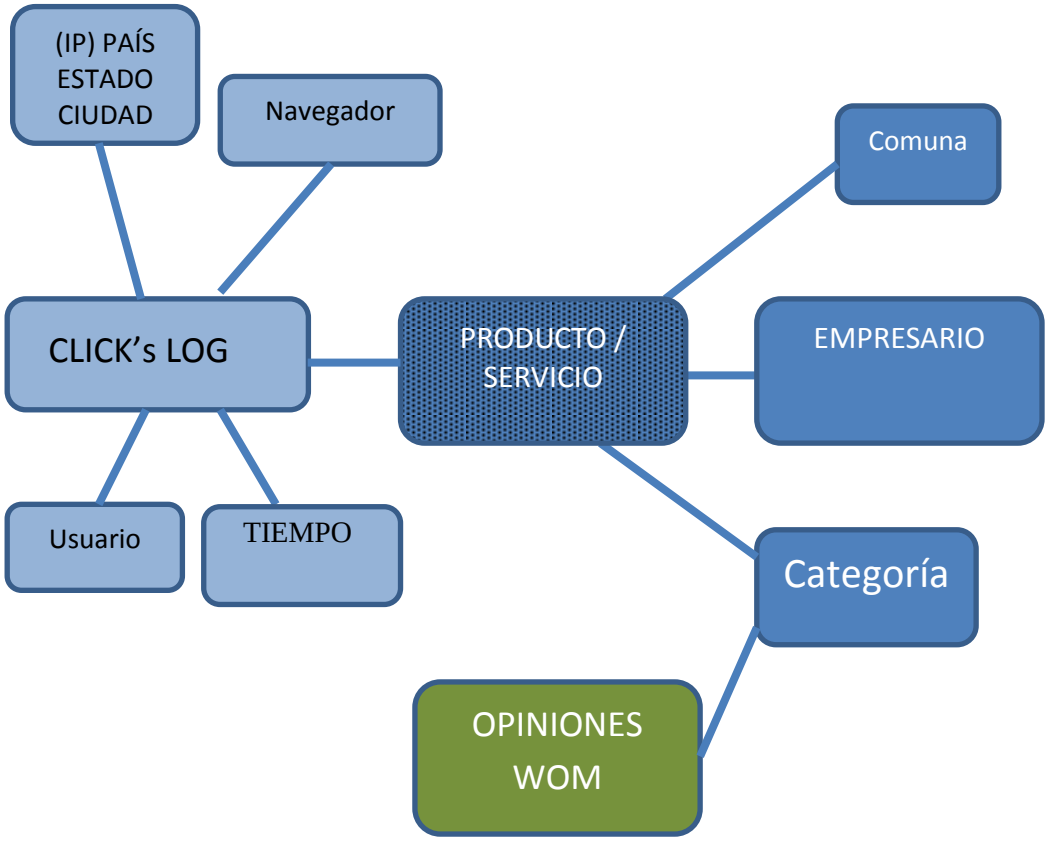


Fig.- Esquema Modelo Final DataWebHouse

5.5 Interacción de Módulos y de Datos.

5.5.1 Interacción con módulo de AWS.

El sitio web cuenta con un módulo adaptativo, que genera una lista de productos a recomendar a partir del análisis del comportamiento de los usuarios dentro del sitio, para ello el módulo AWS revisa la información del sitio web desde la base de datos web y genera su recomendación, pero no cuenta con una Base de Datos propia, lo que hace casi imposible la interacción con ese componente, a menos que se pueda influir en la lista de usuarios para que reformule su resultado, la otra posibilidad es modificar los parámetros con los que opera el AWS desarrollado, pero eso está fuera de lo considerado para el presente trabajo.

De cierto modo, lo que nos resta es considerar que el efecto del recomendador de productos influye en el sitio web, y eso se captura en el análisis del LOG.

5.5.2 Interacción con WOM.

El módulo de análisis de opiniones en la web funciona analizando las opiniones vertidas en el sitio de TripAdvisor para los productos o servicios relativos al turismo de la X Región. Aunque las conclusiones, y datos, que aporta este módulo son de extrema relevancia para el objetivo del proyecto, sólo están disponibles para los servicios de “Restaurant” y “Hotel” lo que implica que el cruce posible con otros datos sólo puede hacerse en esos dos casos; sin embargo, lo positivo es que la forma de construcción del módulo WOM permite agregar otros procesos de análisis de opinión para otros productos o servicios, pero no es competencia de este trabajo la creación de dichos anexos.

Este módulo cuenta con Base de Datos propia, pero su procesamiento es manual, por lo que la renovación de los resultados dependerá de la periodicidad con la que los analistas realicen ese procesamiento.

Para la interacción con el DataWebHouse se realiza una extracción de las tablas del WOM que contienen los resultados de sus análisis directamente, eso no tiene mayor complejidad, sin embargo el hecho de que el proceso sea manual hace que no quede establecida la periodicidad del traspaso de datos, lo que puede generar inconvenientes ya que cada proceso borra los resultados anteriores.

Un elemento adicional, que puede perfeccionarse mejorando el módulo WOM, es que los resultados corresponden a las categorías de producto y no al producto directamente, lo que hace que los análisis sean indirectos.

5.6 Diseño de Interfaces por medio de ETL.

Las tareas que realizan el traslado de los datos se manejan desde el sistema operativo, en donde se dejan establecidas las tareas de carga y sus horarios.

Las tareas de traspaso de datos internas a cada base de datos se realizan con procedimientos almacenados que son activados desde el mismo sistema operativo.

