

**INCORPORACIÓN DE INFORMACIÓN AL
ALMACÉN DE DATOS ESTADÍSTICOS**

DESCRIPCIÓN GENERAL DE METODOLOGÍA

PRESENTACIÓN

El presente documento pretende ser obra de consulta sencilla que refiera las etapas a seguir en el proceso de incorporación de información al Almacén de Datos Estadísticos.

Siendo que este documento brinda un panorama general de la metodología; omite los detalles técnicos y la descripción minuciosa que se incluyen en la Metodología de Incorporación de Información al Almacén de Datos Estadísticos.

INTRODUCCIÓN

Las crecientes necesidades de información por parte de usuarios internos o externos, así como la cada vez más acelerada respuesta que requieren en sus consultas, inducen al INEGI a una constante actualización. Data Warehouse (DWH) o Almacén de datos, provee palpables ventajas respecto al almacenamiento de la información del instituto.

En el presente documento se exponen los procesos a seguir para la integración de la información al almacén institucional y la interacción necesaria entre las áreas que generan o integran la información y el área responsable del almacén, con la finalidad de que el almacén sea una útil y eficiente herramienta de análisis de la información estadística en la atención de requerimientos de información y la toma de decisiones.

La organización del documento se basa en los modelos de datos usados en el Almacén de datos: Relacional (Nivel Detalle), ROLAP (Modelo Estrella) y MOLAP (Cubo).

Como inicio se presentan las generalidades del Almacén de Datos Estadísticos.

En la segunda sección se definen los lineamientos y estándares que se deben seguir en el análisis y diseño de los modelos de datos relacionales (Nivel Detalle), así como los procesos para realizar la carga de información; detallando las etapas y actividades que deben observarse para el desarrollo de las herramientas de extracción, transformación y carga (ETC) requeridas para incorporar datos al DWH.

La siguiente sección considera el análisis, diseño y carga de la información para los modelos de datos ROLAP (Estrella), considerando las diferencias de estos modelos con respecto al nivel detallado.

En la última sección se describen las actividades correspondientes a las áreas participantes en el diseño de cubos para análisis MOLAP, así como los pasos a seguir para la generación de los mismos, ya que constituyen uno de los principales medios de explotación de la información del DWH.

ÍNDICE

Almacén de Datos Estadísticos	5
1. ¿Qué es un Almacén de Datos?	6
2. Objetivo del Almacén de Datos Estadísticos	6
3. Visión del Almacén de datos	6
Modelos de Datos Relacional	8
1. Análisis de la información	9
2. Diseño de la Base de Datos	9
3. Carga de Información	9
Diagrama de Análisis	10
Diagramas de Diseño	11
Modelo Lógico	11
Modelo Físico	12
Diagrama de Carga de Información	13
Modelos de Datos ROLAP	15
1. Análisis de la información	16
2. Diseño de la Base de Datos	16
3. Carga de Información	16
Diagrama de Análisis	17
Diagramas de Diseño	18
Modelo Lógico	18
Modelo Físico	19
Diagrama de Carga de Información	20
Modelos de Datos MOLAP	22
1. Diseño de datos	23
Diagrama de Diseño de Datos	24
Glosario	26

ALMACÉN DE DATOS ESTADÍSTICOS

GENERALIDADES

1. ¿QUÉ ES UN ALMACÉN DE DATOS?

Es un conjunto de datos orientado a temas diseñados para realizar tareas de análisis. Combina información de distintas fuentes y su objetivo es presentar a través de herramientas especializadas una vista integral de la organización en un momento determinado para apoyar el proceso de toma de decisiones. Cuenta con las siguientes características:

- Datos integrados y consolidados.
- Orientado al tema (área de interés).
- De tiempo variante (histórico).
- No volátil.
- Manejo de grandes volúmenes de información.
- Estandarización de la manera de trabajar y registrar información.
- Una sola versión de la verdad en toda la organización.
- Diseñado para tener respuesta inmediata a consultas.
- Enfocado a la información que da valor al negocio.

2. OBJETIVO DEL ALMACÉN DE DATOS ESTADÍSTICOS

Reunir y consolidar las bases de datos de información estadística, que se mantienen en las diferentes áreas funcionales del Instituto como subsistemas de información independientes, en un ambiente integral centralizado, para consulta, explotación y análisis que permita a los usuarios tomar mejores decisiones.

