COLEGIO DE POSGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

POSTGRADO EN SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA

CÓMPUTO APLICADO

DISEÑO DE UNA BODEGA DE DATOS PARA EDUCACIÓN BÁSICA

JOSÉ RAFAEL DURÁN MORENO

TESIS

PRESENTADA COMO REQUISITO PARCIAL

PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS

MONTECILLO, TEXCOCO, ESTADO DE MÉXICO

2018

CARTA DE CONSENTIMIENTO DE USO DE LOS DERECHOS DE AUTOR Y DE LAS REGALIAS COMERCIALES DE PRODUCTOS DE INVESTIGACION

En adición al beneficio ético, moral y académico que he obtenido durante mis estudios en el Colegio de Postgraduados, el que suscribe José Rafael Durán Moreno, Alumno (a) de esta Institución, estoy de acuerdo en ser participe de las regalías económicas y/o académicas, de procedencia nacional e internacional, que se deriven del trabajo de investigación que realicé en esta institución, bajo la dirección de la Profesora Dra.Yolanda Margarita Fernández Ordóñez, por lo que otorgo los derechos de autor de mi tesis Diseño de una bodega de datos para educación básica

y de los productos de dicha investigación al Colegio de Postgraduados. Las patentes y secretos industriales que se puedan derivar serán registrados a nombre del colegio de Postgraduados y las regalías económicas que se deriven serán distribuidas entre la Institución, El Consejero o Director de Tesis y el que suscribe, de acuerdo a las negociaciones entre las tres partes, por ello me comprometo a no realizar ninguna acción que dañe el proceso de explotación comercial de dichos productos a favor de esta Institución.

Montecillo, Mpio. de Texcoco, Edo. de México, a 27 de noviembre de 2018

Firma del
Alumno (a)

Dra. Yolanda Margarita Fernández Ordóñez

Vo. Bo. de la Consejera o Directora de Tesis

La presente tesis, titulada: Diseño de una bodega de datos para educación básica, realizada por el alumno: José Rafael Durán Moreno, bajo la dirección del Consejo Particular indicado, ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

MAESTRO EN CIENCIAS

CÓMPUTO APLICADO

CONSEJO PARTICULAR

CONSEJERO:

DRA. YOLANDA MARGARITA FERNÁNDEZ ORDÓÑEZ

ASESOR:

DR. DAVID HEBERT DEL VALLE PANIAGUA

ASESOR:

DRA. ANTONIA M. MACEDO CRUZ

Montecillo, Texcoco, México, noviembre de 2018

“DISEÑO DE UNA BODEGA DE DATOS PARA EDUCACIÓN BÁSICA”

José Rafael Durán Moreno, M. en C.

Colegio de Postgraduados, 2018

RESUMEN

Esta investigación estudia y resalta las ventajas de una estructura de almacenamiento de información digital tipo bodega de datos (“datawarehouse"), como apoyo a los procesos de producción de información en el Instituto Nacional Para la Evaluación de la Educación (INEE) en México.

El sector educativo enfrenta importantes problemas en todos los niveles de la educación básica, tales como el acceso a la escolaridad obligatoria de niños y jóvenes, la asistencia a la escuela, el avance educacional y la deserción de alumnos. El resultado esperado de este proyecto de investigación es que al diseñar una bodega de datos o datawarehouse (DW) que almacene datos históricos como se propone, se mejore la producción de indicadores requeridos para abordar problemas del sector. La tesis propone mostrar que el acceso más intuitivo, eficaz y libre de errores en la construcción de indicadores contribuye sustancialmente a mejor toma de decisiones en todas las instancias interesadas. La fuente oficial de datos para el proyecto se concentra en el Sistema de Estadísticas Continuas Formato 911 de la Secretaría de Educación Pública. Los datos de este repositorio se conjuntan en registros administrativos provenientes de todas las escuelas de educación básica del país en cada ciclo escolar. Este proyecto propone cómo abordar el análisis, modelación e integración de datos del SECF911 para su incorporación en una bodega de datos en el INEE. En particular, la modelación de datos facilitará considerablemente la comprensión de las nomenclaturas de elementos que en su estado original son crípticas para el usuario analista.

Palabras clave: Bodega de datos, bases de datos, INEE, ETL, estructura de datos

“DESIGN OF A DATAWAREHOUSE FOR BASIC EDUCATIONAL”

José Rafael Durán Moreno, M. en C.

Colegio de Postgraduados, 2018

ABSTRACT

This research studies and highlights the advantages of a storage structure of digital information as a datawarehouse, as the means to support the information production processes in the National Institute for Education Evaluation (INEE -Instituto Nacional Para la Evaluación de la Educación) in Mexico.

The educational sector faces important problems at all levels of basic education, such as access of children and youth to compulsory education, school attendance, scholar advancement and school quitting. The expected result of this research project is that upon designing a data warehouse to store historical data as proposed, the production of indicators to address educational sector problems will improve. The thesis intends to show that a more intuitive, efficient and error free access to information contributes to improved decision making in all concerned instances.

The official source of data for this project is the one gathered by the Mexican Education Ministry within the continuous statistics system through the Format 911 (SECF911). The data in this repository are gathered from administrative registries from all schools of the basic educational system of the country for each scholarly cycle. This project shows how to approach the analysis, modelling and integration of data from the SEC911 format for inclusion in a datawarehouse for INEE. In particular, the data modelling will considerably ease the comprehension of the cryptically naming of items which in their original state are difficult to interpret by the users.

Key words: data warehouse, data bases, INEE, ETL, data structure.

DEDICATORIA Y AGRADECIMIENTOS

Especial dedicatoria a mis padres, quienes siempre me han apoyado incondicionalmente, gracias por sus consejos, fortaleza para continuar con lo que desempeño día a día y sobre todo por mantener junto conmigo las metas que me he fijado a lo largo de mi vida.

A mis hermanos, quiénes han sido participes de cosas importantes en conjunto, además de apoyo incondicional.

Gracias al Instituto Nacional para la Evaluación de la Educación, lugar donde adquirí gran experiencia laboral y de donde surge la idea de este trabajo, en especial al Dr. Héctor Robles, M.C. Mónica Pérez, M.C. Raúl René Rojas, quienes realizaron la gestión correspondiente para la obtención de los datos requeridos para el desarrollo de este trabajo.

Agradezco al Colegio de Postgraduados por haberme brindado la oportunidad de recibir una formación profesional, que con su especializada planta docente forman alumnos con recursos valiosos para el desempeño profesional y al CONACYT por el apoyo en el desarrollo de esta etapa profesional.

Gracias a quiénes que con sus contribuciones como ciudadanos hacen posible el financiamiento para llevar a cabo este tipo de estudios, no me queda más que estar en disposición de ayudar a quienes así lo requieran.

A los miembros del consejo particular, gracias por sus puntuales observaciones y contribuciones para que en medida de lo posible este trabajo se llevase a cabo.

A mis profesores, culmino satisfactoriamente esta etapa, ese conocimiento que nos transmiten permite afrontar más retos en esta vida, de ustedes no solo aprendí en el aula, si no también fuera de ella para ser mejor como persona cada día.

CONTENIDO

LISTA DE FIGURAS

LISTA DE CUADROS

Capítulo 1 INTRODUCCIÓN

El INEE[1] es un organismo cuya misión es ofrecer a las autoridades educativas nacionales, en particular a la Secretaría de Educación Pública, herramientas para evaluar el sistema y subsistemas educativos. Los datos con que cuenta el instituto se utilizan cotidianamente para construir indicadores educativos que permiten conocer la situación de la educación nacional. Esto apoya la aproximación a conocimiento detallado de problemas y retos que enfrenta el sector educativo y contribuyen a resolverlos.

El INEE se creó para ofrecer a las autoridades educativas nacionales, en particular a la Secretaría de Educación Pública, herramientas idóneas para evaluar diferentes elementos de los sistemas educativos. Los datos que se suministran al INEE se utilizan cotidianamente para construir indicadores educativos, los cuales se distribuyen a distintas organizaciones públicas y privadas interesadas en conocer la situación de la educación nacional, con objeto de aproximarse al conocimiento detallado de retos que enfrenta del sector educativo.

El manejo de información es una práctica muy común en distintas organizaciones, para el caso de este trabajo se considera el INEE, donde se procesan datos de educación básica, nivel medio superior y superior. Estos datos son utilizados para calcular indicadores.

El denominado Panorama Educativo de México (PEM[2]) es un producto generado anualmente a partir de la fuente original que es importante difundir. El PEM permite conocer diversos aspectos del estado de la educación en el país. Algunos de los indicadores contenidos en el PEM permiten desarrollar investigaciones no solo del aspecto educativo del país, ampliando eventualmente otras líneas de investigación, tales como las relacionadas con aspectos sociales de la educación o con la interpretación de los resultados mostrados.

El INEE cuenta con un portal[3] que muestra algunos indicadores educativos y tablas, por ejemplo, los índices relacionados con edad, como número de alumnos que cursan algún nivel con

[1] https://www.inee.edu.mx
[2] https://www.inee.edu.mx/index.php/publicaciones-micrositio
[3] http://www.inee.edu.mx/indicadores_/index.html

cierta edad, de ahí se clasifican los que cursan en edad idónea y los que presentan algún tipo de rezago según los años de atraso (ver cuadro 1:1)

Cuadro 1 Porcentaje de alumnos en rezago grave. Fuente Panorama Educativo de México 2013

Porcentaje nacional de alumnos en rezago grave y avance regular para el grupo de edad 15 a 17 años (2011-2012, 2012-2013 y 2013-2014)

Ciclo escolar	Rezago grave	Avance regular
2011-2012	9.4	90.6
2012-2013	8.4	91.6
2013-2014	7.9	92.1

Con la construcción de una bodega de datos podría desarrollarse una aplicación web que mejorase la presentación de la información y, sobre todo, que facilitase al usuario analista el proceso de búsqueda para unificar y decantar datos. Esto sería una de las consecuencias deseables de contar con una organización más apropiada de los datos origen que la que actualmente se utiliza. Los conceptos e informaciones necesarias para construir una bodega de datos se presentan en la sección 2.5.1.

La acumulación de datos en el INEE ha alcanzado un gran tamaño, gracias a las nuevas tecnologías que recolectan y almacenan cada vez volúmenes más grandes. Esta acumulación resulta del seguimiento histórico, que organizada de manera adecuada tendría la posibilidad de facilitar el análisis de distintos escenarios, incluso predictivos, de situaciones que podrían prevalecer en años venideros en el importante ámbito educativo. En el almacén de datos el proceso de actualización consiste exclusivamente en agregar valores recientes a las variables almacenadas. Una bodega de datos es un almacén donde se acumulan datos que por ser históricos ya no se alteran con el tiempo. El contar con datos históricos permite realizar estudios y análisis que apoyen la toma de decisiones más acertadas.

El procesamiento de datos con distintas fuentes de origen motivó al desarrollo de este trabajo, buscando reemplazar el uso del modelo convencional. En esta forma de operar se

consolida la información en cada procedimiento requerido, tareas solicitadas continuamente son ejecutadas de manera repetitiva.

1.2 Objetivos

Uno de los objetivos es proponer una estructura de datos que permita el manejo más intuitivo de la información, debido a la interacción de multiusuarios. Esta estructura de datos permitirá recuperar los datos de manera eficiente. Se hace énfasis sobre la estructura y se omiten otras etapas en la consolidación de este proceso, por ejemplo, la validación de datos, la cual es realizada por otras instituciones. Posterior a la estructura de datos se desarrolla una interfaz para la generación de consultas específicas y reportes. Esta interfaz no requiere de una formación especializada respecto al manejo de datos, está orientada solo a la explotación de la información.

La difusión de indicadores educativos se realiza a través de la edición impresa del PEM y en la página web oficial del instituto, sin embargo, puede mejorarse el acceso a la información de modo que los usuarios interesados en consultar los indicadores los encuentren de una manera más sencilla, teniendo la aplicación web para consulta y con ello incrementar la difusión y tener un mayor volumen de consulta a los resultados que emite la Dirección General para la Integración y Análisis de la Información (DGIAI).

Se tienen distintas fuentes de información por lo que se debe recurrir a un proceso de consolidación de los datos cada vez que se tienen un requerimiento de información o consulta, es decir, se opera de manera tradicional.

No existe un modelo de bases de datos o estructura que permita el fácil manejo y acceso a los datos.

En el año 2016 se desarrolló el proyecto Portal de indicadores y estadísticas educativas que tuvo como objetivo facilitar y agilizar el acceso de indicadores que se presentan en el Panorama Educativo de México a los diferentes usuarios que exploran la información, que son tomadores de decisiones o algunos solo consultan. En este portal que se desarrolló se presentan algunos resultados, pero en algunos casos el usuario no logra obtener lo que busca, sino debe construirlo o calcularlo a partir de todos los datos o un subconjunto de ellos.

La consolidación de los datos consiste en construir una sola tabla o base de datos con información que provienen de distintas fuentes.