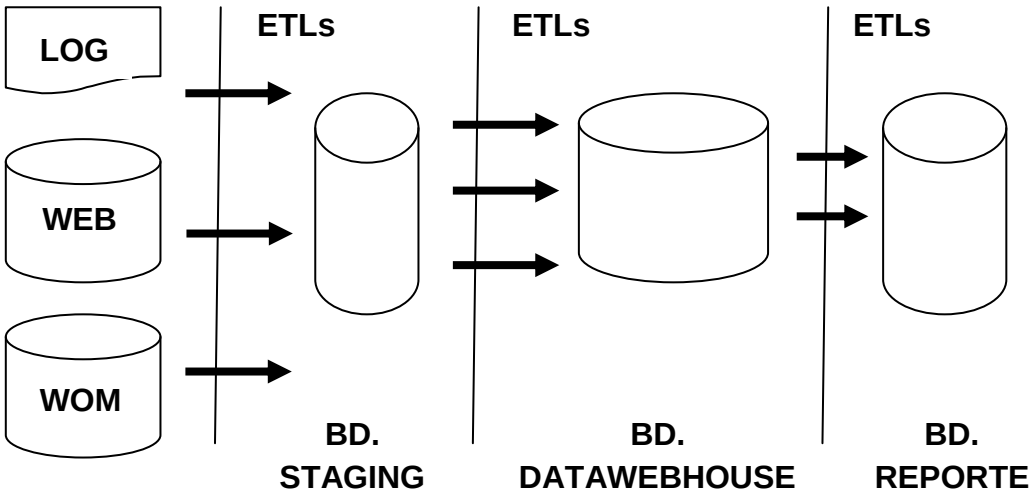


Fig.- Esquema de Traspaso

5.7 Reportes y Lista de Indicadores.

Finalmente, el módulo de reportes se realiza con una base de datos para ello, llamada “REPORTE”, que contiene tablas sin modelamiento y de las cuales se extraen los reportes que se puedan realizar en base a la información disponible. Esta forma de no-modelar los datos de salida obedece a la tendencia actual de privilegiar una rápida respuesta de los reportes, dejándole el trabajo “pesado” a los procesos de carga pre-establecidos, en donde se aplican las reglas del negocio y la orientación del análisis.

5.7.1 Reportes basados en el WEB LOG.

El sitio construido para el proyecto W.H.A.L.E. es ***www.patagonialoslagos.cl***, respecto del cual se extrae el LOG de navegación a diario y se carga a la base de datos inicial “Staging”, allí se inserta y procesa en una tabla que acumula todos estos movimientos. A partir de esta información se realizaron distintos análisis del comportamiento de los usuarios, buscando entregar indicadores y reportes tales como:

- Tiempo estadía en el sitio.
- Velocidad de Carga.
- Ranking de Productos más pinchados.
- Desde dónde provienen las visitas al sitio (países).

El primer punto de análisis relevante es que la mayoría de las visitas (hasta el momento) son anónimas y no con usuarios registrados, lo que no permite detectar la navegación de buena manera, y por lo tanto los indicadores de navegación web, como la cantidad de visitantes por ejemplo, es una estimación a partir de la IP de origen de la solicitud WEB; con la cantidad de solicitudes desde la misma IP en tramos de 1 hora (suponemos que no pasarán más de 1 hora en el sitio) se estima el tiempo de estadía promedio, etc…

Otro de los elementos interesantes, es la procedencia de las solicitudes de navegación del sitio, gana por lejos Estados Unidos, le sigue CHINA, y luego CHILE, pero así también llama la atención “pinchazos” de un solo registro, lo que implica que no hay navegación dentro del sitio, incluso existen accesos a imágenes internas del sitio de manera directa; por lo anterior, se adjudicó este comportamiento a robots de navegación web y no a usuarios reales.

Por otra parte, un elemento claro de identificar es la solicitud de información relativa a un producto o servicio del sitio, con lo que se cruzó ese “interés” del usuario con el correspondiente producto, y a partir de allí se rescató información más amplia de la base de datos. Para esto se utilizó la interacción con la Base de Datos del sitio web, de la cual se importaron las características del Producto y los Empresarios.

Esta combinación de la tabla de registro de navegación web y los elementos propios del negocio (del sitio web) implican el uso del término Data WebHouse (según Kimball, 2000), y en base a ello es que se resuelven otros reportes del tipo:

- Ranking de Productos, por empresario, más pinchados.
- Ranking de Productos, por categorías , más pinchados.
- Ranking de Productos, por comuna, más pinchados.

Es en este punto donde se hace el cruce con los resultados del módulo WOM, ya que sus resultados se combinan a través de la categoría de productos:

- Ranking de Productos, para Hoteles, más pinchados.
- Ranking de Productos, para Restaurant , más pinchados.
-

5.7.2 Interfaces de visualización.

La tarea de desplegar los reportes se desarrolló con la herramienta JASPERSOFT, que permite en su versión comunitaria desplegar reportes vía web con

las condiciones de seguridad necesarias, recordemos que la información que se desprende del presente desarrollo apunta a un uso de las autoridades regionales del turismo de la X Región.

Capítulo 6.- CONCLUSIONES.

1.- El desafío de compatibilidad de llevar la información hacia el DataWareHouse, con procesos ETL, se cumple para la información del sitio web, del WEB LOG, y para el módulo de opiniones WOM.

2.- El centro del modelo de datos del Data Webhouse terminó siendo la entidad "producto", que tiene mucho sentido si consideramos que el sitio web es una vitrina virtual de productos, por lo que lo propuesto en el presente trabajo se podría aplicar con cualquier otro rubro o industria.

3.- En la evolución de los análisis del log de visitas, sujeto a la individualización de los usuarios, en el futuro se podría detectar las secuencias de productos clave.

4.- Se cumple la premisa de la interacción de los componentes pero parcialmente, ya que no todos los componentes cuentan con una base de datos para relacionar.

5.- La información de las opiniones resulta ser cualitativa lo que dificulta la combinación con la información cuantitativa de las visitas a los productos del sitio web, por lo que aunque se pueden mostrar juntos no se combinan.

6.- El módulo de recomendación cuenta con algortimos propios de análisis y clasificación, por lo que en una futura versión debería absorber los análisis de las otras componentes o cambiarlo por una estructura abierta que permita intervención.