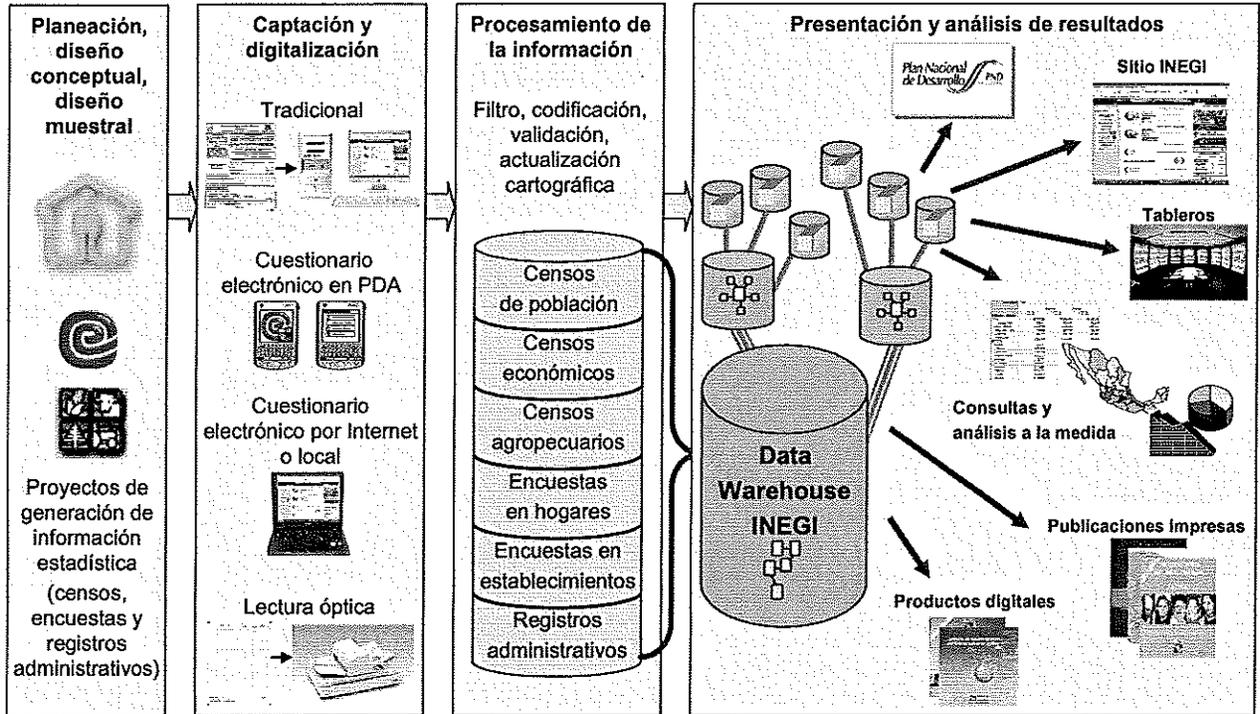
3. VISIÓN DEL ALMACÉN DE DATOS

El flujo de generación de la información que publica el INEGI está conformado por 4 etapas principales:

1. Planeación, diseño conceptual, diseño muestral
2. Captación y Digitalización
3. Procesamiento de la información
4. Presentación y análisis de resultados

La incorporación de información al almacén de datos estadísticos se fija al inicio de la etapa de presentación y análisis de resultados. De esta manera la información ya procesada ingresa al almacén de datos para de allí hacer la presentación y análisis de resultados por medio de las publicaciones, del sitio en Internet de INEGI y de diversas herramientas de consulta y análisis.

En seguida se muestra de manera gráfica el flujo de la información incluyendo la incorporación de información al almacén de datos.



MODELOS DE DATOS RELACIONAL

(NIVEL DETALLE)

1. ANÁLISIS DE LA INFORMACIÓN

Este proceso es la más importante y fundamental en el desarrollo de un Data Warehouse, ya que es aquí, donde se realiza la identificación y revisión de la información a incorporar o los incrementos a la información ya incorporada, con la finalidad de generar la documentación suficiente para la Etapa de Diseño de la Base de Datos.

El análisis comienza con recabando las generalidades del proyecto así como la documentación que se ha generado durante el desarrollo del mismo, esta información será el insumo para hacer la revisión de los datos que se generan del proyecto, al igual que sus metadatos.

Se hace una confrontación del proyecto con la información contenida en el almacén de datos para hacer los ajustes necesarios a fin de facilitar la comparabilidad de la información durante las consultas.

Durante la etapa de análisis se generan documentos cuyas estructuras están predefinidas en la vista detallada de la metodología y que permiten contar con el registro uniforme de los resultados de esta etapa.

2. DISEÑO DE LA BASE DE DATOS

El diseño de la base de datos constituye una labor de suma importancia y detalle, ya que en ella se define la estructura en la que ha de residir la información que se integra al almacén. Dicha estructura se conforma por tablas relacionadas entre sí, además del conjunto de reglas que debe cumplir la información que en ellas reside. El diseño puede hacerse con apoyo de herramientas de software, que lo hacen más confiable y menos laborioso.

Con base a los resultados del análisis de la información se diseña un modelo de datos que refleja los detalles y reglas de la información del proyecto y que habrá de convertirse en un modelo físico que describa las estructuras de almacenamiento, las relaciones entre ellas y métodos que se utilizarán para el acceder y manipular los datos de modo eficiente.

El diseño de la base de datos incluye la generación de la documentación conteniendo el detalle del trabajo realizado.

3. CARGA DE INFORMACIÓN

La información que ha de incorporarse al almacén de datos es comparada con los criterios captados en el análisis para detectar alguna posible inconsistencia, que en caso de existir será corregida a fin de tener los datos preparados para ser cargados en las bases de datos del almacén.