En educación básica se consideran los niveles preescolar, primaria y secundaria. Cada nivel tiene tipos de servicio, que son educación inicial, general, indígena, comunitario en preescolar; general, comunitario e indígena en primaria y en secundaria se cuenta con técnica, general, telesecundaria, para trabajadores y comunitaria. Cada entidad federativa cuenta con un archivo con formato dbf por cada nivel escolar y tipo de servicio (mencionado anteriormente), estos supuestamente deberían tener una estructura homogénea o similar entre ellos. Entonces para obtener, por ejemplo, el número de alumnos en nivel primaria, se deben unir todos los archivos por tipo de servicio del nivel primaria (3) y a su vez poder identificar cada uno, posteriormente se buscan aquellas variables que contienen la información sobre número de alumnos y así finalmente obtener la cifra para un determinado ciclo escolar o cualquier otro tipo de desagregación.

Este tipo de procedimiento ejemplifica lo que debe realizar un analista cuando se presenta algún requerimiento de información usando el mencionado sistema tradicional. A partir de esa forma operacional un tanto tortuosa, es que se tuvo la motivación de buscar otra manera de facilitar el manejo de datos. No debiera contarse con analistas expertos para poder aprovechar el contenido o realizar investigación, sino a través de recursos tales como los que ofrece el SQL (Structured Query Language[4]) (Figura 1.1)

```
#MANIPULACION DE DATOS

SELECT      ---> SELECT * FROM TABLA1;
INSERT INTO ---> INSERT INTO TABLA1 VALUES (1,'NOMBRE');
DELETE      ---> DELETE FROM TABLA 1 WHERE ID=5;
UPDATE      ---> UPDATE TABLA1 SET NOMBRE='NOMBRE_NUEVO' WHERE ID=10;
```

Figura 1.1 Ejemplo de sentencias SQL

[4] Lenguaje estándar de bases de datos relacionales

Ante la situación actual y los problemas de manejo de información descritos en la sección anterior, este documento aborda los siguientes puntos que se analizaron y que se plantean como una propuesta de solución en el resto del documento.

Se hizo una revisión de literatura sobre la evolución del manejo de información a partir de los años 60. Se resumen en el capítulo 2.

En el capítulo 3 se revisan las nociones fundamentales de las bodegas de datos y la pertinencia de considerar esta estructura para abordar el tratamiento de datos que se realiza en el INEE.

En el capítulo 4 se menciona la necesidad de asegurar la calidad de datos previamente a su inserción en la bodega de datos del INEE.

En el capítulo 5 se analizan los procesos mediante los cuales se construyó la bodega de datos del INEE, que incluyen las tareas dentro de procesos tales como la extracción y transformación, consolidación y centralización de los datos del INEE. Se muestra la metodología propuesta desde los procesos de importación de datos hasta la carga de los insumos a un gestor de base de datos.

Capítulo 2 REVISIÓN DE LITERATURA

2.1 De los datos al conocimiento

En la década de los 60´s las computadoras tenían limitaciones en cuanto al volumen de procesamiento y la capacidad de procesamiento. Las aplicaciones denominadas de tipo administrativo donde se procesan volúmenes de datos, se restringían al almacenaje y extracción puntual básica mediante reportes pre-programados, por lo tanto no fácilmente no modificables. No se generaban datos tan masivos como en la actualidad.

En los años 80 con el surgimiento de las bases de datos relacionales emanadas de la propuesta de E.F. Codd de organizar conceptualmente la información a partir de tablas relacionales el mundo del procesamiento de datos tuvo un gran salto (Silberscahtz et al, 1990). Ciertamente existían sistemas computarizados de manejo de archivos pero cada uno funcionaba bajo sus maneras propias de organizar los datos. El acceso era un tanto tortuoso y difícil a base de apuntadores que construían listas enlazadas para acceder a los datos y producir información. El modelo relacional que propuso E. F. Codd unificaba en una visión conceptual los acervos de datos, dejando a los implementadores de software el cómo se haría el acceso. Surgieron los primeros sistemas gestores de bases de datos (SGBD), el primero de ellos en el System R de IBM. Enseguida aparecieron otros sistemas y eventualmente se unificó también la manera de acceso que el modelo relacional propuso mediante el álgebra relacional. Surge entonces como estándar que implementa el álgebra relacional el SQL o Structured Query Language promoviendo la popularización de las bases de datos relacionales debido a los logros en el acceso y a las oportunidades que se abrieron.

Actualmente hay sistemas de software como R, SAS ®, Python, MariaDB, MySQL®, PostgreSQL que permiten manipular millones de registros, esto gracias a la simplicidad lógica de

los modelos de bases de datos. Un modelo de bases de datos prescribe cómo deben estar organizados los datos, qué reglas deben satisfacer para cuidar su integridad, facilitar su manejo y tener coherencia en sus resultados ("SAS, R, or Python Survey 2016," 2016).

Cada tipo o modelo de bases de datos tiene una estructura particular para almacenar los datos y forma de operar. Una base de datos es un conjunto de datos estructurados y definidos a través de un proceso específico, que busca evitar la redundancia, y que se almacenará en algún medio de almacenamiento masivo, como un disco (Reinosa et al., 2012).

La computación en la nube o cloud computing, es aquella en la cual se ofrecen los recursos informáticos como servicios a través de internet sin que los usuarios tengan conocimiento de la infraestructura que hay detrás (Joyanes, 2012).

Diversas áreas de trabajo cuentan con grandes bases de datos, de las cuales resulta útil extraer conocimiento que sirva de ayuda al usuario final. En esta dirección, las técnicas adaptativas son muy importantes en estos sistemas ya que permiten crear un modelo de la información contenida en la base de datos, que represente el conocimiento actual, con capacidad suficiente para adaptarse automáticamente a los cambios de la información disponible, sin necesidad de realizar toda la tarea de extracción de conocimiento nuevamente (Hasperué, 2013).

La toma de decisiones ha ganado popularidad en distintos dominios, apareciendo distintas técnicas que ayudan a las personas encargadas de la toma de decisiones.

Estas técnicas se incorporan en los denominados sistemas de soporte a decisiones. Una solución aceptable para estos sistemas debería poder ofrecer al usuario final distintas alternativas para realizar la toma de decisiones y de ser posible de manera online. Luego, a partir de la información obtenida debería actuar de acuerdo con experiencias pasadas, recordando cuales fueron respuestas efectivas y filtrando las indeseadas.

En muchas situaciones, el método tradicional de convertir los datos en conocimiento consiste en un análisis e interpretación realizada de forma manual. El especialista en la materia analiza los datos y elabora un informe o hipótesis que refleja las tendencias o pautas de estos. Por ejemplo, un grupo de médicos analiza la evolución de enfermedades infecto-contagiosas entre la población para determinar el rango de edad más frecuente de las personas afectadas. Este conocimiento es usado por la autoridad sanitaria competente para establecer políticas de

vacunación. Esta forma de actuar es lenta, cara y altamente subjetiva. De hecho, el análisis manual es impracticable en dominios donde el volumen de los datos crece exponencialmente. Consecuentemente, muchas decisiones importantes se realizan, no sobre la base de los datos disponibles sino siguiendo la propia intuición del usuario al no disponer de las herramientas necesarias. Existen herramientas analíticas que han sido empleadas para analizar los datos y tienen su origen en la estadística.

El objetivo principal de este trabajo es cambiar el sistema tradicional por el uso de una bodega de datos. De manera sucinta, una bodega de datos es una captura o repositorio de datos provenientes de distintas fuentes con fines analíticos y de acceso, la cual cuenta con una estructura multidimensional, es decir, se tiene una tabla central de hechos y una o más tablas de dimensiones a través de las cuales se puede filtrar o manipular la información almacenada en la tabla de hechos. Se puede definir una base de datos como una tabla o conjunto de tablas relacionadas bajo un contexto específico, en el capítulo 2 y 3 se mencionan más conceptos sobre bases de datos relacionales (Figura 2.1)

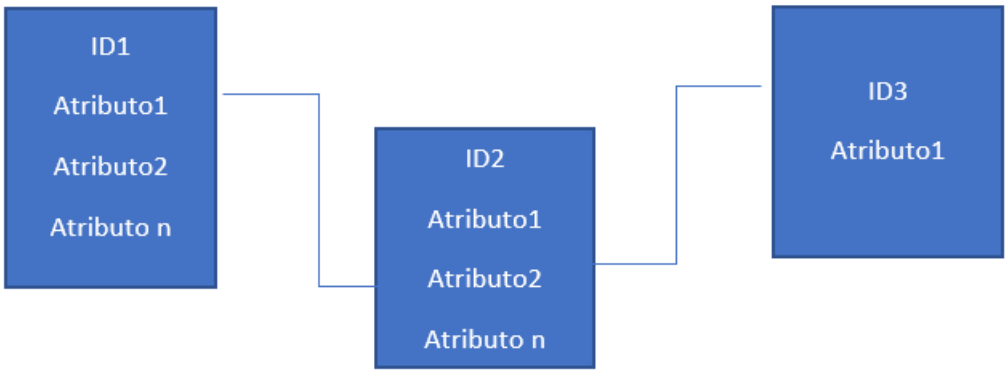

Figura 2.1 Modelo de bases de datos relacional

Una bodega de datos cuenta con una estructura multidimensional, debido a que se tienen dos grandes estructuras principales, la tabla de hechos y las tablas de dimensiones, estas permiten la creación de cubos y que las consultas hechas al sistema gestor de bases de datos (SGBD) sean más rápidas.

La tabla de hechos contiene datos cuantitativos que serán empleados para analizar la bodega de datos y responder a las necesidades o requerimientos previamente establecidos. Estos hechos pueden ser filtrados, manipulados o agrupados a través de condiciones establecidas en las tablas de dimensiones. Cabe mencionar que además de contener datos cuantitativos, la tabla de hechos aloja las claves primarias de las tablas de dimensiones. Con estas claves se conforma la clave primaria de la tabla de hechos, resultando ser una clave compuesta.

Las tablas de dimensiones alojan datos cualitativos, a través de los cuales se puede filtrar o manipular la información almacenada en la tabla de hechos (Bernabeu, 2017).

Los datos a procesar contemplan todas las escuelas del país, de este modo se conocen las cifras de número de alumnos, docentes y escuelas a nivel nacional, un ejemplo de filtrado de información es el número de alumnos en nivel primaria con sostenimiento público en el Estado de México. Esto se logra con las dimensiones, nivel educativo, tipo de sostenimiento y ubicación geográfica.

La bodega de datos se compone de la tabla de hechos y las tablas de dimensiones, sin embargo, existen algunas variantes respecto a su diseño relacional. Algunos de los modelos más utilizados son estrella y copo de nieve. Estos modelos de datos se definen a detalle en el capítulo 3.

Dentro de la bodega de datos se cuenta con los siguientes tipos de estructura:

- Básica: archivos planos que proporcionan datos en bruto que se almacenan junto con metadatos. Los usuarios finales pueden acceder a ellos para su análisis, generación de informes y minería.

- Básica con un área de ensayo: ésta proporciona un lugar donde los datos se pueden limpiar antes de entrar en el almacén. Es posible personalizar la arquitectura del almacén para diferentes grupos dentro de la organización

- Básica con área de ensayo y data marts: diseño particular, se puede tener data marts separados para distintos objetivos, los usuarios finales pueden acceder a datos de uno o todos los data marts de la organización.

En un principio los almacenes de datos estaban diseñados para datos estructurados y no podían mezclarse con los no estructurados con fines analíticos, hoy día, es posible.

La bodega de datos se caracteriza por ser:

- **Integrada**: los datos deben integrarse en una estructura consistente, es decir, se eliminan las inconsistencias entre los diversos sistemas operacionales. El proceso que permite la consolidación, se denomina integración de datos y se compone por distintas tareas, una de las más reconocidas es el proceso ETL (Extracción, Transformación y Carga).

- **Temática**: solo se integran los datos necesarios para el proceso de generación de conocimiento o extracción, esto se especifica cuando se fijan los requisitos, necesidades y objetivos de la organización.

- **Histórica**: se cargan los datos que toman las variables en el tiempo, lo que permite hacer comparaciones entre distintos periodos.

- **No volátil**: existe este almacén de datos para ser leído y no modificado. La información es permanente, lo que implica que la actualización no es más que la incorporación de los últimos valores que tomaron las distintas variables contenidas en él.

La información puede ser vista como un cubo, es decir, la interacción de tres aspectos o dimensiones, por ejemplo, en este caso, zona geográfica, nivel educativo, información a visualizar (alumnos, profesores, escuelas)

2.2 Conceptos de Bases de datos

En esta sección se mencionan algunos conceptos de bases de datos relacionales, qué es una base de datos, cómo se conforma y algunos procesos que se pueden aplicar a las bases de datos. Se hace referencia a solo dos modelos de datos multidimensionales, estrella y copo de nieve, que son los contemplados en el presente trabajo.

Una base de datos es un conjunto de datos estructurados y definidos a través de un proceso específico, que busca evitar la redundancia y que se almacenará en algún medio masivo.