7.- Lo anterior induce a sugerir que el diseño del proyecto del sitio web considere originalmente las componentes involucradas, y no sólo como módulos independientes, así se elevaría en nivel de compatibilidad entre los datos conjuntos.

BIBLIOGRAFÍA.

[1] KIMBALL, R and Margy Ross, R. *The Data Webhouse Toolkit*. Wiley Computer Publisher, New York, 2000.

[2] Adaptive Web Sites: A Knowledge Extraction from Web Data Approach. IOS Press, Amsterdam. Velasquez, J.D., Palade, V. (2008).

[3] TEIXEIRA, C. y G. DAVID. 2006. Higher Education Web Information System Usage Analysis with a Data Webhouse. Computational Science and Its Applications – ICCSA.

[4] Pang, B. y Lee, L. (2008) Opinion mining and sentiment analysis. En Foundations and Trends in information Retrieval. Vol. 2, pp. 1-135.

[5] Ku, L.-W. y Chen, H.-H (2007). Mining Opinions from the web: Beyond Relevance Retrieval. Journal of American Society for Information Science and Technology, 58(12), pp. 1838-1850.

[6] "Desarrollo de una plataforma tecnológica genérica basada en web intelligence de apoyo al diseño y aplicación de mejores estrategias de creación de valor en la industria de los servicios: Experiencia demostrativa en el Clúster del Turismo de la Región de Los Lagos de Chile", Formulario de presentación, XVIII Consurso Nacional de Proyectos de Investigación y desarrollo FONDEF 2010.

[7] J.D. Velásquez and V.L. Rebolledo (2010). Web Intelligence. En H. BIGDOLI (Ed.) Then Handbook of Technology Management. Pp. 639-673, Willey.

[8] A. JOSHI and R. KISHNAPURAM (2000). On mining web access logs. In ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 63-69.

[9] G.D. KOBLINC (1998). Web mining: Estado actual de investigación. Departamento de Ciencias de la Computación, Universidad de Buenos Aires - Argentina.

[10] R. KOSALA and H. BLOCKEEL (2000). Web mining research: A survey. En SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM.

[11] J.D. Velásquez; S.Ríos; A. Bassi; H. Yasuda and T. Aoki (2005). Towards the identification of keywords in the web site text content: A methodological approach. International Journal of Web Information Systems. Vol. 1, N°1, pp. 11-15.

[12] R.F. Dell, P.E. Román and J.D. Velásquez (2008). Web User Session Reconstruction Using Integer Programming.; p.4; AUSTRALIA; Procs. of The 2008 IEEE/WIC/ACM International Conference on Web Intelligence, Sydney, Australia.

[13] R.F. Dell, P.E. Román and J.D. Velásquez (2009). Optimization Models for Construction and Analysis of Web User Sessions; p. 1; Procs. 11th Informs Computing Society Conference, Charleston, South Carolina, USA.

[14] J.D. Velásquez; A. Bassi; H. Yasuda and T. Aoki (2004). Mining web data to create online navigation recommendations. In Procs. 4th IEEE Int. Conf. on Data Mining. Pp. 551-554.

[15] J.D. Velásquez and V. Palade(2007). Building a Knowledge Base for Implementing a Web-Based Computerized Recommendation System. International Journal of Artificial Intelligence Tools. Vol. 16, N°5, pp. 793-828.

[16] J.D. Velásquez and V. Palade(2007). A Knowledge Base for the maintenance of knowledge extracted from web data. Journal of Knowledge-Based Systems. Vol. 20, N°3, pp. 238-248.

[17] IALE. (2010). Observatorio de Inteligencia Regional de Turismo en la Región de Los Lagos, Informe Final.

[18] Laesser, C. (2008). Predicting Online Travel Purchases: The Case of Switzerland. EMAC Conference, 26, pág.28. Nantes.

[19] BONTEMPO, C. y G. ZAGELOW. 1998. The IBM Data Warehouse Architecture. Communications of the ACM, Volume 41. New York, USA.

[20] MCFADDEN, F.R. 1996. Data warehouse for EIS: some issues and impacts. Wailea, HI, USA.

[22] HU, X. y N. CERCONE. 2004. A data warehouse/online analytic processing framework for web usage mining and business intelligence reporting. Journal International Journal of Intelligent Systems - Granular Computing and Data Mining, Volume 19 Issue 7. New York, USA.