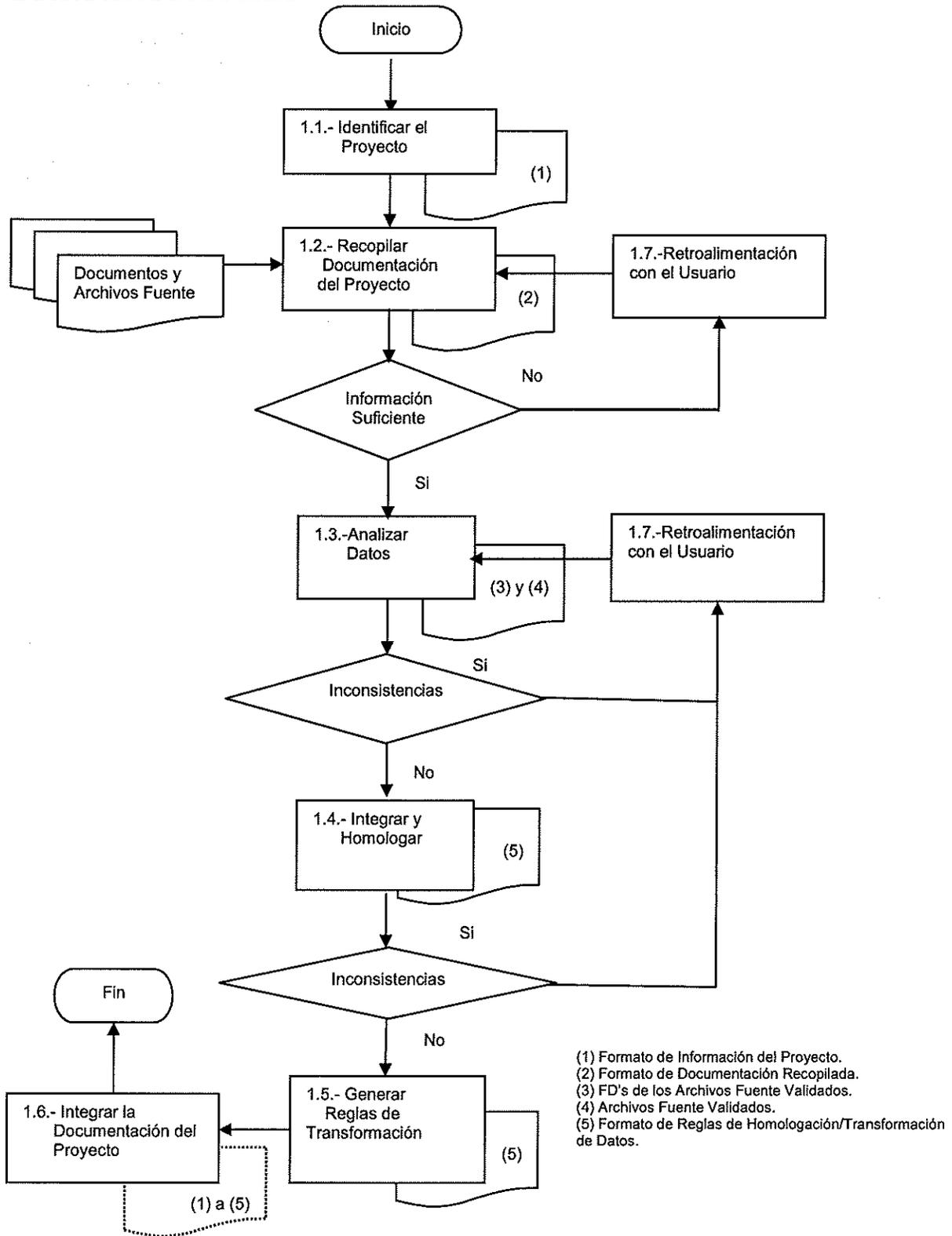
Los formatos en que se recibe la información son variados y de igual forma el tratamiento que se hace para su incorporación, que va desde el desarrollo de programas, hasta la utilización de software comercial que tienen esa finalidad.

La información de un modelo de datos relacional (nivel detalle) se conforma de las tablas de registros y los catálogos que se describen a el contenido de los registros de información.