La seguridad en la base de datos es un tema muy importante. En la mayoría de los casos son funciones manejadas por una sola persona, quien es el responsable del acceso de los usuarios y los alcances de acceso que se tendrán

La información que se observa dentro de la mecánica funcional de un problema que se solucionará, se representa por una combinación de estructuras que satisfagan como solución al problema. Esta información que está sustentada por un modelo diseñado con la función de solucionar la problemática observada conforma la estructura de la base de datos. Esta base de datos, con sus estructuras internas, será pensada para que cumpla con las formas normales. Esto último asegura que los datos mantienen su estado relacional, su dependencia y que no habrá pérdida de información respecto de lo observado en el problema (Reinosa et al., 2012)

Sistema gestor de bases de datos

Consiste en una colección de datos interrelacionados y un conjunto de programas para acceder a dichos datos. El objetivo principal de un SGBD es proporcionar una forma de almacenar y recuperar la información de una base de datos de manera que sea tanto práctica como eficiente (Silberschatz, 2014)

Base de datos

Una base de datos es un conjunto de datos estructurados y definidos a través de un proceso específico, que busca evitar la redundancia, y que se almacena en algún medio de almacenamiento masivo, como un disco. Existen diversos modelos de bases de datos, se usa el que satisfaga las necesidades de la organización. Anteriormente predominaba el uso de las bases de datos relacionales, que contienen datos estructurados y almacenados en arreglos tabulares, actualmente ha incrementado la implementación de las bases de datos no relacionales. Estas utilizan una variedad de modelos de datos para acceder y administrar datos, como gráficos, documentos, datos en memoria, además de que están optimizados para aplicaciones que requieren grandes volúmenes de datos.

Modelo de datos

Un modelo de datos brinda distintos conceptos y permite definir las reglas y las estructuras para el almacenamiento de datos para después manipularlos. Hace referencia a una representación de la realidad, ya que un modelo no necesariamente es una ecuación, sino la abstracción de algún problema en específico.

En los modelos basados en registros, los datos se estructuran en un conjunto de campos que conforman un registro. Tal es el caso de un registro "Estudiante" conformado por los campos como apellido, nombres, fecha y lugar de nacimiento.

Base de datos relacional

El concepto fundamental, en el modelo relacional, es que los datos se representan de una sola manera, en el nivel de abstracción que es visible al usuario, y es, específicamente, como una estructura tabular-conformada por filas y columnas-o como una tabla con valores, estas estructuras se relacionan entre ellas mediante unos indicadores.

Relación

No es más que una representación en dos dimensiones, o de doble entrada, constituida por filas, tuplas y columnas o atributos. Dentro del diseño de la base de datos, las relaciones representan a las entidades que se modelaron. Las entidades poseen atributos que las distinguen y cada uno de ellos está logado a un dominio en particular. Por ejemplo, la entidad "Escuela"

Fila o tupla

La tupla en una relación es una colección ordenada de elementos diferentes. Cada tupla es una única combinación de estos elementos; por ende, las tuplas no se repiten dentro de la misma relación.

Columna o atributo

Es una propiedad que caracteriza a cada entidad, como el color o tamaño de un artículo, el apellido de un estudiante, etcétera. Los elementos que componen la estructura de la entidad se denomina atributos y serán irrepetibles en la entidad. Puede ser denominado como columna, campo, atributo o variable. Cada columna representa un tipo de datos, puede ser numérico, carácter, fecha, hora, moneda u otros datos.

Dato

Es la mínima unidad que se almacena en una relación, indivisible en el concepto original de modelo almacenado en la intersección de una fila y de una columna.

Cardinalidad

Así se denomina al número de tuplas o filas de una relación. Es dinámica, ya que depende del agregado o la eliminación de filas a través del tiempo.

Dominio

Es el conjunto de valores posibles de un atributo en la relación, puede ser un conjunto cerrado, donde solo puede tomar un valor de un conjunto, un valor dentro de un rango numérico, entre otros.

Clave primaria

Es la clave candidata que el diseñador elige para identificar a cada entidad que se almacenará en una base de datos relacional. Es decir que la clave primaria cumple con las condiciones exigidas a la clave candidata, esto es unicidad y minimalidad.

Clave foránea

Es un atributo, o combinación de atributos, que permite la combinación de datos de las distintas relaciones del sistema (Reinosa et al, 2012)

2.3 Álgebra relacional

Es un lenguaje de consulta procedimental. Consta de un conjunto de operaciones que toman como entrada una o dos relaciones y producen como resultado una nueva relación.

Las operaciones fundamentales del álgebra relacional son selección, proyección, unión, diferencia de conjuntos, producto cartesiano y renombramiento. Las operaciones se expresan en el lenguaje estándar SQL.

2.3.1 Operaciones fundamentales

Las operaciones antes mencionadas son denominadas unarias, debido a que solo requieren u operan con una sola relación.

2.3.1.1 Operación selección

Selecciona filas que satisfacen un predicado dado.

2.3.1.2 Operación proyección

Es una operación unaria que devuelve su relación de argumentos, excluyendo algunos argumentos. Dado que las relaciones son conjuntos, se eliminan todas las filas duplicadas (Silberschatz, 2014

2.4 Structured Query Language (SQL)

El lenguaje estructurado de consultas apoya a la creación y mantenimiento de la base de datos relacional y a la gestión dentro de una base de datos. Es de sintaxis universal, que permite la recuperación de información, creación de bases de datos, tablas, filtrado de información.

2.4.1 Lenguaje de definición de datos (LDD)

Proporciona órdenes para la definición de esquemas de relación, borrado de relaciones, creación de índices y modificación de esquemas de relación.

2.4.2 Lenguaje de manipulación de datos

Incluye un lenguaje de consultas, basado tanto en el álgebra relacional como en el cálculo relacional de tuplas. Incluye también órdenes para insertar, borrar y modificar tuplas de bases de datos.

2.5 Bodega de datos

En este capítulo se presenta en detalle cómo se estructura una bodega de datos, para en seguida en el capítulo 5 mostrar la propuesta central de este trabajo, que es cómo se conforma la base de datos para el INEE.

2.5.1 ¿Qué es una bodega de datos?

Conviene considerar los antecedentes de las bodegas de datos antes de entrar en definiciones. Diversas instituciones sobre todo aquéllas que colectan grandes cantidades de datos se han interesado en maneras eficientes de extracción de información histórica. Generalmente los volúmenes importantes de información se recaban cotidianamente en bases de datos relacionales o en otros medios como los sistemas tabulares de Excel. Estas formas de los repositorios no permiten extracción de informaciones resumidas para análisis de tendencias, patrones o reportes resumidos de valores de variables seleccionadas dinámicamente. Sin embargo la necesidad de obtener resúmenes, frecuentemente como indicadores construidos bajo demanda de usuarios, ha dado lugar a propuestas de otras formas de organizar y de tener acceso los datos para obtener resúmenes de manera expedita y con minimización de errores. Las instituciones bancarias y comerciales ejemplifican este tipo de necesidades ya que a través de sus bases de datos transaccionales, cotidianamente hacen crecer los volúmenes de datos. Las instituciones financieras en muchos países aportan a instancias de sus gobiernos indicadores necesarios para

la supervisión, regulación e implementación de políticas monetarias y económicas. Estos indicadores han venido integrándose en lo que se conoce como inteligencia de los negocios. Los tomadores de decisiones en diversos tipos de organizaciones usan este tipo de resúmenes de información para conocer el estado real de sus actividades, analizar tendencias, proponer escenarios y decidir sobre el curso de sus acciones a nivel corporativo. Esto se engloba en el concepto de inteligencia del negocio.

Pero no son únicamente las organizaciones comerciales las que enfrentan esta necesidad de partir de sus datos hacia la información y eventualmente al conocimiento que permita tomar decisiones. Es el caso de organizaciones como el Instituto Nacional de la Educación que recaba grandes volúmenes de datos al final de cada ciclo escolar a todos los niveles de educación básica en México, y a la cual se le solicitan cotidianamente indicadores que permitan aproximarse a la situación de distintos aspectos de la educación y contribuir a analizar problemas que enfrenta el sector educativo. Los conceptos fundamentales del manejo de información en bases de datos y en bodegas de datos de presentan a continuación.

Es una base de datos corporativa que replica los datos transaccionales una vez seleccionados, depurados. La función de la bodega de datos es aislar los sistemas operacionales de las necesidades de información para la gestión, de forma que cambios en aquéllos no afecten a éstas y viceversa, únicamente cambiarán los mecanismos de alimentación, no la estructura y contenidos (Reinosa et al., 2012)

Organiza y orienta los datos desde la perspectiva del usuario final. Muchos sistemas operativos organizan sus datos desde la perspectiva del negocio, para mejorar la rapidez de acceso y actualización de los datos (Abella et al., 2016)

La mayoría de los almacenes de datos contienen información histórica que se retira con frecuencia de los sistemas operativos porque ya no es necesaria para las aplicaciones operacionales y de producción. Por la necesidad de administrar tanto la información histórica como las actuales, una bodega de datos es mayor que las bases de datos operacionales (Reinosa et al., 2012)

Una bodega de datos es una base de datos con estructura multidimensional. William Harvey Inmon reconocido mundialmente como el padre del datawarehouse[5] (bodegas de datos), menciona: *"Un datawarehouse es una colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte del proceso de toma de decisiones de la gerencia".*

Es orientada al negocio porque solo ingresarán datos relevantes para el análisis y toma de decisiones y es integrada porque implica que todos los datos provenientes de orígenes heterogéneos deben ser analizados a fin de asegurar su calidad y limpieza para luego ser consolidados en la bodega de datos.

2.5.2 Características de una bodega de datos

Se menciona que un almacén de datos es una base de datos con estructura mutidimensional, debido a que se compone de una tabla de hechos y una o más tablas de dimensiones, que es conformado por datos de distintas fuentes y estructura heterogénea. El almacén de datos permite utilizar los datos para contribuir en la toma de decisiones.

También puede ser entendido como un almacén de datos, pero si fuese estrictamente este concepto se conservarían problemas como en los centros de información. Se caracteriza por ser: integrado, temático, variante en el tiempo y no volátil.

2.5.3 Integridad

Es integrado porque requiere de una consolidación de los datos, el proceso que permite esta consolidación se denomina integración de datos y cuenta con diversas técnicas y subprocesos para llevar a cabo sus tareas. Una de estas técnicas es el proceso ETL (Extracción, Transformación y Carga de datos)

A continuación, se listan los orígenes de datos más comunes:

[5] Término en inglés que hace referencia a bodega de datos.

Producidas por tipo de usuarios

Operacional: produce datos diariamente, en gran cantidad, muchos de los cuales son poco relevantes para el análisis por sí mismos, la granularidad de estos datos es muy fina, ejemplo, venta de productos.

Medio: utiliza los datos operacionales para producir otros datos nuevos que tienen implicancia a corto-medio plazo, ejemplo, control de stock, requerirá una o varias compras a fin de realizar el abastecimiento.

Gerencial: utiliza datos altamente procesados, en general, este perfil de usuario será el destinatario de la bodega de datos, siempre produce retroalimentación que permite generar nueva información para el análisis.

Producidas por áreas o departamentos de la organización

Cada sub área produce sus propios datos, estos serán compartidos con otras áreas.

Producidas por diferentes fuentes de datos

Fuentes internas: datos que genera la empresa en sus actividades diarias

Fuentes externas: datos que complementan y suplementan los datos internos, por ejemplo, datos climáticos, análisis de tendencia, estadísticas, censos, (Bernabeu, 2017)

2.5.4 Temático

Solo se integran los datos necesarios para el proceso de generación de conocimiento o de su extracción.

2.5.5 Variante en el tiempo

En la bodega de datos, la información actual es almacenada junto a los datos históricos y cada dato tiene su respectivo identificador en el tiempo, a través del cual se podrá tener acceso a diferentes versiones o estado de la misma situación. Esto permite conocer la situación de la

organización en el pasado y presente, para casos futuros también es posible con el uso de la modelación, pero sobre todo saber el porqué de cada uno de los escenarios analizados.

2.5.6 No volátil

La información solo será útil para el análisis y esto contribuirá a la toma de decisiones, teniendo como requisito principal una estabilidad, es decir, una vez que los datos ingresan, no cambian. En este entorno solo existen dos acciones posibles:

- **Insertar:** esta acción realiza de forma programada los procesos de integración de datos
- **Consultar:** es la única acción que los usuarios pueden realizar sobre la base de datos.

2.6 Componentes de una bodega de datos

Se tiene una base de datos multidimensional que tiene dos componentes principales, la tabla de hechos y las tablas de dimensiones. Las tablas y su organización dependerán principalmente del modelo que se haya elegido.

2.6.1. Hechos

Tabla central que contiene las variables cuantitativas para realizar análisis, además contiene las claves primarias de las tablas de dimensiones, las cuales conforman la clave primaria de la tabla de hechos.

2.6.2 Dimensiones

Tablas que nos permiten filtrar y manipular la información, están relacionadas directamente con la tabla de hechos. Para esta investigación, se contempló un modelo estrella, es decir, solo existen conexiones directas entre la tabla de hechos y cada una de las tablas de dimensiones. El tener dimensiones anidadas implica trabajar con una estructura diferente que podría complicar la ejecución de consultas o acceso a la información.

2.7 Modelos de datos para bodega de datos

Existen varios modelos de datos, la elección depende de los objetivos planteados, organización de la información, niveles de normalización dentro de las bases de datos, por ejemplo, para una base de datos en la tercera forma normal, se tendrá un modelo copo de nieve, donde las tablas de dimensiones además de tener una relación directa con la tabla de hechos, tienen relaciones adicionales con más tablas.