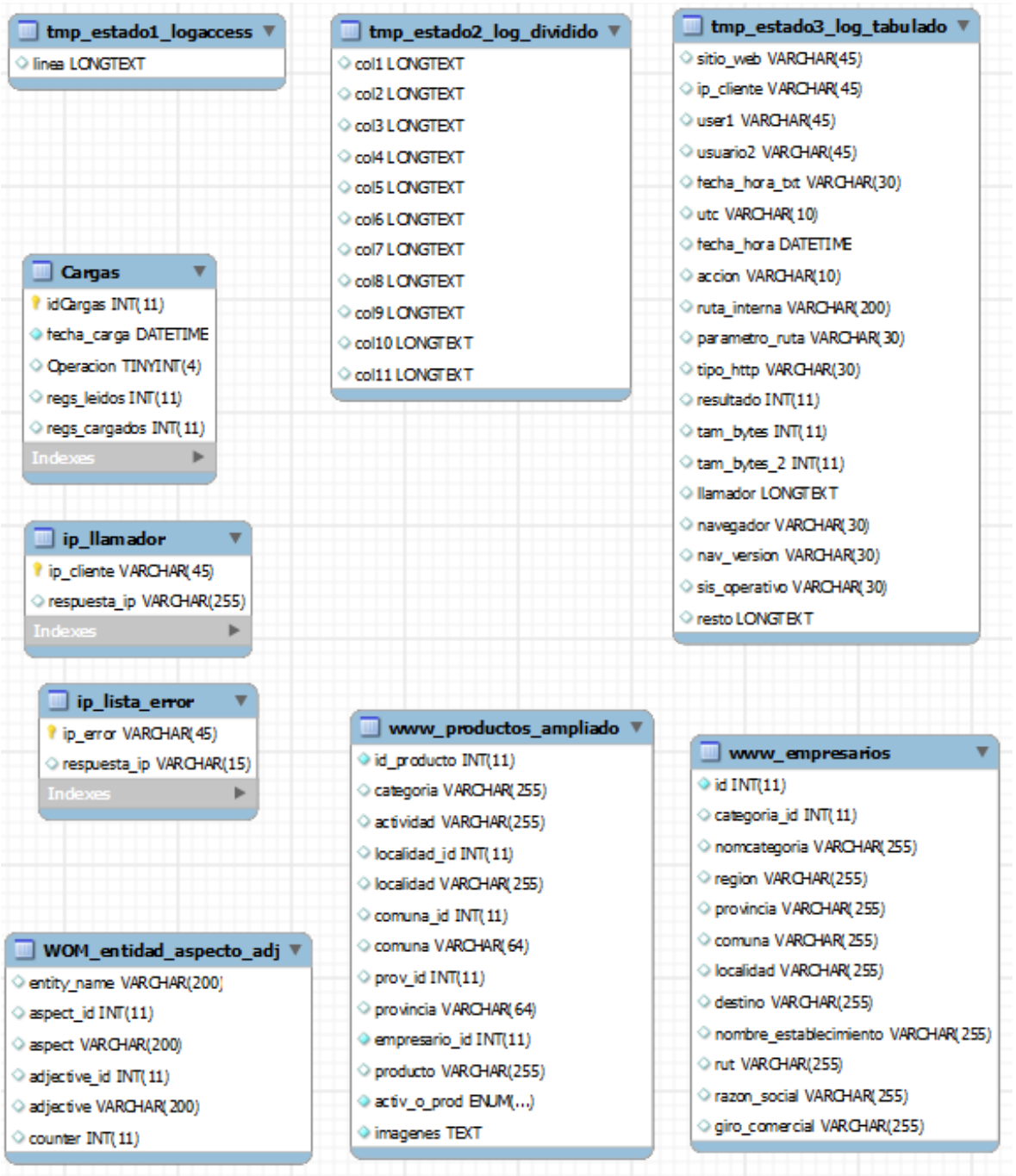
[23] KIMBALL, M. y M. ROSS, 2002. The Data Warehouse Toolkit. Segunda Edición. John Wiley y Sons.

[24] BALLARD, C.; D. HERREMAN; D. SCHAU; R. BELL; E. KIM, y A. VALENCIC. 1998. Data Modeling Techniques for Data Warehousing. IBM

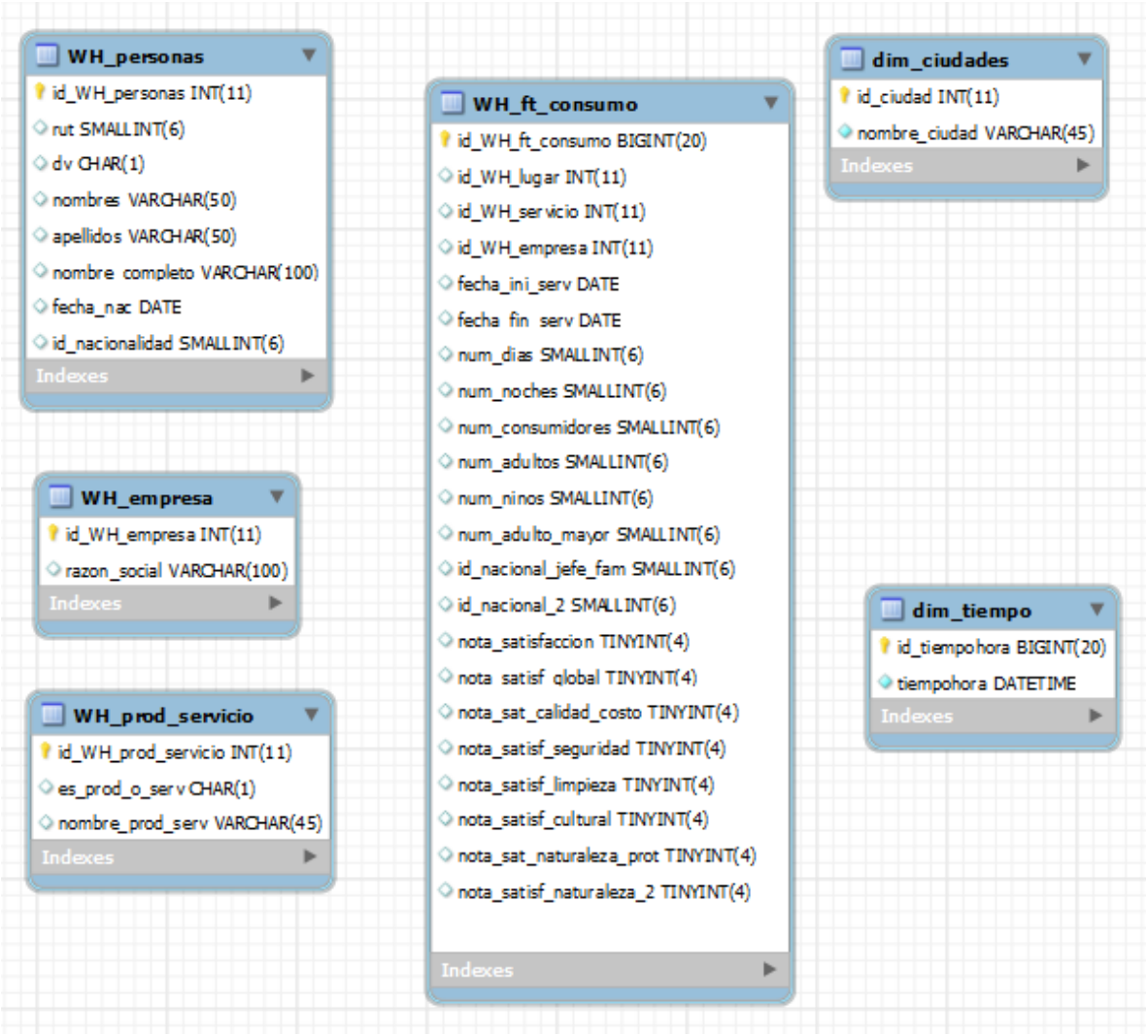
[25] VELÁSQUEZ, J.D. y V. PALADE. 2008. Adaptative Web site: a knowledge extraction form Web data approach. IOS Press, Netherlands

ANEXOS.