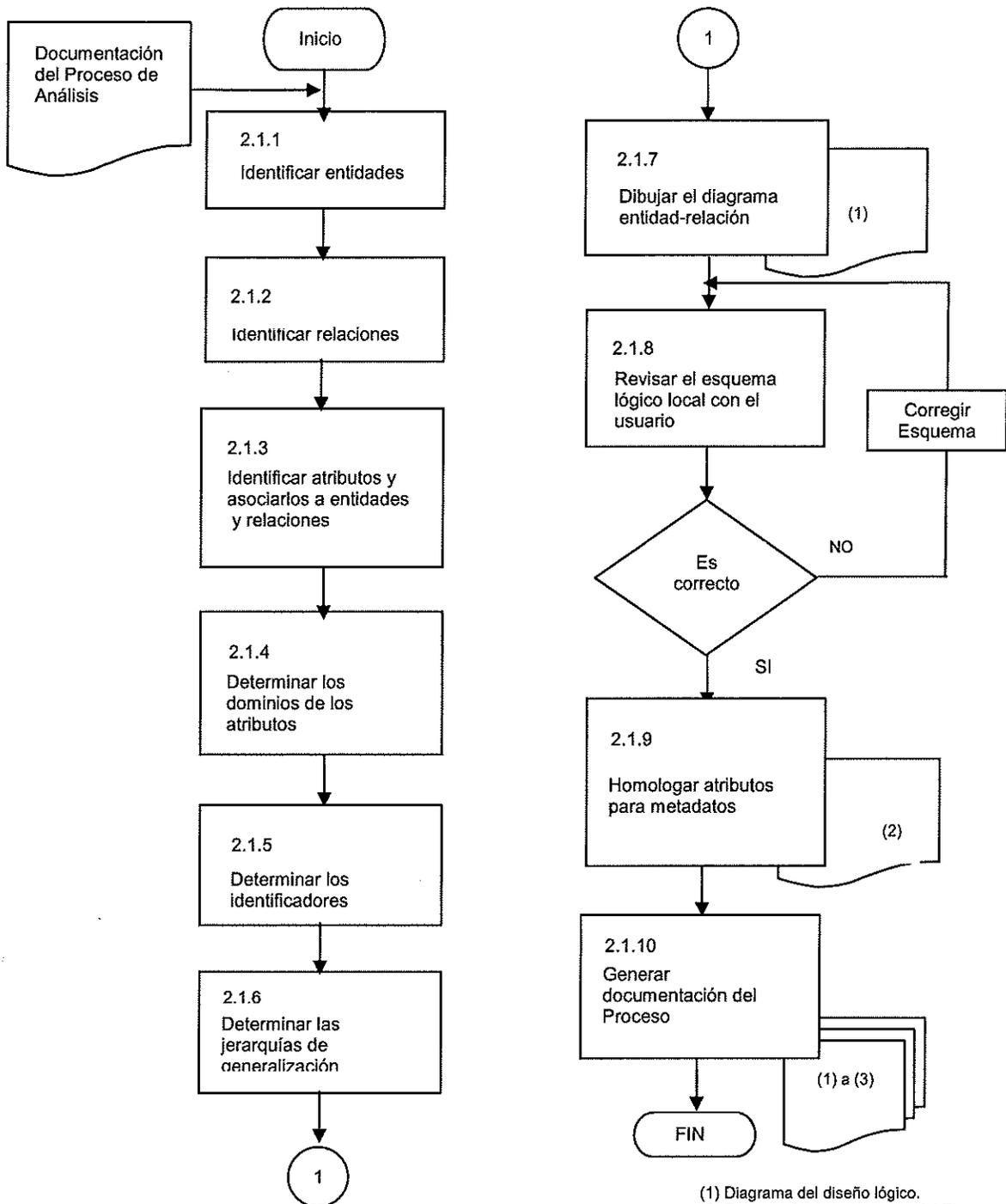
La carga de la información considera una revisión de los datos y sus estructuras y la aplicación de la afinación a la base de datos si es necesario, todo esto para garantizar el mejor rendimiento del almacén de datos.

De igual forma que en las etapas anteriores, a partir de la carga de información se generan documentos que sirven como referencia de esta etapa de la incorporación.