2.7.1 Modelo estrella

Una tabla de hechos en el centro conectada a un conjunto de tabla de dimensiones. El modelo estrella consolida hechos con relación a unas dimensiones o filtros. Se debe tener un modelo desnormalizado. Se implementa un diseño relacional, en donde la tabla de hechos representa la tercera forma normal y las dimensiones ilustran la segunda forma normal (Figura 2.2)

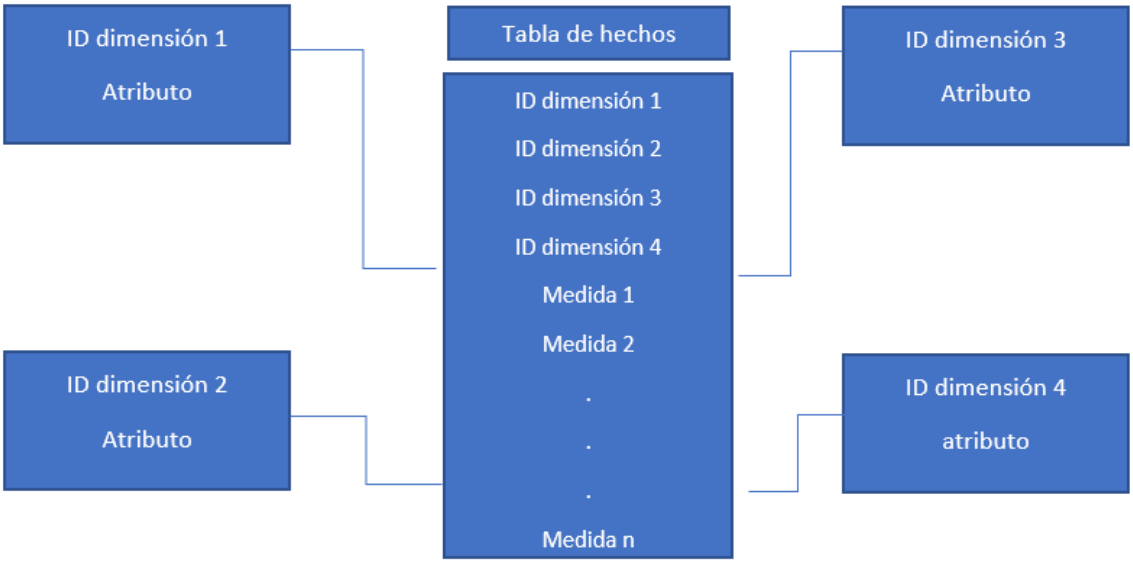

Figura 2.2 Modelo estrella. Elaboración propia

2.7.2 Modelo Copo de nieve

Una variación del esquema estrella en donde alguna jerarquía se normaliza en un conjunto de tablas de dimensiones más pequeñas, la precisión está orientada en facilitar el

mantenimiento de dimensiones. En este modelo las tablas de dimensiones representan relaciones normalizadas en su tercera forma. (Figura 2.3)

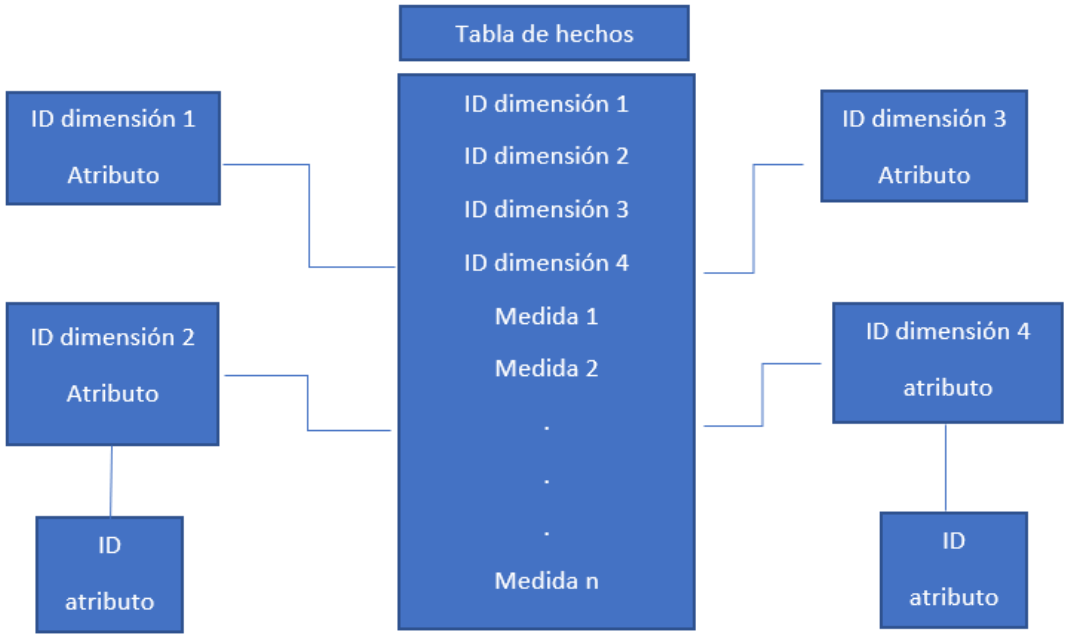

Figura 2.3 Modelo copo de nieve. Elaboración propia

2.8 Conclusiones

El uso de las bases de datos permite crear representaciones de ciertos casos de estudios o fenómenos, que a través del uso de los modelos ayuda a entender la abstracción de los problemas.

Procesar la información con estructuras ya estudiadas, contribuye a la disminución de los tiempos de procesamiento de los datos. Es importante conocer o establecer cuáles son los requerimientos a las necesidades de una organización, con base a esas necesidades se propone el uso de software especializado que conlleva a realizar tareas específicas, esto hace mención a que muchas veces se procesa información con software inadecuado como lo es el caso de Excel, que solo es una hoja de cálculo y no tiene un uso eficiente para la elaboración de reportes que dependen de grandes volúmenes de datos. Para el manejo de datos que siguen algún modelo

previamente creado, se recurre a un SGB. El uso de un SGBD permite almacenar y recuperar la información de una base de datos de manera que sea práctica y eficiente, además que la integración de SQL con su lenguaje de definición y manipulación de datos permite la implementación de algún modelo especifico, ya que ese lenguaje crea las entidades y relaciones que contempla el modelo

Las características de los datos en uso indican qué modelo resulta mejor para su procesamiento. Cuando se tienen datos históricos, que no sufren modificaciones después de ser colectados o almacenados (no volátiles) y además provienen de distintas fuentes, entonces el uso de una bodega de datos es lo adecuado para su manejo, ya que la bodega de datos a través de una estructura dimensional permite facilitar la recuperación de información y con esto automatiza procesos que facilitan la generación de conocimiento.

Capítulo 3 CONSTRUCCIÓN DE LA BODEGA DE DATOS PARA EDUCACIÓN BÁSICA

Antes de consolidar la bodega de datos, debe verificarse la calidad de datos, que consiste en asegurar su consistencia, evitar redundancia en la información, limpiar los datos para que presenten una coherencia o lógica, no basta con almacenar, ya que, de no tener calidad de datos, no se realizarán contribuciones adecuadas en el proceso de toma de decisiones

3.1 Procesamiento previo de datos

Permite llevar la información a condiciones óptimas para su explotación. Se aplican ciertos lineamientos fijados por el administrador de la base de datos, ya que esto permite estandarizar toda la información para reducir errores de captura en los campos.

Tipos de procesos de limpieza

Eliminación de caracteres especiales, espacios, elementos duplicados.

3.2 Validación de datos

Proceso de corrección, ejecutar un procesamiento previo, debido al estado de su origen, pueden tener inconsistencias. Deben cumplirse ciertas reglas lógicas, que los datos tengan una coherencia, esto tiene una relación directa con la definición de los valores posibles de cada campo.

3.2.1 Formato de datos por campo

Puede llevarlo o no, en ocasiones es mejor omitir caracteres y homologar la información como dejar todo en minúsculas, sin acentos, etc. Posteriormente se puede asignar formato como nombre propio (primera letra de cada palabra en mayúscula y el resto en minúsculas) o solo la primera letra en mayúscula.

3.2.2 Homologación de información

Después del proceso de limpieza se puede homologar datos que tienen pequeñas variantes en escritura, pero hacen referencia a lo mismo. Puede generarse una estructura similar en los distintos orígenes de datos.

3.3 Información de calidad

El modelo de calidad de datos representa cimientos sobre los cuales se construye un sistema para la evaluación de un producto de datos. En un modelo de calidad de datos se establecen las características de calidad de datos que se deben tener en cuenta a la hora de evaluar las propiedades de un producto de datos terminado

La Calidad del Producto de Datos se puede entender como el grado en que los datos satisfacen los requisitos definidos por la organización a la que pertenece el producto. Son precisamente estos requisitos los que se encuentran reflejados en el modelo de Calidad de Datos mediante sus características (Exactitud, Completitud, Consistencia, Credibilidad, Actualidad, Accesibilidad).

Las características de Calidad de Datos están clasificadas en dos grandes categorías:

- Calidad de Datos Inherente: Se refiere al grado con el que las características de calidad de los datos tienen el potencial intrínseco para satisfacer las necesidades establecidas y necesarias cuando los datos son utilizados bajo condiciones específicas. Desde el punto de vista inherente, la Calidad de Datos se refiere a los mismos datos, en particular a:
    - Valores de dominios de datos y posibles restricciones (Reglas de Negocio gobernando la calidad requerida por las características en una aplicación dada).
    - Relaciones entre valores de datos (Consistencia).
    - Metadatos.
- Calidad de Datos Dependiente del Sistema: Se refiere al grado con el que la Calidad de Datos es alcanzada y preservada a través de un sistema informático cuando los datos son utilizados bajo condiciones específicas. Desde el punto de vista dependiente del sistema,

la Calidad de Datos depende del dominio tecnológico en el que los datos se utilizan, y se alcanza mediante las capacidades de los componentes del sistema informático tales como: dispositivos hardware (Respaldo Software para alcanzar la Recuperabilidad), y otro software (Herramientas de migración para alcanzar la Portabilidad).

Esta estructura de datos facilitara el acceso, análisis y exploración de los datos. Su manejo resultara más intuitivo a comparación del sistema convencional que es usado actualmente.

3.4 Definición de requisitos

Para el caso del INEE, los datos más relevantes son aquellos que permiten calcular los indicadores que se publican en el PEM. A continuación, se muestran algunos:

AT01 ¿Cuántos niños y jóvenes se matriculan en educación básica o media superior?

AT02 ¿Cómo avanzan los alumnos en su trayectoria escolar?

AT03 ¿Cuántos alumnos de una generación escolar terminan oportunamente cada nivel educativo?

AT04 ¿Cuántos alumnos de una generación escolar terminan oportunamente cada nivel educativo?

CS01 ¿Cuáles son las poblaciones objetivo de la educación básica, media superior y para adultos?

CS02 ¿Cuál es la asistencia de la población infantil y juvenil a la educación básica y media superior?

CS03 ¿En qué medida el sistema educativo cubre la necesidad social de educación?

CS04 ¿Cómo es el contexto socioeconómico en que opera el sistema educativo nacional?

CS05 ¿Cómo es el rezago en la escolarización?

PG01 ¿En cuántas escuelas de educación básica todos los docentes atienden más de un grado?

PG02 ¿La organización de las escuelas por zonas posibilita su atención adecuada y oportuna?

AR01 ¿Cuáles son las características de los alumnos, docentes y directores de educación básica y media superior?

AR02 ¿Existe una disponibilidad mínima de recursos informáticos dedicados a la enseñanza en las escuelas de educación básica?

AR03 ¿Cuánto gasta el estado y la sociedad en la formación integral de la población, especialmente en educación obligatoria?

Adicional a esto existen requerimientos de información aún más específicos, cuya información no siempre está disponible en el subconjunto de los datos más utilizados, es por ello que se partió de dos grandes subconjuntos, los datos que contribuyen a la elaboración del PEM y los que no. Después de analizar los requerimientos específicos más frecuentes, se generó un tercer subconjunto para agregarlo al primero.

3.5 Datos a procesar

Una bodega de datos es una colección de datos orientada al negocio (Inmon, 2008), esto significa que solo se contemplarán datos relevantes para el análisis y toma de decisiones.

La información que se procesa para crear el PEM es proporcionada por la Secretaría de Educación Pública (SEP). Envía los archivos al INEE en distintos formatos (.csv, .dbf,.sas7bdat), puede ser un archivo por cada tipo de servicio (general, indígena, comunitario, etc.) y nivel (preescolar, primaria, secundaria) o de manera adicional la misma información antes mencionada, pero por entidad federativa.

Los datos son a nivel escuela, cada registro representa una escuela en el país, cuenta con su clave de centro de trabajo, información geográfica, tipo de sostenimiento, cifras sobre alumnos en distintas formas, por edad, grado, sexo, además de información respecto a docentes como totales por escuela, por sexo.

Se realizan dos suministros de datos de manera anual, al inicio y fin de cada ciclo escolar. Estos datos son acompañados de las cifras oficiales y los metadatos para realizar un cotejo de lo que obtuvo la SEP y de lo que obtiene u obtendrá el INEE.

Posterior a la entrega de datos, se realiza un proceso de validación, se toma una muestra de los indicadores para realizar los respectivos cálculos, después de eso se efectúa una comparación con las cifras oficiales que emite la Secretaría de Educación Pública. Si los cálculos coinciden con las cifras oficiales se liberan los archivos para que todos los usuarios tengan acceso.