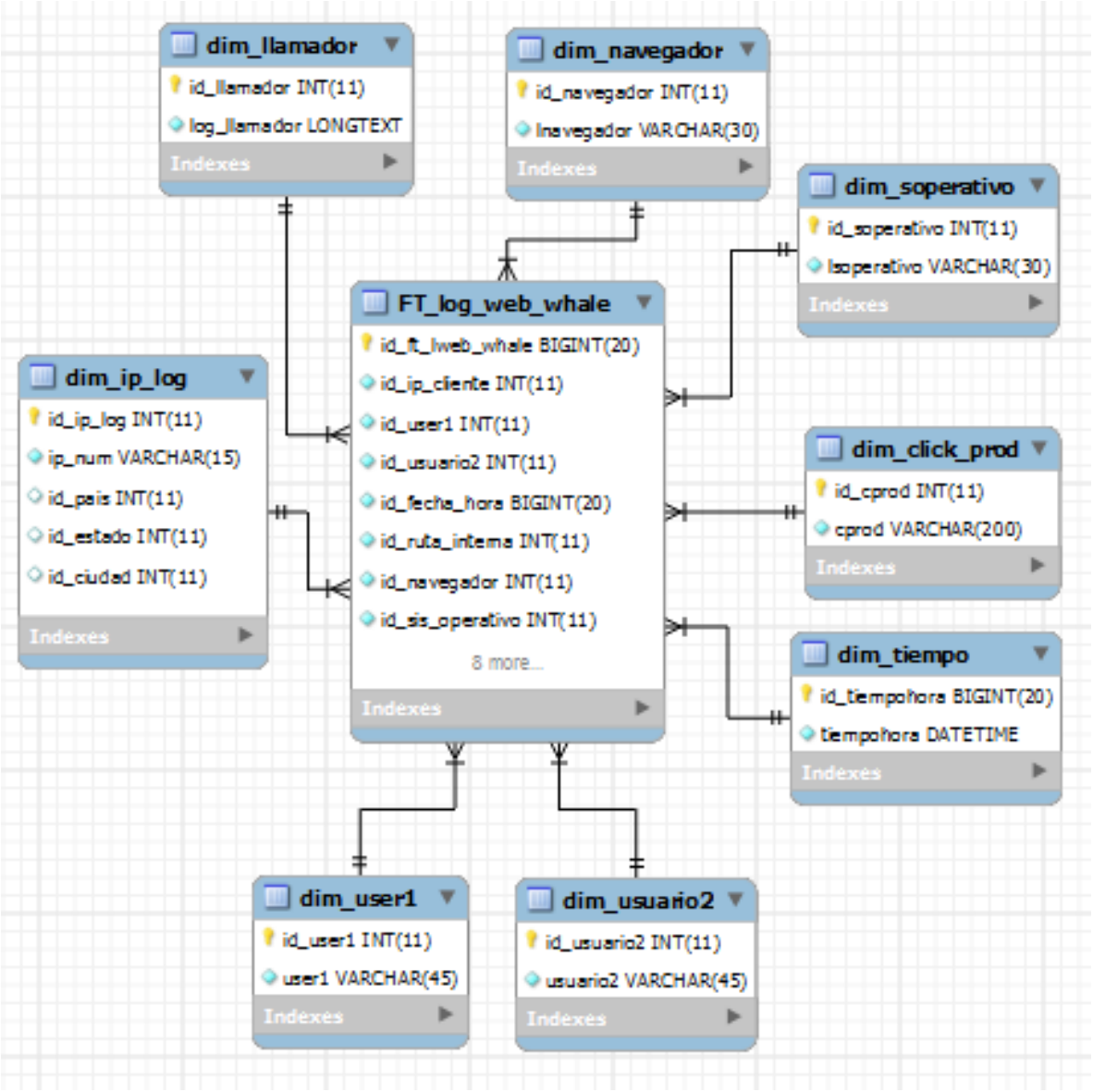
A1.- Modelo de Datos de la Base Staging_Area.