DIAGRAMA DE ANÁLISIS

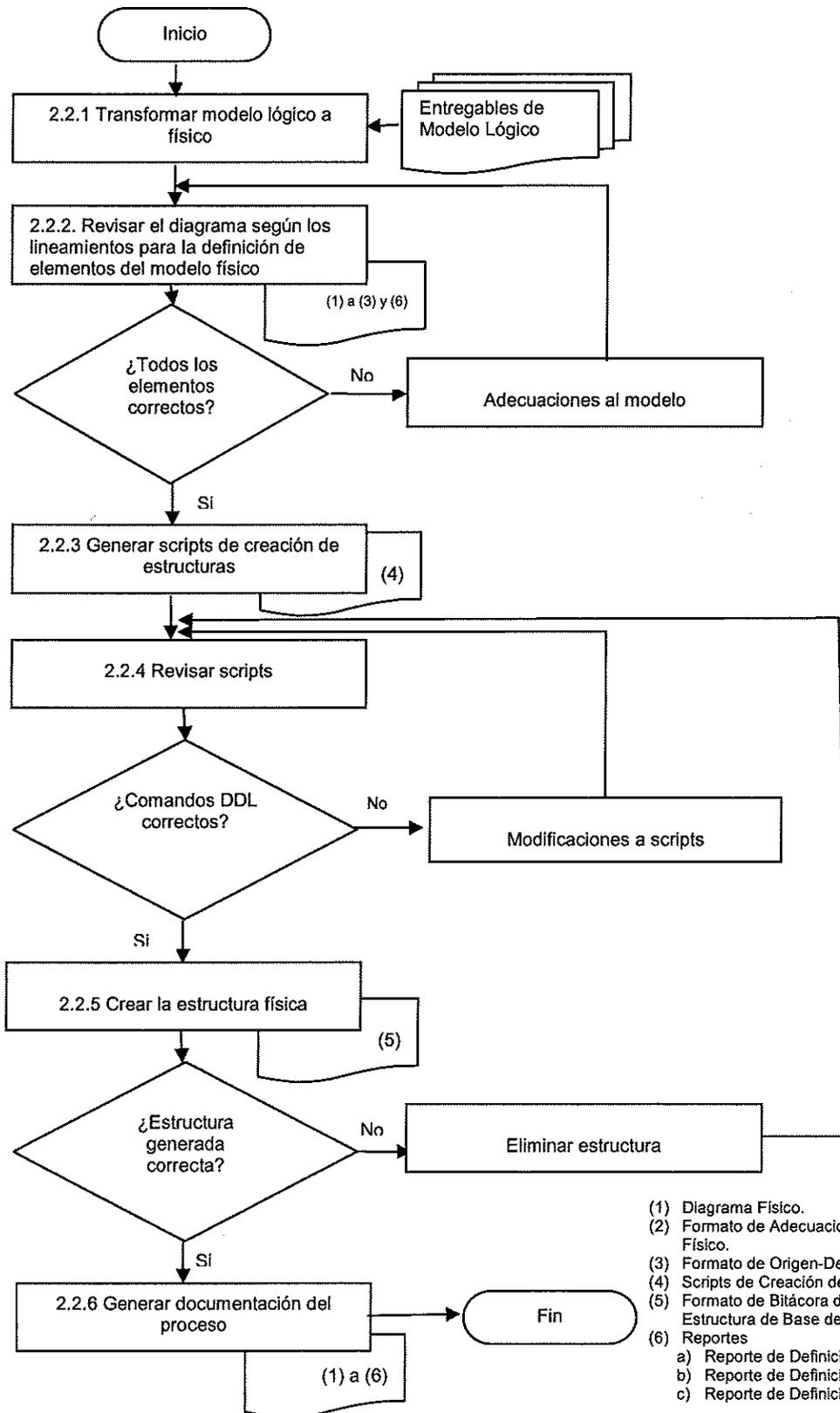


DIAGRAMAS DE DISEÑO MODELO LÓGICO



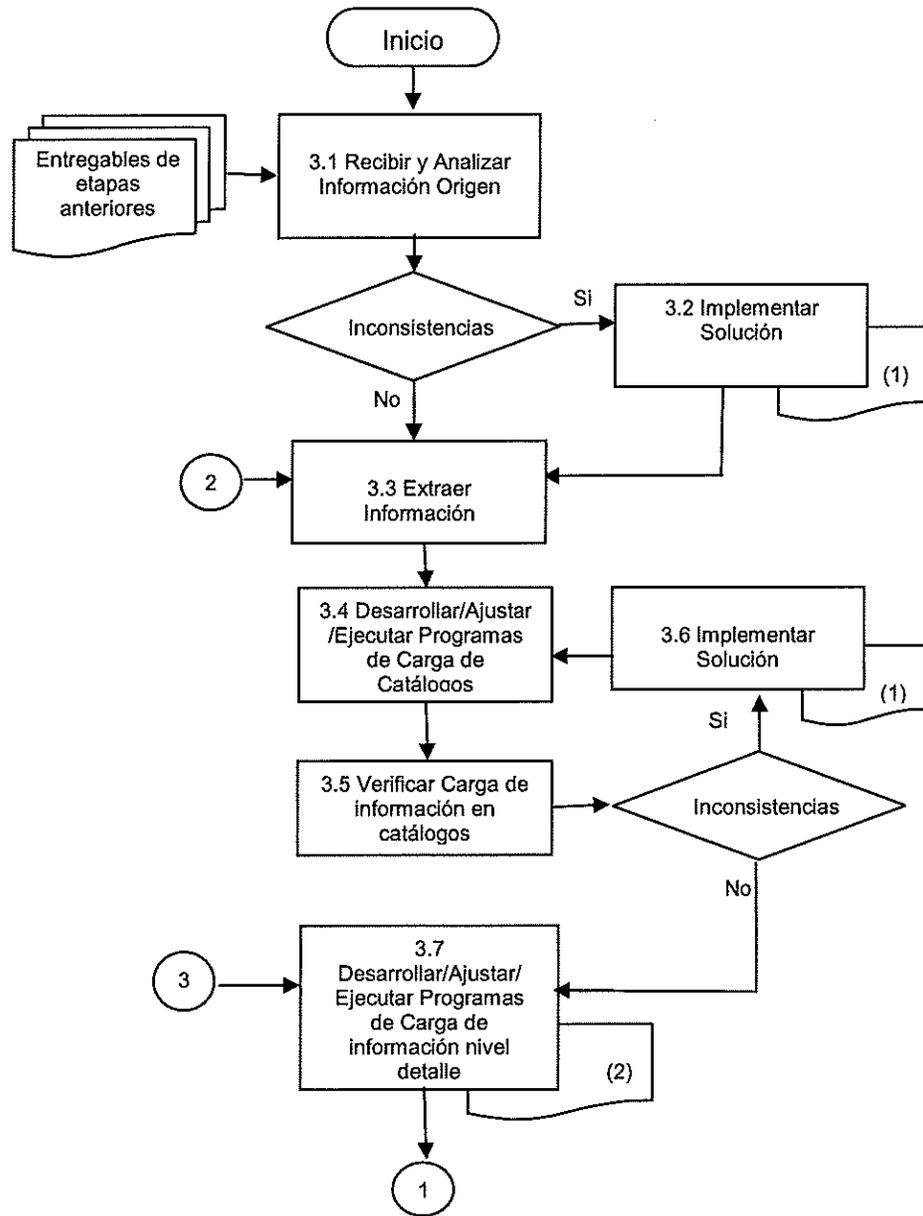
- (1) Diagrama del diseño lógico.
- (2) Formato de Homologación del Elemento de Dato.
- (3) Reportes
 - a) Definición de entidades.
 - b) Entidades y sus atributos.
 - c) Descripción de atributos.
 - d) Atributo en dominios.