3.6 Diseño del modelo

El modelo estrella fue el elegido, que consiste en crear una tabla de hechos y nueve tablas de dimensiones, por lo que se construyó una tabla de hechos y nueve tablas de dimensiones. La tabla de hechos contiene las claves primarias de cada una de las tablas de dimensiones, además de las variables de interés que se especificaron en pasos previos. Algunas variables dentro de la tabla de hechos son alumnos, docentes y número de escuelas por entidad federativa, alumnos por grado, edad y sexo, etc.

Se consideraron nueve tablas de dimensiones que contienen información respecto a:

- Entidad Federativa
- Municipio
- Localidad
- Nivel educativo
- Tipo de servicio
- Tipo de sostenimiento
- Tamaño de localidad
- Grado de marginación
- Ciclo escolar

Cada una de las tablas de dimensiones contiene su clave primaria y aquellos campos que permiten el filtrado de la información o construcción de cubos (Figura 3.1)

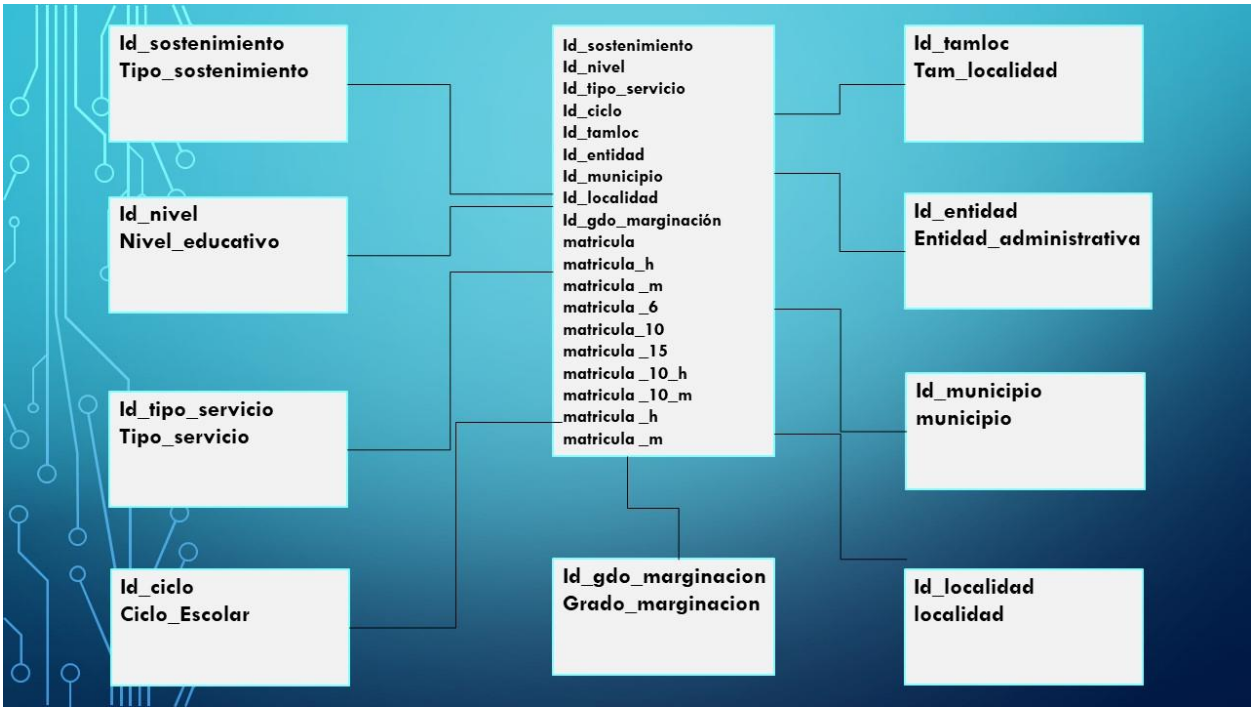

Figura 3.1 Modelo estrella para la bodega de datos INEE

3.7 Elección de software

Después de plantear el modelo recibir los datos, se revisaron los metadatos e identificaron los formatos, se procedió con la elección de software. Los archivos recibidos tienen formato .dbf, un formato propio de bases de datos, el cual puede ser manejado con software como Access de la paquetería Office, también pueden ser procesados con SAS (uno de los softwares más utilizados dentro del instituto). El elegido fue PostgreSQL, que es un sistema gestor de bases de datos relacionales. Esta elección se realizó para alojar las bases de datos en un SGBD como tal, desde este paso se notó una diferencia en el procesamiento de información, ya que se obtienen los resultados más rápidos que en el software utilizado anteriormente (SAS). El procesamiento en SAS era rápido si se tenía una computadora con recursos elevados en memoria RAM y procesador, el manejo con PostgreSQL fue rápido aún en una computadora de bajos recursos computacionales. También se eligió Pentaho, que es una plataforma de código abierto, útil en el proceso de construcción de la bodega de datos. Finalmente se seleccionó R, un software gratuito

que brinda la opción de ser utilizado a través de la escritura de código o a través del uso de interfaces gráficas.

3.8 Definición de medios de conexión

Los datos son almacenados en el servidor del instituto, de este modo solo se ingresarán las rutas para que los programas tengan acceso a la información. Todas las fuentes de datos están disponibles dentro de la organización, por lo que no fue necesario incluir conexiones web, alguna IP externa o permisos adicionales. Para el inicio del desarrollo del proceso de construcción de la bodega de datos, se almacenaron los datos de manera local en la misma computadora donde estaban montados los programas requeridos. Se especificó la ruta de archivos y verificó la conexión con PostgreSQL.

Para trasladar los archivos a PostgreSQL se usó como medio de conexión Pentaho, que con su interfaz gráfica facilitó este paso.

Otro medio de conexión definido fue R con PostgreSQL, esta conexión fue a través de la librería RPostgreSQL de R, que permite manipular los datos desde R, no es necesario crear copias en memoria, ya que se saturaría el software y resultaría obsoleto, solo se crea una conexión y se despliegan resultados en R.

3.9 Extracción, Transformación y Carga (ETL)

En este proceso se realizan tres tareas principales, extracción, transformación y carga. Estas tareas van enfocadas a la manipulación, control, integración, depuración de datos, carga y actualización de la bodega de datos.

3.9.1 Extracción

Proceso enfocado a la obtención de datos relevantes y mantenerlos en un almacenamiento intermedio. Para este caso se utilizó la herramienta Spoon de Pentaho para realizar la comunicación con el SGBD que fue PostgreSQL

Primero se identificó la ruta donde se almacenaron todos los archivos (C:\Users\Admin\Datos\archivos_dbf) (Figura 3.2)

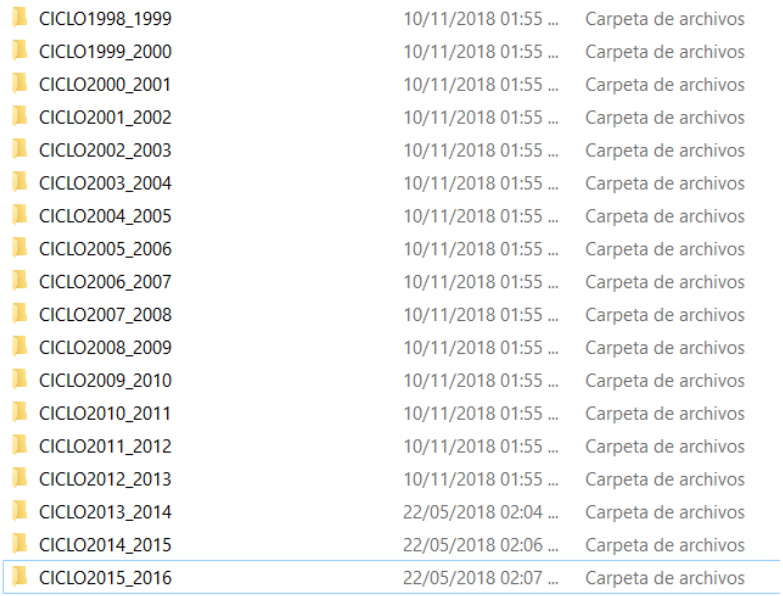

Figura 3.2 Almacenamiento local de archivos

Dentro de cada carpeta se encuentran los archivos .dbf, uno por cada tipo de servicio (Figura 3.3)

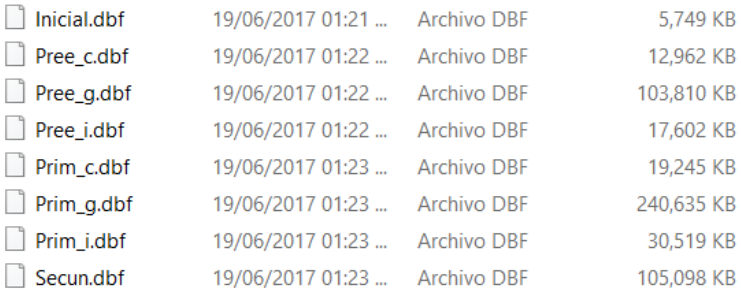

Figura 3.3 Archivos por tipo de servicio en formato dbf

Para cargar las bases de datos a PostgreSQL se utilizó Spoon, cuya herramienta no requiere de un proceso de instalación, solo se descargó el archivo “pdi-ce-8.1.0.0-365” de la liga https://sourceforge.net/projects/pentaho/files/Data%20Integration/ . Después de la descarga se descomprimió el archivo, posteriormente en Windows se abrió shell (CMD) para introducir la ruta hasta la carpeta “data-integration”, ya estando dentro de dicha ruta se tecleó “spoon.bat” para iniciar. Estos pasos pueden ser evitados cada vez que se cargue el proceso, colocando un acceso directo e indicando que ejecute el último comando mencionado. Después de enviar la orden de abrir Spoon.bat, se muestra la siguiente interfaz (Figura 3.4)

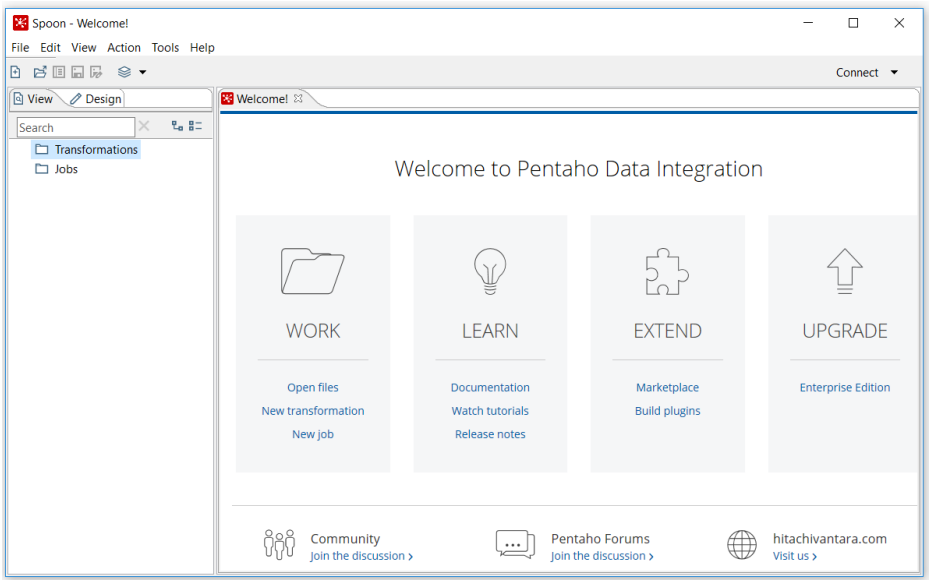

Figura 3.4 Pantalla de inicio de Spoon.

En la parte izquierda presionar en “Transformations” para que abra la siguiente pestaña (Figura 3.5)

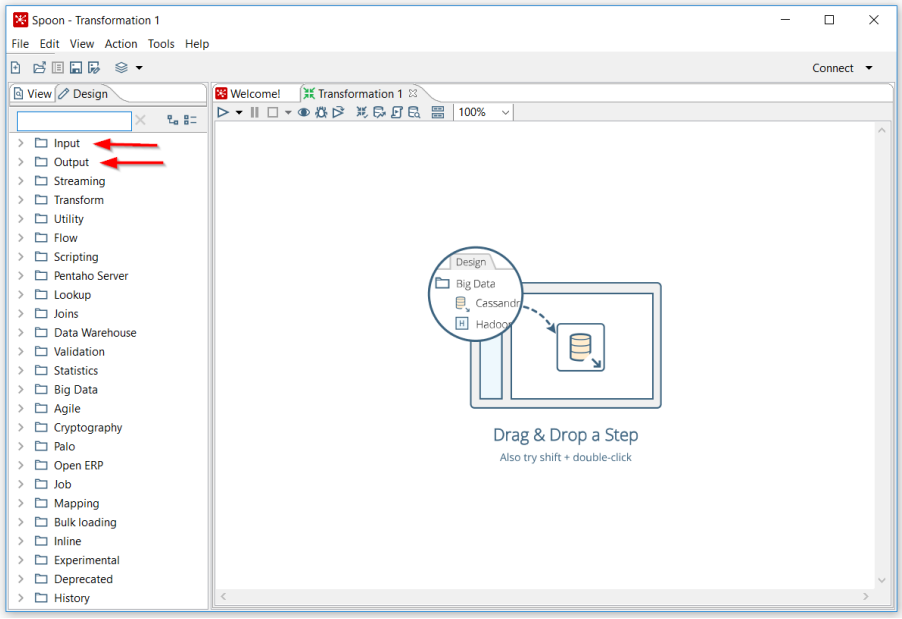

Figura 3.5 Ventana de edición para transformación.