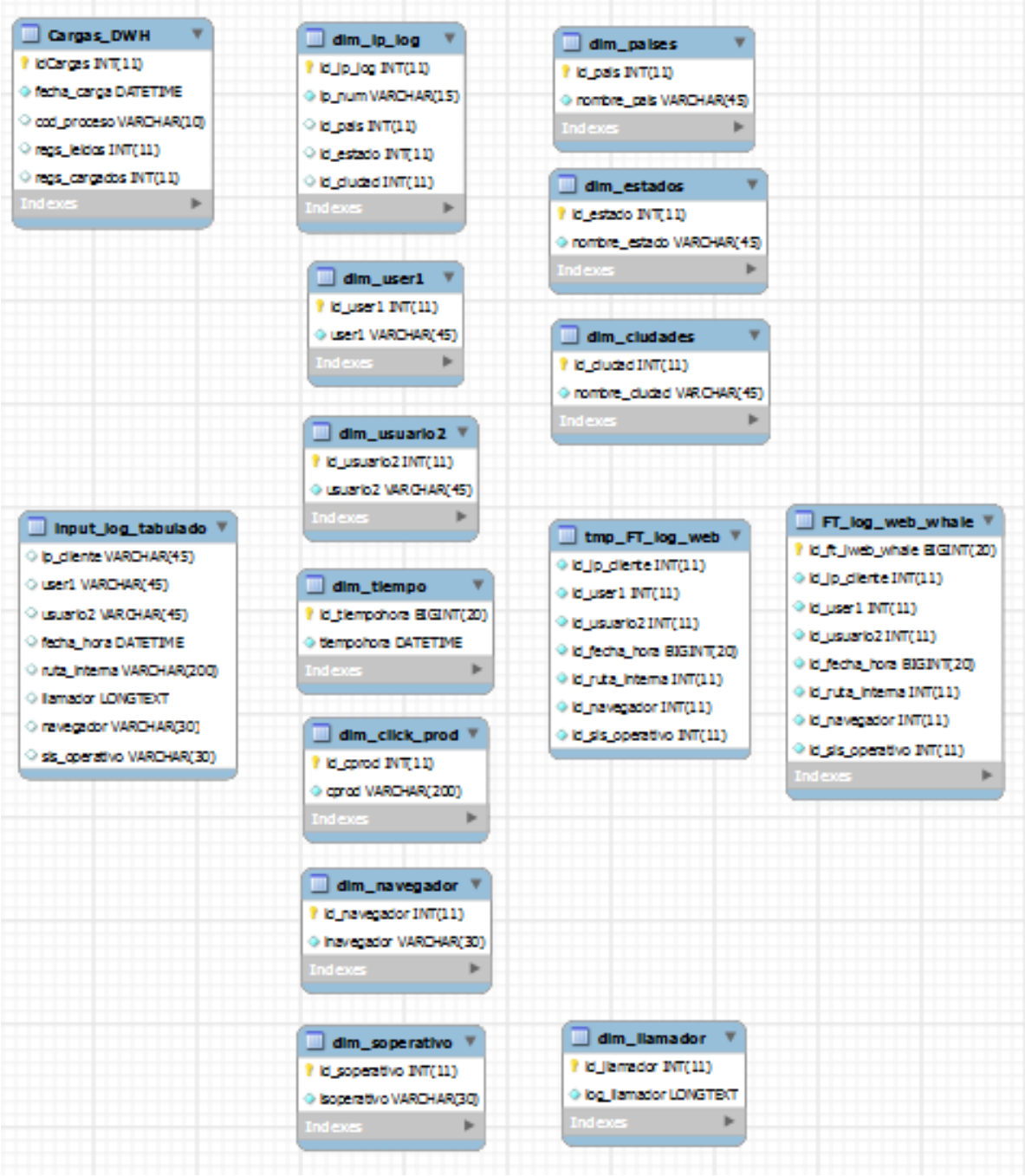
A3.- Modelo de Datos Inicial DW.



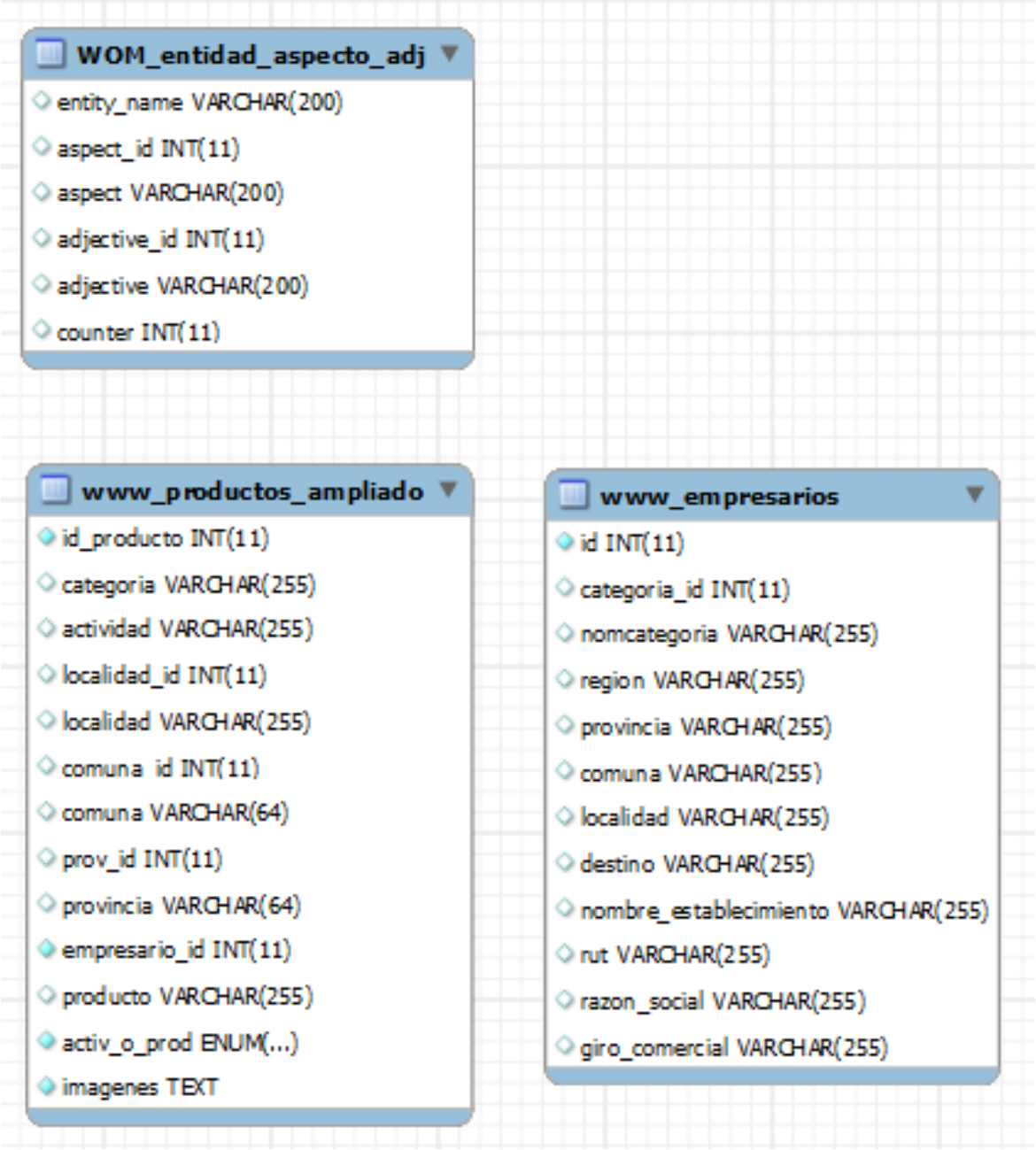
A4.- Modelo de Datos Estrella para el LOG WEB.