DIAGRAMAS DE DISEÑO MODELO FÍSICO

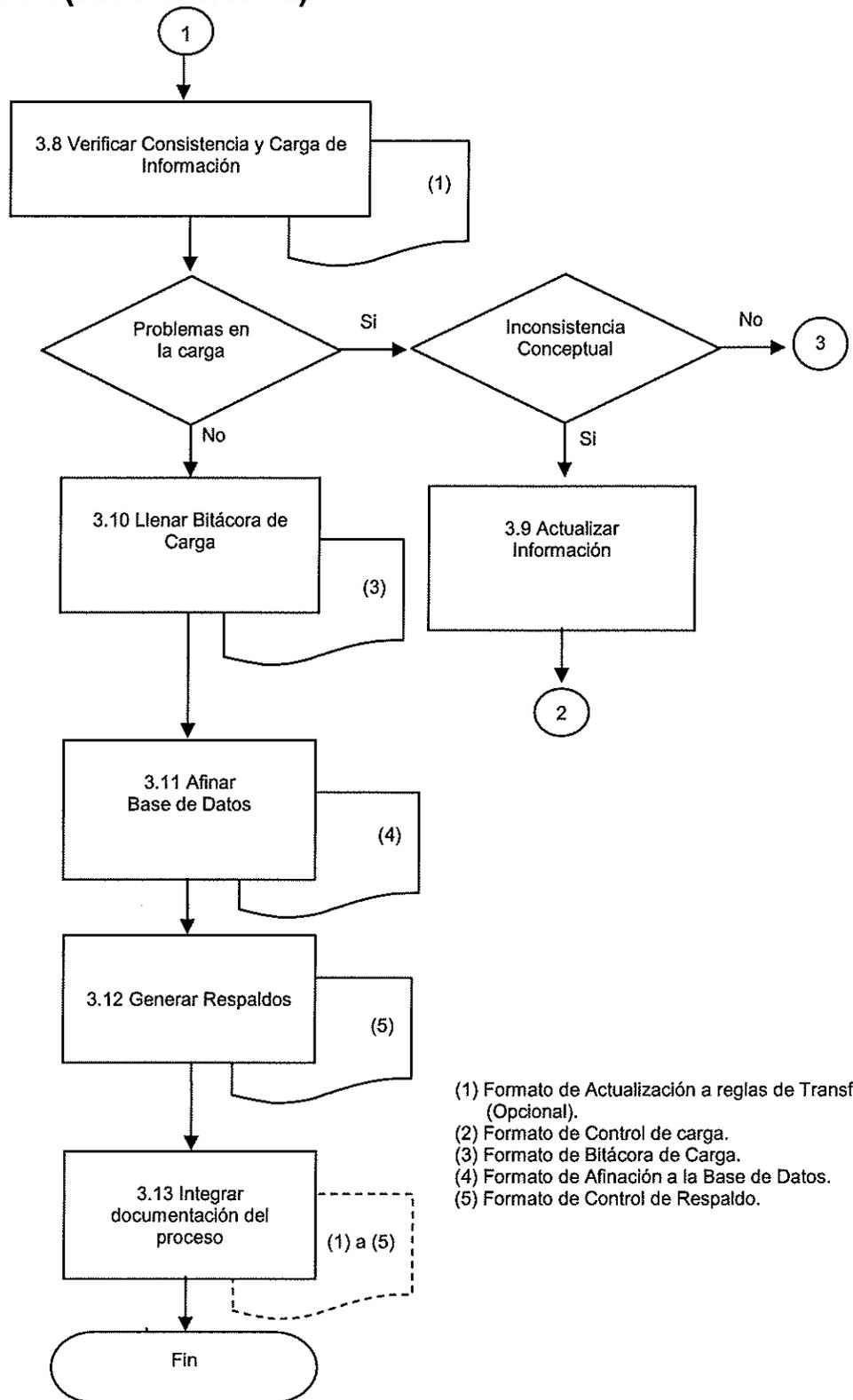


- (1) Diagrama Físico.
- (2) Formato de Adecuaciones del Modelo Físico.
- (3) Formato de Origen-Destino.
- (4) Scripts de Creación de Estructura.
- (5) Formato de Bitácora de Generación y Modificaciones de Estructura de Base de Datos.
- (6) Reportes
 - a) Reporte de Definición de Tablas.
 - b) Reporte de Definición de Columnas.
 - c) Reporte de Definición de Restricciones.

DIAGRAMA DE CARGA DE INFORMACIÓN (PRIMERA PARTE)



CARGA DE INFORMACIÓN (SEGUNDA PARTE)



- (1) Formato de Actualización a reglas de Transformación (Opcional).
- (2) Formato de Control de carga.
- (3) Formato de Bitácora de Carga.
- (4) Formato de Afinación a la Base de Datos.
- (5) Formato de Control de Respaldo.