El primer paso fue desplegar la opción input para seleccionar y arrastrar 

hacia la pestaña Transformation. Después de esto se dio doble clic en el ícono para que se mostrara el asistente donde se asigna un nombre a este paso, enseguida se especifica la ruta de los archivos (Figura 3.6)

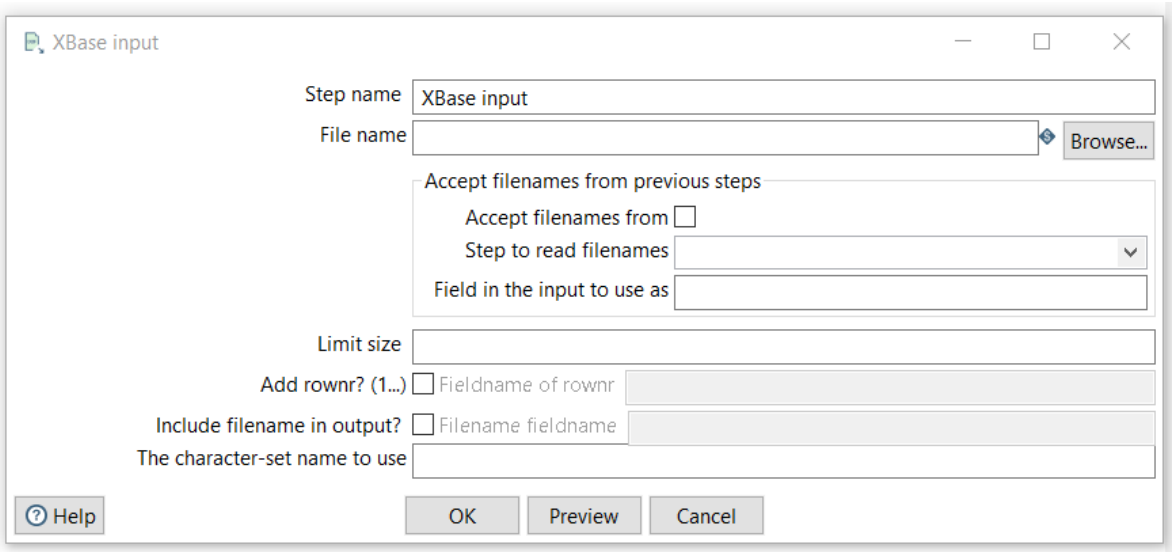

Figura 3.6 Asistente para edición de datos de un objeto XBase en entrada de datos.

Antes de dar clic en aceptar se puede hacer uso de la opción “Preview” donde se puede especificar algún número de registros a visualizar previamente a la carga, esto con el fin de asegurar que no hay errores (Figura 3.7).

Rows of step: input_sec (20 rows)

#	CLAVECCT	N_CLAVECCT	TURNO	N_ENTIDAD	MUNICIPIO	N_MUNICIPI	LOCALIDAD	N_LOCALIDA
1	01DES0001O	LIC. BENITO JUAREZ	1	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES
2	01DES0001O	LIC. BENITO JUAREZ	2	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES
3	01DES0002N	MOISES SAENZ	1	AGUASCALIENTES	001	AGUASCALIENTES	0239	GENERAL JOSE MARIA I
4	01DES0003M	LIC. ADOLFO LOPEZ MATEOS	1	AGUASCALIENTES	007	RINCON DE ROMOS	0001	RINCON DE ROMOS
5	01DES0003M	LIC. ADOLFO LOPEZ MATEOS	2	AGUASCALIENTES	007	RINCON DE ROMOS	0001	RINCON DE ROMOS
6	01DES0004L	IZCOATL	1	AGUASCALIENTES	003	CALVILLO	0001	CALVILLO
7	01DES0004L	IZCOATL	2	AGUASCALIENTES	003	CALVILLO	0001	CALVILLO
8	01DES0005K	IGNACIO ALLENDE	1	AGUASCALIENTES	007	RINCON DE ROMOS	0030	PABELLON DE HIDALGC
9	01DES0006J	AMADO NERVO	7	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES
1	01DES0007I	22 DE OCTUBRE	7	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES
1	01DES0008H	JOSE CLEMENTE OROZCO	1	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES
1	01DES0008H	JOSE CLEMENTE OROZCO	2	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES
1	01DES0009G	CONGRESO DE ANAHUAC	2	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES
1	01DES0009G	CONGRESO DE ANAHUAC	1	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES
1	01DES0010W	LEYES DE REFORMA	1	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES
1	01DES0010W	LEYES DE REFORMA	2	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES
1	01DES0011V	CONVENCION DE AGUASCALIENTES	1	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES

Figura 3.7 Visualización previa de datos especificados en objeto XBase

Después de especificar los datos de entrada, se continuó con la indicación del tipo de archivo de salida. En el apartado de Output se seleccionó la opción 

, que despliega la siguiente interfaz (Figura 3.8)

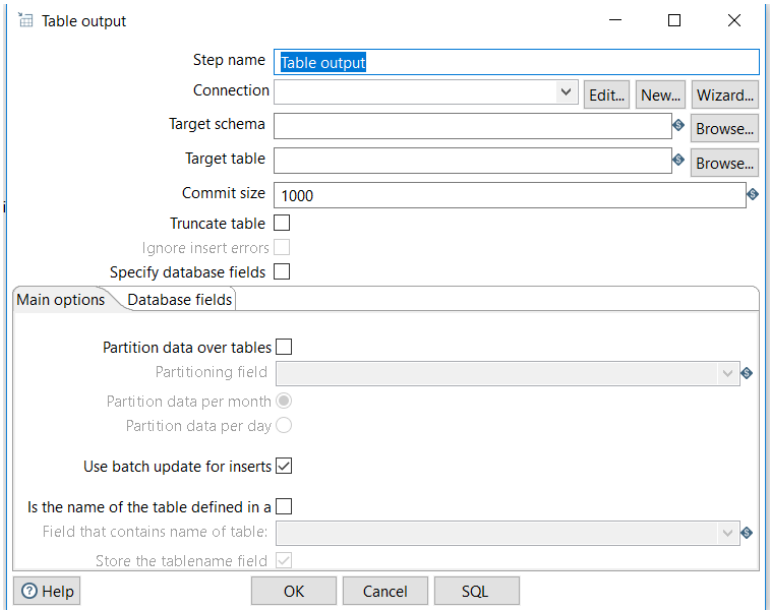

Figura 3.8 Asistente para edición de datos de un objeto Table output en entrada de datos.

En este paso se especificaron más detalles:

- Asignar un nombre a la conexión para poder ser usada posteriormente (**conection_postgres).**
- Especificar el tipo de conexión, en este apartado se eligió la opción **PostgreSQL** de la lista disponible.
- Indicar el nombre del host, para este caso se trabajó con **localhost.**
- Escribir el nombre del base de datos donde se alojarán los datos.
- Introducir el nombre de usuario, que es **postgres** y la respectiva contraseña.

Después de los pasos anteriores se observó lo siguiente (Figura 3.9)

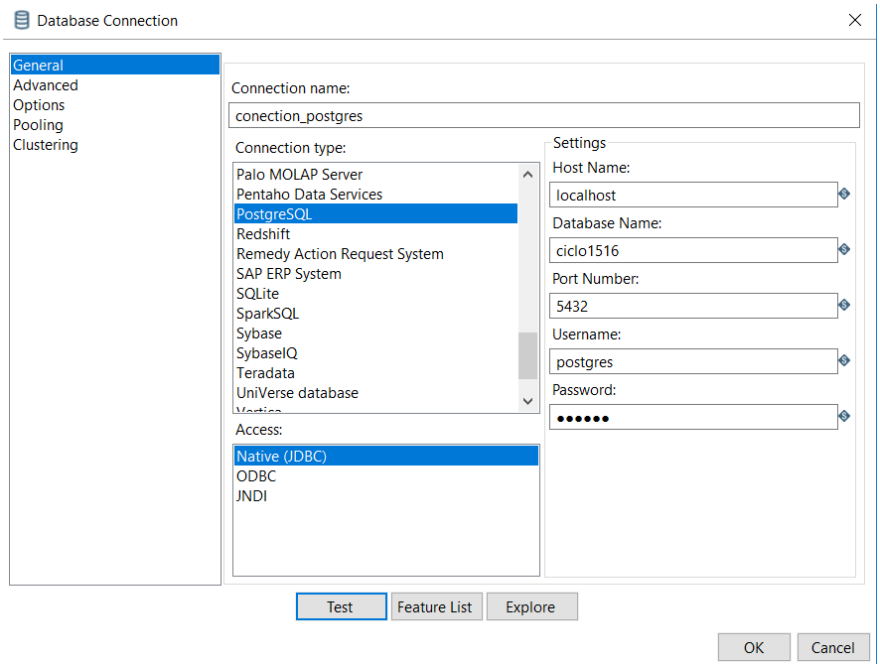

Figura 3.9 Asistente para la creación de conexión a un SGBD.

Como recomendación se utilizó la opción **Test** para verificar que la conexión sea correcta (Figura 3.10)

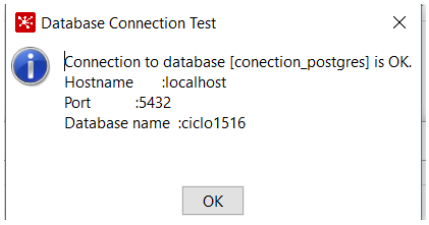

Figura 3.10 Verificación del estatus de conexión

Para que los pasos especificados en input y output tengan una secuencia, basta con poner el cursor sobre cualquiera de los dos objetos y arrastrar la flecha hacia el otro (Figura 3.11)

Figura 3.11 Conexión de un objeto de entrada y uno de salida.

Las opciones utilizadas fueron elegidas para realizar la lectura específica sobre un archivo .dbf y para generar una salida en una tabla que será almacenada en PostgreSQL, estas opciones pueden variar según sea el tipo de archivo a leer y el destino de los datos, como puede ser un archivo de texto plano (.txt), un archivo delimitado por comas (.csv), etc.

La extracción de todos los archivos quedó especificada así (Figura 3.12)

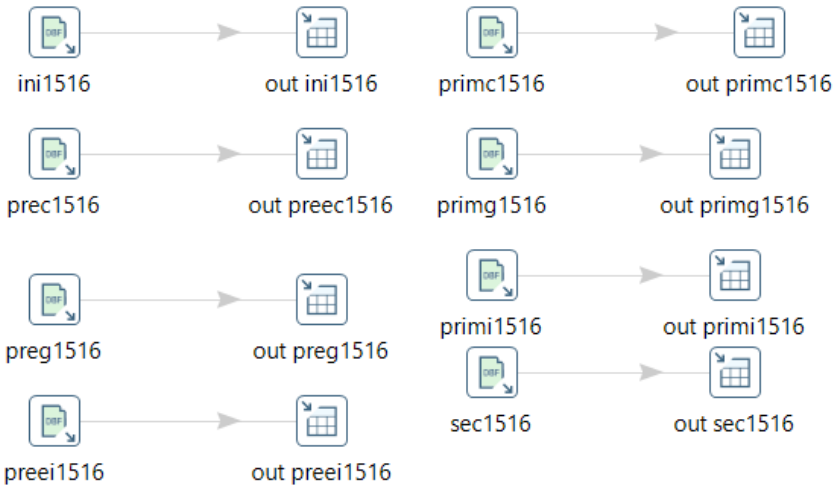

Figura 3.12 Representación de entras y salidas para el ciclo escolar 2015-2016.

Que se almacenó con la extensión. ktr para su posterior uso. Después de ejecutar el procedimiento cada paso debe mostrarse así 

, lo que significa que no existieron errores, además en la venta LOG se observa el mensaje “Spoon - The transformation has finished!!”, con hora y fecha indicada. Para verificar que los archivos han sido cargados se

abre PostgreSQL a través del pgAdmin. Cuando se abre la base de datos de cada ciclo se observa lo siguiente (Figura 3.13).

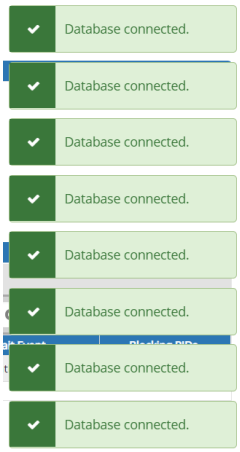

Figura 3.13 Conexión exitosa de la base de datos en PostgreSQL.