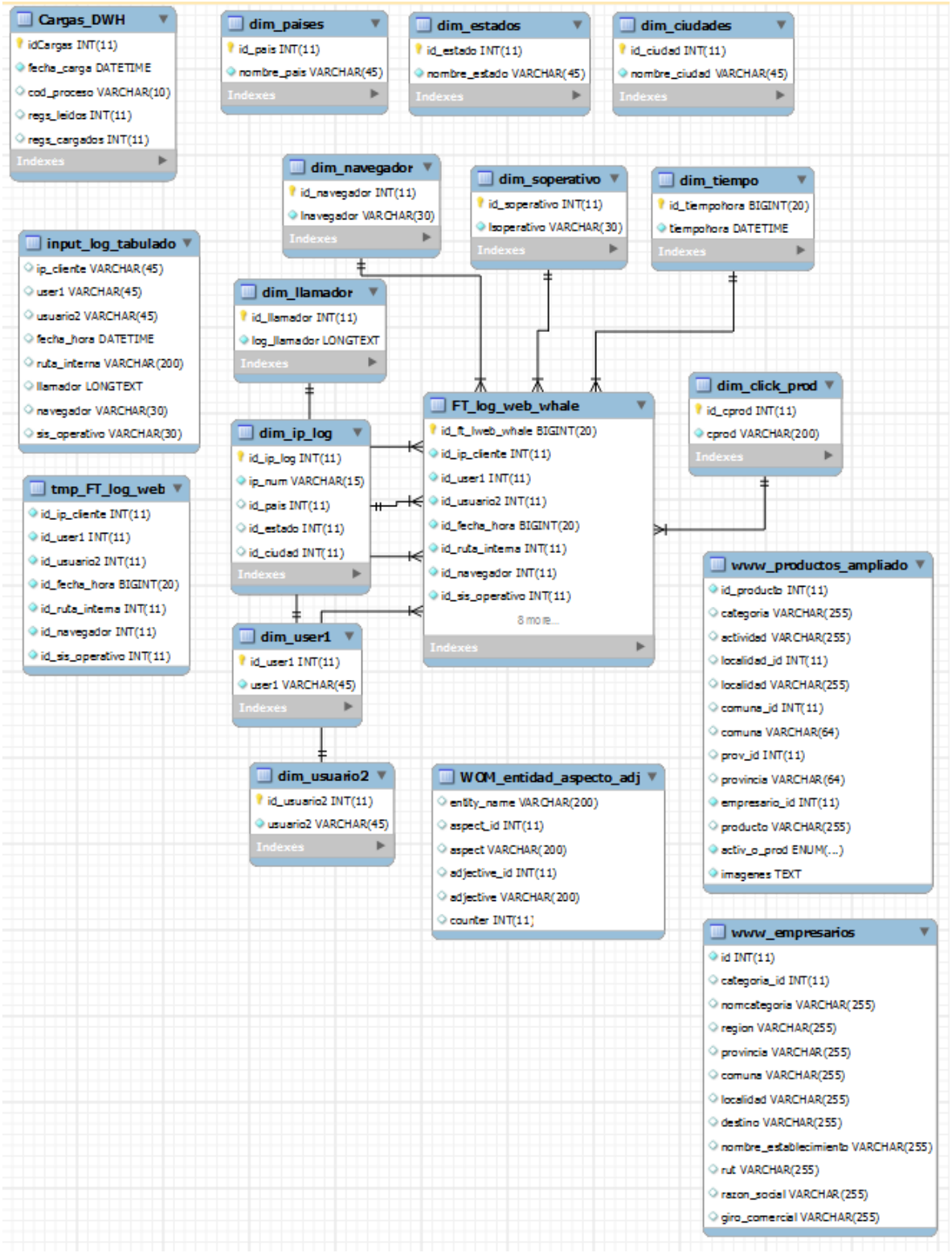
A5.- Tablas con el paso de Datos desde el LOG WEB.



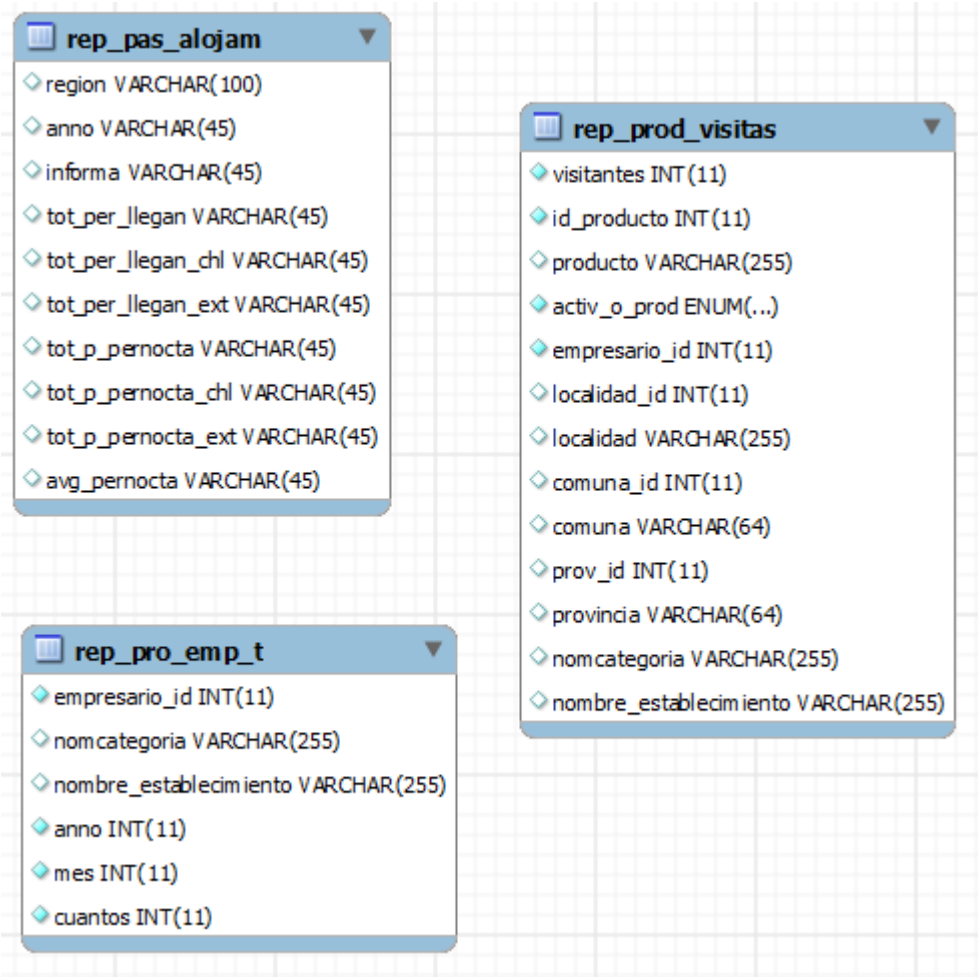
A6.- Tablas con datos importados desde WOM y sitio patagonia.