MODELOS DE DATOS ROLAP

(ESTRELLA)

1. ANÁLISIS DE LA INFORMACIÓN

Para los proyectos cuya información se ha incorporado en un modelo de datos relacional (nivel detalle), se reutiliza gran parte del análisis realizado para dicho modelo. En caso contrario, se inicia recabando las generalidades del proyecto así como la documentación que se ha generado durante el desarrollo del mismo, esta información será el insumo para hacer la revisión de los datos que se generan del proyecto, al igual que sus metadatos.

El análisis incluye una confrontación del proyecto con la información contenida en el almacén de datos para hacer los ajustes necesarios a fin de facilitar la comparabilidad de la información durante las consultas.

A diferencia del modelo de datos relacional (nivel detalle), el análisis del modelo de datos ROLAP consiste en identificar y determinar los elementos que conformarán al modelo a partir de las necesidades de consulta de la información y definir así una estructura multidimensional de datos.

Durante la etapa de análisis se generan documentos cuyas estructuras están predefinidas en la vista detallada de la metodología y que permiten contar con el registro uniforme de los resultados de esta etapa.

2. DISEÑO DE LA BASE DE DATOS

El diseño de las estructuras de Base de datos se construye observando los lineamientos que se detallan para la información Relacional a nivel registro. De tal forma que al diseñar un esquema ROLAP, puede consultarse la sección homóloga a ésta, ubicada dentro de este documento.

3. CARGA DE INFORMACIÓN

El modelo de dato ROLAP se compone de dos tipos de tablas: Hechos y Dimensiones.

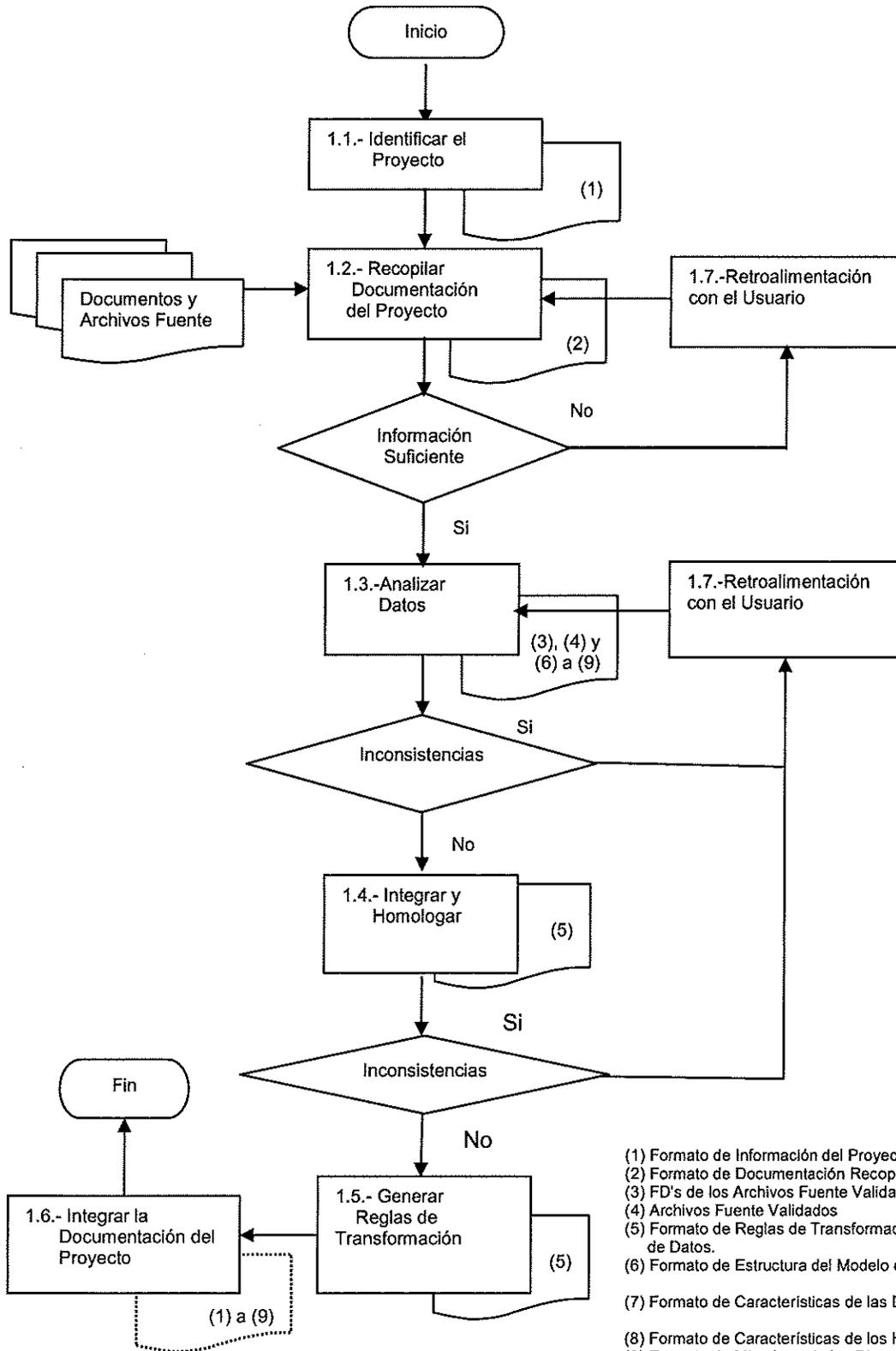
Es importante tomar en cuenta que la información a cargar a las tablas de dimensiones puede derivarse de la información de los catálogos del modelo relacional o bien de información que así se determinó por el usuario o como resultado de la etapa de análisis hecha previamente.