Y dentro de cada ciclo se observan las tablas que fueron cargadas con Pentaho (Figura 3.14)

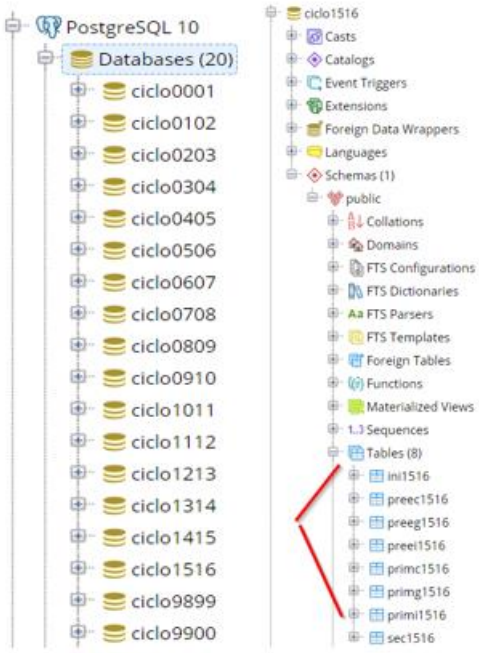

Figura 3.14 Visualización de tablas

Para comprobar que están los datos, se da clic derecho sobre alguna tabla (Figura 3.15), como ejemplo se usó secundaria del ciclo 2015-2016, donde se solicitó observar los primeros 100 registros (Figura 3.16)

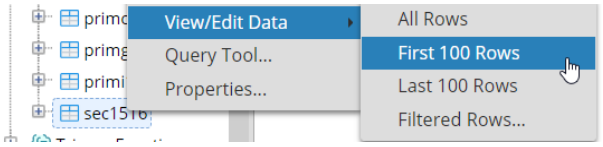

Figura 3.15 Secuencia de pasos para visualización de datos de una tabla en PostgreSQL.

Data Output | Explain | Messages | Notifications | Query History

	clavecct character varying (10)	n_clavecct character varying (100)	turno	n_entidad character varying	municipio character varying (3)	n_municipi character varying (80)	localidad character varying (4)	n_localida character varying (140)
1	01DES0001O	LIC. BENITO JUAREZ	1	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES
2	01DES0001O	LIC. BENITO JUAREZ	2	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES
3	01DES0002N	MOISES SAENZ	1	AGUASCALIENTES	001	AGUASCALIENTES	0239	GENERAL JOSE MARIA MOR...
4	01DES0003M	LIC. ADOLFO LOPEZ MATEOS	1	AGUASCALIENTES	007	RINCON DE ROMOS	0001	RINCON DE ROMOS
5	01DES0003M	LIC. ADOLFO LOPEZ MATEOS	2	AGUASCALIENTES	007	RINCON DE ROMOS	0001	RINCON DE ROMOS
6	01DES0004L	IZCOATL	1	AGUASCALIENTES	003	CALVILLO	0001	CALVILLO
7	01DES0004L	IZCOATL	2	AGUASCALIENTES	003	CALVILLO	0001	CALVILLO
8	01DES0005K	IGNACIO ALLENDE	1	AGUASCALIENTES	007	RINCON DE ROMOS	0030	PABELLON DE HIDALGO
9	01DES0006J	AMADO NERVO	7	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES
10	01DES0007I	22 DE OCTUBRE	7	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES
11	01DES0008H	JOSE CLEMENTE OROZCO	1	AGUASCALIENTES	001	AGUASCALIENTES	0001	AGUASCALIENTES

Figura 3.16 Visualización de datos de una tabla en PostgreSQL

3.9.2 Transformación

La transformación es uno de los pasos más importantes en el proceso, ya que es la fase donde se da paso al modelo propuesto, creando las tablas indicadas en dicho modelo, homologación de la información debido a que algunas fuentes de datos tienen una estructura distinta.

La información puede ser vista como un cubo, es decir, la interacción de tres aspectos o dimensiones, por ejemplo, en este caso, tamaño de localidad, grado de marginación, nivel educativo (preescolar, primaria y secundaria), tipo de servicio (comunitario, indígena, general, etc), tipo de sostenimiento (público y privado), ciclo escolar (1998-1999 a 2015-2016), entidad federativa (administrativa), municipio, localidad. Estas dimensiones ayudan a filtrar y manipular la información (alumnos, profesores, escuelas).

La manipulación de los datos para su respectiva transformación fue realizada con R, conectándolo con PostgreSQL donde ya se encuentran almacenadas las tablas. Para ello se hizo uso de las siguientes librerías:

- RPostgreSQL
- Tidyverse

Primero se creó una conexión con PostgreSQL desde R, con el siguiente script

```
library(RPostgreSQL)#librería para conectar R a PostgreSQL
library(tidyverse)#esta lib contiene a dplyr,es para manipular datos
#cargar driver de postgres
drv<-dbDriver("PostgreSQL")
#crear conexión
con1516<-dbConnect(drv,
          host="localhost",
          port=5432,
          user="postgres",
          password="rafael",
          dbname="ciclo1516")
```

La variable que contiene al número de alumnos, es el primer filtro en los datos, ya que se debe especificar que se contemplen las escuelas con ALUMNOS>0, debido a que en ocasiones se encuentran algunas escuelas sin datos de alumnos y afectarían al conteo de escuelas. El siguiente código muestra cómo se realizó este paso

```
ini<-dbGetQuery(con1516,"SELECT * FROM ini1516 WHERE v132>0")
pree_c<-dbGetQuery(con1516,"SELECT * FROM preec1516 WHERE v19>0")
pree_g<-dbGetQuery(con1516,"SELECT * FROM preeg1516 WHERE v63>0")
pree_i<-dbGetQuery(con1516,"SELECT * FROM preei1516 WHERE v63>0")
prim_g<-dbGetQuery(con1516,"SELECT * FROM primg1516 WHERE v347>0")
prim_i<-dbGetQuery(con1516,"SELECT * FROM primi1516 WHERE v344>0")
prim_c<-dbGetQuery(con1516,"SELECT * FROM primc1516 WHERE v363>0")
sec<-dbGetQuery(con1516,"SELECT * FROM sec1516 WHERE v164>0")
```

Se realizó renombramiento de variables, ya que gran parte de la información viene nombrada como v1, v2, …, vn. Por ejemplo, en preescolar general v63 contiene el número total de alumnos por escuela, en este caso se cambió el nombre a ALUMNOS y así con las demás variables.

El tipo de sostenimiento no aparece en los archivos, deben ser calculados a partir de otras variables. Para el caso de tipo de servicio se extra el tercer carácter de la clave de centro de trabajo (clavecct), en algunos casos se requieren instrucciones adicionales para la asignación de tipo de sostenimiento.

Para el caso del tipo de servicio se hace una codificación a partir de las variables n_renglon y renglon, una contiene valores numéricos y se agrupan con base a un catálogo de claves, adicionalmente se usa n_renglon para complementar la codificación. Se muestra código utilizado (Figura 3.17)

```
ini=ini%>%mutate(sostenimiento=ifelse(substr(ini$clavecct,3,3)=='P','privado','publico'),t_servicio='inicial',nivel='preescolar')
pree_g=pree_g%>%mutate(sostenimiento=ifelse(substr(pree_g$clavecct,3,3)=='P' | n_renglon=='PREESCOLAR GENERAL PARTICULAR','privado','publico'),
                       t_servicio='general',nivel='preescolar')
pree_i=pree_i%>%mutate(sostenimiento=ifelse(substr(pree_i$clavecct,3,3)=='P','privado','publico'),t_servicio='indigena',nivel='preescolar')
pree_c=pree_c%>%mutate(sostenimiento=ifelse(substr(pree_c$clavecct,3,3)=='P','privado','publico'),t_servicio='comunitario',nivel='preescolar')
prim_g=prim_g%>%mutate(sostenimiento=ifelse(substr(prim_g$clavecct,3,3)=='P' | n_renglon=='PRIMARIA GENERAL PARTICULAR','privado','publico'),
                       t_servicio='general',nivel='primaria')
prim_i=prim_i%>%mutate(sostenimiento=ifelse(substr(prim_i$clavecct,3,3)=='P','privado','publico'),t_servicio='indigena',nivel='primaria')
prim_c=prim_c%>%mutate(sostenimiento=ifelse(substr(prim_c$clavecct,3,3)=='P','privado','publico'),t_servicio='comunitario',nivel='primaria')
sec=sec%>%mutate(sostenimiento=if_else(substr(sec$clavecct,3,3)=='P' | n_renglon=="SECUNDARIA GENERAL PARTICULAR",'privado','publico'),
                 t_servicio=if_else(sec$renglon %in% c("0436","0437","0441","0446","0448","0449","0460","0470","0471"),'general',
                                    if_else(sec$renglon %in% c('0490','0491','0498','0499','0510'),'para trabajadores',
                                            if_else(sec$renglon %in% c('0530','0531','0532','0533','0534','0540','0550','0751'),'telesecundaria',
                                                    if_else(sec$renglon %in% c('0570','0571','0572','0576','0578','0579','0590','0610','0615',
                                                                               '0620','0630','0650','0680','0710'),'tecnica',
                                                            'comunitario')))),nivel='secundaria')
```

Figura 3.17 Código para creación de variables a partir de otras contenidas en la tabla.

La instrucción mutate nos permite añadir columnas, las instrucciones anteriores son ejecutadas secuencialmente, cada una separada por %>%. Esta función fue utilizada también para la creación de los nuevos nombres de variables, en un solo paso se hacen cálculos y se renombran aquellas que no dependen de alguna operación. A continuación, se muestra el código utilizado para obtener matrícula por edad, matrícula por edad y sexo (Figura 3.18)

```
#matricula por edad
sec<-mutate(sec,alumnos=v164,mat_11=sum(v1,v18),mat_12=sum(v2,v10,v19,v27,v45,v53,v61,v69),
            #matricula por edad
                mat_13=sum(v3,v11,v20,v28,v46,v54,v62,v70,v86,v93,v100,v107),
                mat_14=sum(v4,v12,v21,v29,v47,v55,v63,v71,v87,v94,v101,v108),
                mat_15=sum(v5,v13,v22,v30,v48,v56,v64,v72,v88,v95,v102,v109),
                mat_16=sum(v6,v14,v23,v31,v49,v57,v65,v73,v89,v96,v103,v110),
                mat_17=sum(v7,v15,v24,v32,v50,v58,v66,v74,v90,v97,v104,v111),
                mat_18ymas=sum(v8,v16,v25,v33,v51,v59,v67,v75,v91,v98,v105,v112),
            #matricula por edad y ysexo
            mat_11_h=sum(v122),
            mat_11_m=sum(v139),
            mat_12_h=sum(v123,v131),
            mat_12_m=sum(v140,v148),
            mat_13_h=sum(v124,v132),
            mat_13_m=sum(v141,v149),
            mat_14_h=sum(v125,v133),
            mat_14_m=sum(v142,v150),
            mat_15_h=sum(v126,v134),
            mat_15_m=sum(v143,v151),
            mat_16_h=sum(v127,v135),
            mat_16_m=sum(v144,v152),
            mat_17_h=sum(v128,v136),
            mat_17_m=sum(v145,v153),
            mat_18ymas_h=sum(v129,v137),
            mat_18ymas_m=sum(v146,v154),
```

Figura 3.18 Código para la creación de variables utilizando funciones.

Para las variables que fueron renombradas no se requieren más operaciones adicionales, pero en el caso de las que fueron utilizadas para generar nuevas variables, deben ser eliminadas porque ya no son útiles en el esquema, para ello se creó un vector con las variables a eliminar, enseguida se hizo una selección omitiendo el vector creado (Figura 3.19)

```
drop.cols<-paste(c("v"),1:757,sep="")  #crear un vector con nombres variabes antiguas
sub2<-sec1%>%select(-one_of(drop.cols)) #borrar variables antiguas y conservar nuevas
```

Figura 3.19 Eliminación de un conjunto de variables.

Para la creación de las entidades se utilizaron las siguientes instrucciones (Figura 3.20)

```
#creacion de la tabla de dimension entidad federativa
dim_entidad<-distinct(select(sec1,n_entidad),n_entidad)
dim_entidad=mutate(dim_entidad,id_entidad=paste(c("ent"),1:32,sep="_"))
dim_entidad

#creacion de la tabla de dimension nivel
dim_nivel<-distinct(select(sec1,nivel),nivel)
dim_nivel=mutate(dim_nivel,id_nivel=paste(c("niv"),1,sep="_"))
dim_nivel

#creacion de la tabla dimensión tipo de servicio
dim_tservicio<-distinct(select(sec1,t_servicio),t_servicio)
dim_tservicio=mutate(dim_tservicio,id_t_servicio=paste(c("tserv"),1:5,sep="_"))
dim_tservicio

#creacion de la tabla dimensión ciclo escolar
aux1<-1998:2015
aux2<-1999:2016
ciclo<-paste(aux1,aux2,sep="-")
id_ciclo<-paste(c("ciclo"),1:18,sep="_")
dim_ciclo<-data.frame(id_ciclo,ciclo)
dim_ciclo
```

Figura 3.20 Creación de tablas de dimensiones.

3.9.3 Carga

Al finalizar la transformación de las distintas variables y creación de las tablas de dimensiones, se cargaron las tablas al SGB, esto con el siguiente código (Figura 3.21)

```
#Cargar tabla de hechos y tablas de dimensiones a PostgreSQL
dbWriteTable(con_load, "dim_entidad",value = dim_entidad)
dbWriteTable(con_load, "dim_municipio",value = dim_municipio)
dbWriteTable(con_load, "dim_localidad",value = dim_localidad)
dbWriteTable(con_load, "dim_nivel",value = dim_nivel)
dbWriteTable(con_load, "dim_tservicio",value = dim_tservicio)
dbWriteTable(con_load, "dim_sostenimiento",value = dim_sostenimiento)
dbWriteTable(con_load, "dim_tam_loc",value = dim_tam_loc)
dbWriteTable(con_load, "dim_gdo_marg",value = dim_gdo_marg)
dbWriteTable(con_load, "dim_ciclo",value = dim_ciclo)
dbWriteTable(con_load, "hechos",value = hechos)
```

Figura 3.21 Proceso de carga de tablas a PostgreSQL

Se especifica la conexión a la nueva base de datos, se crearon dos conexiones, debido a que se almacenarán los datos en bases distintas, una para carga y otra para comunicar al proceso de transformación (Figura 3.22)

```
con_load<-dbConnect(drv,
                    host="localhost",
                    port=5432,
                    user="postgres",
                    password="*****",
                    dbname="dw_ebasica")
```

Figura 3.22 Creación de conexión para la carga de datos.