A7.- Modelo de Datos ampliado - Data WebHouse.



A10.- Tablas creadas para reportes básicos. (REPORT)



B.- Anexos con ETL.

B1.- ETL de trasporte de datos del LOG hacia el Data Webhouse.

```
Se incluye sólo un extracto del código automático generado.
// ======================================================================

package whale_traspasa.traspasa_log_a_dwh_0_1;

String insert_tMysqlOutput_1 = "INSERT INTO `" + "input_log_tabulado"+ "`
(`ip_cliente`,`user1`,`usuario2`,`fecha_hora`,`ruta_interna`,`llamador`,`naveg
ador`,`sis_operativo`) VALUES (?,?,?,?,?,?,?,?)";
…
String dbquery_tMysqlInput_1 = "SELECT  `ip_cliente`,  `user1`,  `usuario2`,
`fecha_hora`,  `ruta_interna`,  `llamador`,  `navegador`,  `sis_operativo`
FROM `area_staging`.`tmp_estado3_log_tabulado`   where   `accion` = 'GET'  ";

globalMap.put("tMysqlInput_1_QUERY", dbquery_tMysqlInput_1);
…

/**********************************************************************/
```

B2.- ETL de trasporte de datos de productos entre el sitio Patagonia y el Data Webhouse.

```
Se incluye sólo un extracto del código automático generado.

// ======================================================================

package whale_traspasa.traspasa_prod_ampliado_0_1;

String insert_tMysqlOutput_1 = "INSERT INTO `" + "productos_ampliado"+ "`
(`id_producto`,`categoria`,`actividad`,`localidad_id`,`localidad`,`comuna_id`,
`comuna`,`prov_id`,`provincia`,`empresario_id`,`producto`,`activ_o_prod`,`imag
enes`) VALUES (?,?,?,?,?,?,?,?,?,?,?,?,?)";
…
String dbquery_tMysqlInput_1 = "SELECT   prod.id as id_producto,   cate.nombre
as categoria, acti.nombre as actividad,   prod.localidad_id, loc.nombre as
localidad,  prod.comuna_id, com.nombre as comuna,  prov.id, prov.nombre as
provincia,   prod.empresario_id,  prod.titulo as producto,  prod.tipo as
activ_o_prod,  prod.images as imagenes   FROM productos as prod    left join
actividades as acti on acti.id = prod.actividad_id    left join categorias as
cate on cate.id = acti.categoria_id    left join localidades as loc on loc.id
= prod.localidad_id   left join comunas as com on com.id = prod.comuna_id
left join provincias as prov  on prov.id = com.provincia_id ";

globalMap.put("tMysqlInput_1_QUERY", dbquery_tMysqlInput_1);
/**********************************************************************/
```

B3.- ETL que procesa datos hacia la BD REPORT.

Se incluye sólo un extracto del código automático generado.

```
// ======================================================================

package whale_traspasa.reporte_prod_visitas_0_1;

String insert_tMysqlOutput_1 = "INSERT INTO `" + "rep_prod_visitas"+ "`
(`visitantes`,`id_producto`,`producto`,`activ_o_prod`,`empresario_id`,`localid
ad_id`,`localidad`,`comuna_id`,`comuna`,`prov_id`,`provincia`,`nombre_establec
imiento`,`nomcategoria`) VALUES (?,?,?,?,?,?,?,?,?,?,?,?,?)";
…
String dbquery_tMysqlInput_1 = "   select  count(distinct logtab.usuario2) as
visitantes , prod.id_producto,           prod.producto, prod.activ_o_prod,
prod.empresario_id ,          prod.localidad_id, prod.localidad,
prod.comuna_id, prod.comuna,          prod.prov_id, prod.provincia
, empre.nombre_establecimiento  , empre.nomcategoria    FROM
area_staging.tmp_estado3_log_tabulado as logtab       inner join
www_productos_ampliado as prod       on prod.id_producto = (replace(substring(
logtab.ruta_interna, instr( logtab.ruta_interna,'producto/ver')+13,
10),'/',''))       left join www_empresarios as empre       on empre.id =
prod.empresario_id    WHERE logtab.ruta_interna like '%producto/ver%'      and
logtab.fecha_hora > DATE_SUB(curdate(), INTERVAL 3 month)  group by
prod.id_producto, prod.empresario_id ,          prod.localidad_id,
prod.comuna_id, prod.prov_id   order by visitantes desc, prod.empresario_id
";

globalMap.put("tMysqlInput_1_QUERY", dbquery_tMysqlInput_1);
/**********************************************************************/
```