Para comprobar que las tablas han sido cargadas, se usó el código siguiente (desde R):

dbGetQuery(con_load, "SELECT * from dim_entidad")

Se especificó la conexión y se envió una sentencia SQL. Para verificar que están las tablas en el gestor se actualiza la base de datos y se observan las tablas (Figura 3.23)

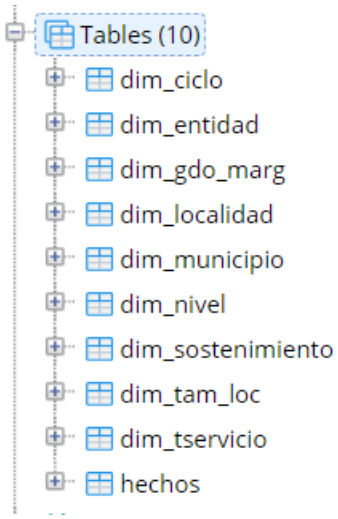

Figura 3.23 Visualización de tablas de dimensiones cargadas a PostgreSQL.

3.10 Interfaz para consulta de información

Después de construir la bodega de datos, se emite el medio de conexión para acceder a ella. Esta estructura permite ser analizada a través del lenguaje SQL, para ello se consulta el catálogo de variables, donde se describen los acrónimos empleados para representar o nombrar a las variables contempladas. A continuación, se muestran algunos ejemplos:

id_entidad: identificador único en la tabla de dimensión entidad, con la cual se pueden hacer consultas junto con la tabla de hechos, respecto a alguna de las 32 entidades federativas o el despliegue de resultados a nivel nacional.

id_nivel: identificador único en la tabla de dimensión nivel, que hace referencia a alguno de sus tres posibles valores, preescolar, primaria, secundaria.

clavecct: identificador de cada escuela, este campo por si solo no resulta ser único, ya que en ocasiones se reporta información de escuelas que cuentan con dos turnos (matutino y vespertino), por lo que la identificación de manera única de una escuela es la unión de la clave de centro de trabajo y el turno.

id_t_servicio: identificador único dentro de la tabla dimensión tipo de servicio, es usado en conjunto con la tabla de dimensión nivel, esto cuando se pretende desplegar resultados por nivel educativo y tipo de servicio.

mat_5: variable referida a la matrícula de alumnos con cinco años de edad

mat_5_h: también es posible obtener datos sobre matricula por edad y sexo, en este caso se muestra el ejemplo de número de alumnos con cinco años de edad y sexo hombre.

mat_h: de manera independiente a la edad se cuenta con variables que proporcionan información con respecto a matricula por sexo, para este caso matricula con sexo hombre.

Posteriormente a consultar el catálogo de variables se pueden realizar procedimientos a través del lenguaje SQL, si el usuario desea conocer alguna cifra solo debe conectarse a la bodega de datos a través de algún SGBD e introducir sentencias SQL.

El fin de este trabajo es presentar una interfaz de consulta, la propuesta de diseño es la siguiente (Figura 3.24)

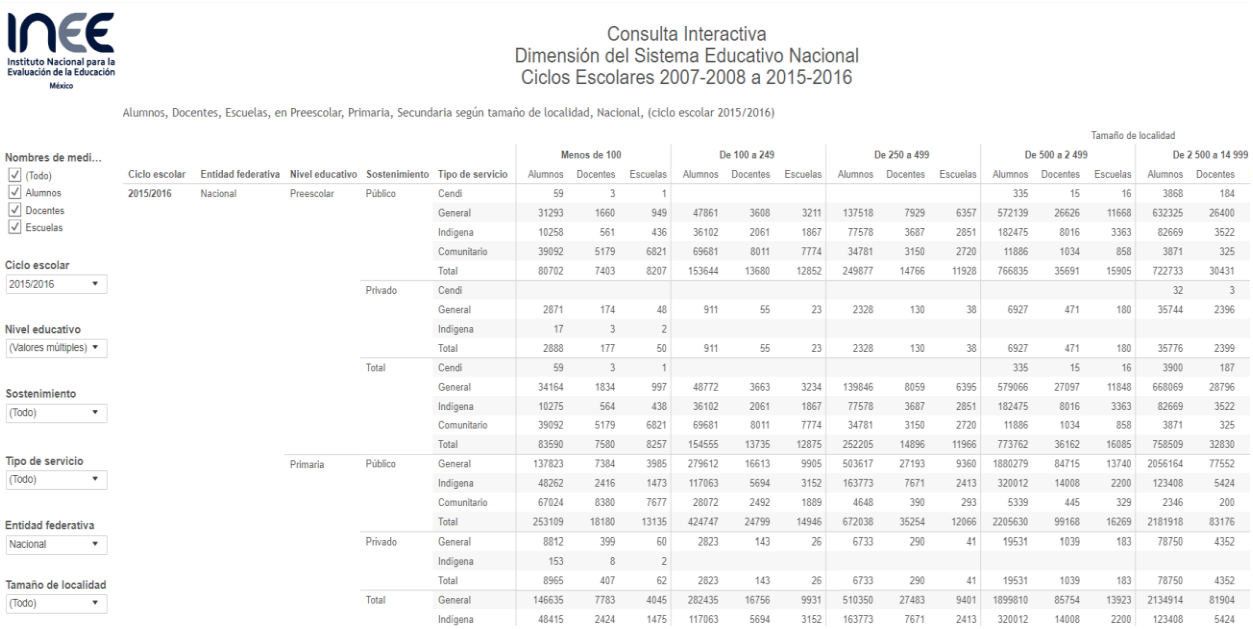

Figura 3.24 Interfaz de consulta

En la figura anterior se muestra la propuesta de interfaz de consulta, donde el usuario no requiere conocimientos previos en programación, basta con presiones algunos botones para el despliegue de resultados.

En este ejemplo se muestran cifras de alumnos, docentes y escuelas a nivel nacional, para los niveles preescolar y primaria en sus tipos de sostenimiento privado y público, además de desplegar los resultados por tamaño de localidad.

Todos los filtros de dimensiones contienen los posibles valores que pueden tomar, además de proporcionar una opción que permita obtener el total de alguna dimensión (Figura 3.25)

Instituto Nacional para la Evaluación de la Educación
México

Consulta Interactiva
Dimensión del Sistema Educativo Nacional
Ciclos Escolares 2007-2008 a 2015-2016

Alumnos, Docentes, Escuelas, en Preescolar & Primaria según tamaño de localidad, Nacional, (ciclo escolar 2015/2016)

Nombres de medi...
- [x] (Todo)
- [x] Alumnos
- [x] Docentes
- [x] Escuelas

Ciclo escolar
2015/2016

Nivel educativo
(Valores múltiples)

Sostenimiento
(Valores múltiples)

Tipo de servicio
(Todo)

Entidad federativa
Nacional

Tamaño de localidad
Total

					Tamaño de localidad		
					Total		
Ciclo escolar	Entidad federativa	Nivel educativo	Sostenimiento	Tipo de servicio	Alumnos	Docentes	Escuelas
2015/2016	Nacional	Preescolar	Público	Cendi	60587	2733	956
				General	3488823	147530	45292
				Indígena	412813	18815	9802
				Comunitario	162029	17905	18654
				Total	4124252	186983	74704
			Privado	Cendi	6035	321	160
				General	681662	43474	14543
				Indígena	17	3	2
				Total	687714	43798	14705
		Primaria	Público	General	12041420	465059	68381
				Indígena	818196	36944	10178
				Comunitario	110366	12138	10511
				Total	12969982	514141	89070
			Privado	General	1280290	60061	8932
				Indígena	153	8	2
				Total	1280443	60069	8934

Figura 3.25 Despliegue de resultados a nivel nacional, por nivel educativo, tipo de sostenimiento y tipo de servicio

Capítulo 4 DISCUSIÓN DE RESULTADOS

Manejar los datos bajo una estructura de bodega de datos permite entender más el contexto de los datos, ya que el proceso de transformación en la creación de la bodega de datos coloca las variables a modo que el usuario comprenda o entienda lo que está trabajando.

Con el modelo implementado el acceso a la información es más rápido (operacionalmente referido al gestor de datos), debido a que no se encuentra normalizada y no existen muchas relaciones entre las tablas.

Procesar los datos con un manejador de bases de datos, disminuye el tiempo de operación sobre la información, de este modo no se consumen muchos recursos computacionales, además en el mejor de los casos utilizar el procesamiento en nube o algún servidor mejorará sustancialmente los tiempos de respuesta en consultas.

Una de las grandes desventajas del uso de bodega de datos es que puede requerir elevados recursos para su mantenimiento, pero la gran ventaja es que un proceso bien implementado evitará muchos errores en el procesamiento, facilitará el acceso a datos, teniendo un impacto positivo en el proceso de toma de decisiones.

4.1 Recomendaciones

La bodega de datos construida es una buena alternativa con respecto al sistema tradicional que es usado actualmente.

Al nivel de alcance del proyecto, los datos pueden ser aprovechados a través de lenguaje de consulta SQL, basta con consultar los metadatos para saber cómo o de qué modo encontrar la información de interés. El avance en la propuesta abarca el proceso completo de ETL, los datos han sido cargados a un gestor de bases de datos,

Como trabajo futuro de este desarrollo, en una siguiente etapa se requiere crear una interfaz gráfica que permite a los usuarios acceder a los datos sin necesidad de programación. Será suficiente presionar algunos botones para el despliegue de resultados, indicadores, tablas, etc.

La recomendación para la interfaz es crearla con el software R y la librería Shiny, la cual está diseñada para la creación de aplicaciones de un modo sencillo, además permite realizar un alojamiento en sus servidores junto con todas las bases de datos que sean requeridas para alimentar la aplicación.

Las aplicaciones web requieren de un hosting o sitio de almacenamiento en nube. Por ejemplo, Shiny ofrece servidores gratuitos, con algunas restricciones en cuanto a tiempo de uso o acceso, número de aplicaciones limitadas, pero también ofrece más ventajas a través de sus opciones comerciales, como mayor interacción de usuarios al mismo tiempo, mayor tiempo de acceso.

Otra recomendación para completar la implementación, si el INEE dispone de espacio en alguno de sus servidores, puede hospedarse ahí la aplicación y no se generarían costos adicionales.

BIBLIOGRAFÍA

Amazon Web Services, 2017. Caso de éxito de AWS: IE Business School.

Artículo 2º. Diario Oficial de la Federación (2003). Estatuto orgánico del Instituto Nacional Para la Evaluación de la Educación. URL:http://www.inee.edu.mx/index.php/normateca-descentralizada/documentos-del-inee/normateca/normas-internas-del-inee/normas-sustantivas-del-inee/presidencia-del-inee/02_Estatuto_Organico_INEE_2nda_modf.pdf/download

Batini C., Ceri S., Navathe S. 2015. Diseño conceptual de bases de datos. Un enfoque de entidades interrelacionales. México. Addison Wesley

Joyanes A.L. (2012). Computación en la nube: estrategias de Cloud Computing en las empresas. Alfaomega, Ciudad de México

Hasperué, W. (2013). Extracción de conocimiento en grandes bases de datos utilizando estrategias adaptativas.

Micrositio de indicadores, INEE. URL: http://www.inee.edu.mx/indicadores_/index.html
Banco de indicadores educativos,
URL: http://www.inee.edu.mx/index.php/bases-de-datos/banco-de-indicadores-educativos

Oppel A., Sheldon R. (2010). Fundamentos de SQL. 3ª edición. México, D.F.

Pressman S. Roger. (2016). Ingeniería de software. Un enfoque práctico (5ed). Madrid. McGraw-Hill

Reed Michael (2013). A definition of a data warehouse. London.

Reinosa, E.J., Alejandro Maldonado, C., Muñoz, R., Esteban Damiano, L., Adrián Abrutsky, M. (2012). Bases de datos. Alfaomega, Buenos Aires.

Robles, H. (2016). Panorama Educativo de México 2015. Ciudad de México. URL: http://publicaciones.inee.edu.mx/buscadorPub/P1/B/114/P1B114.pdf

SAS, R, or Python Survey (2016): Which Tool Do Analytics Pros Prefer?

Silberschatz, A., Stonebraker, M., Ullman, J.D. (1990). Database systems: achievements and opportunities. 1990 ACM SIGMOD Record - Directions for future database research & development: Vol. 19, Iss. 4, December. pp. 6-22.

Silberschatz A., Korth & S. Sudarskhan (2014). Fundamentos de Bases de Datos. 6ta edición. McGraw Hill.

William H Inmon (2018). Building the data warehouse. Indianapolis. Wiley Computer Publishing.