Una vez que se ha efectuado de forma exitosa las dimensiones, se procede a desarrollar los programas para la carga de la información de las tablas de hechos tomando en cuenta si la información ha de cargarse desde archivos proporcionados por el usuario o a partir de la información que existe en el modelo de datos relacional (nivel detalle).

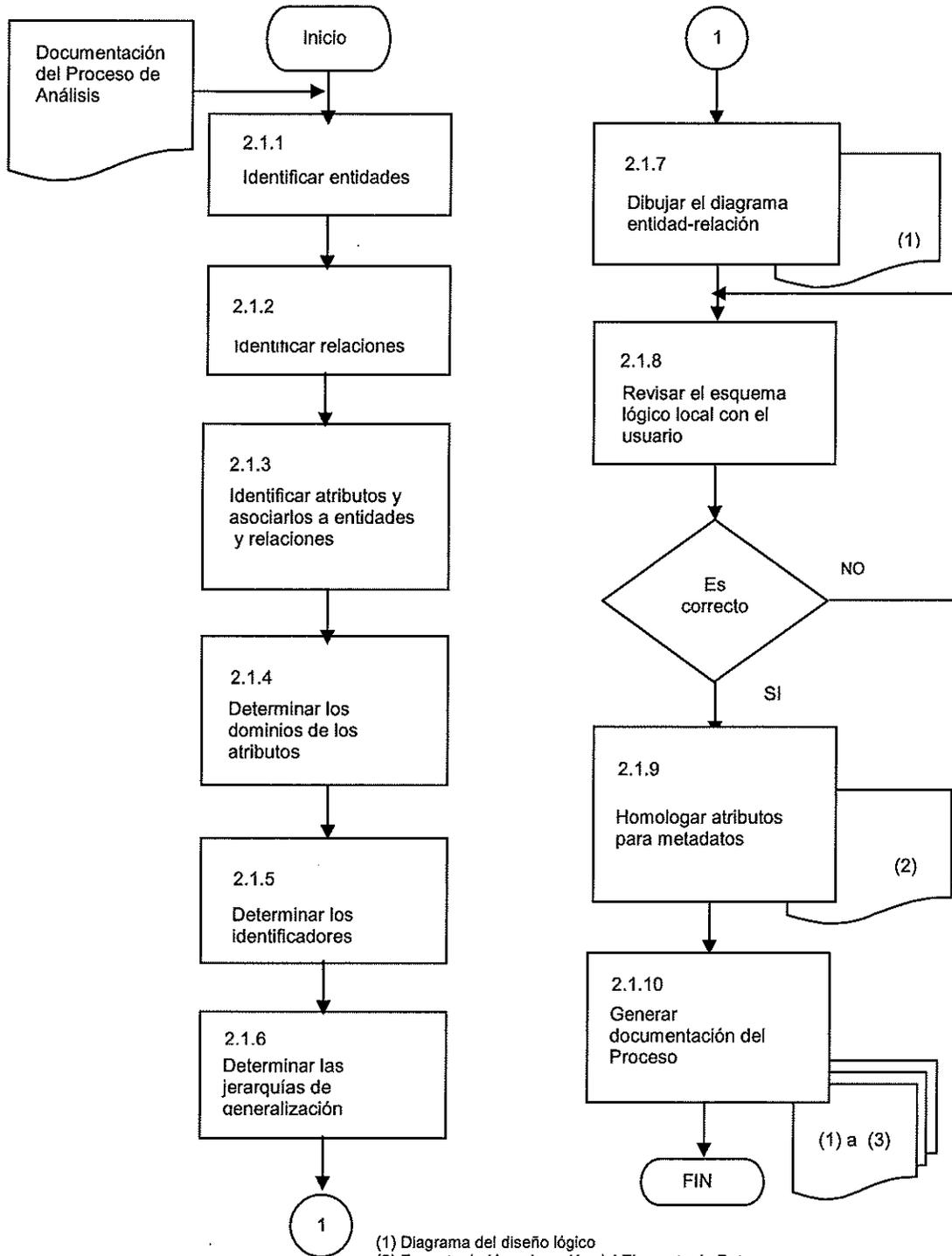
La carga de la información considera una revisión de los datos y sus estructuras y la aplicación de la afinación a la base de datos si es necesario, todo esto para garantizar el mejor rendimiento del almacén de datos.

De igual forma que en las etapas anteriores, a partir de la carga de información se generan documentos que sirven como referencia de esta etapa de la incorporación.

DIAGRAMA DE ANÁLISIS



DIAGRAMAS DE DISEÑO MODELO LÓGICO



- (1) Diagrama del diseño lógico
- (2) Formato de Homologación del Elemento de Dato.
- (3) Reportes
 - a) Definición de entidades.
 - b) Entidades y sus atributos.
 - c) Descripción de atributos.
 - d) Atributos en dominios.

DIAGRAMAS DE DISEÑO MODELO FÍSICO

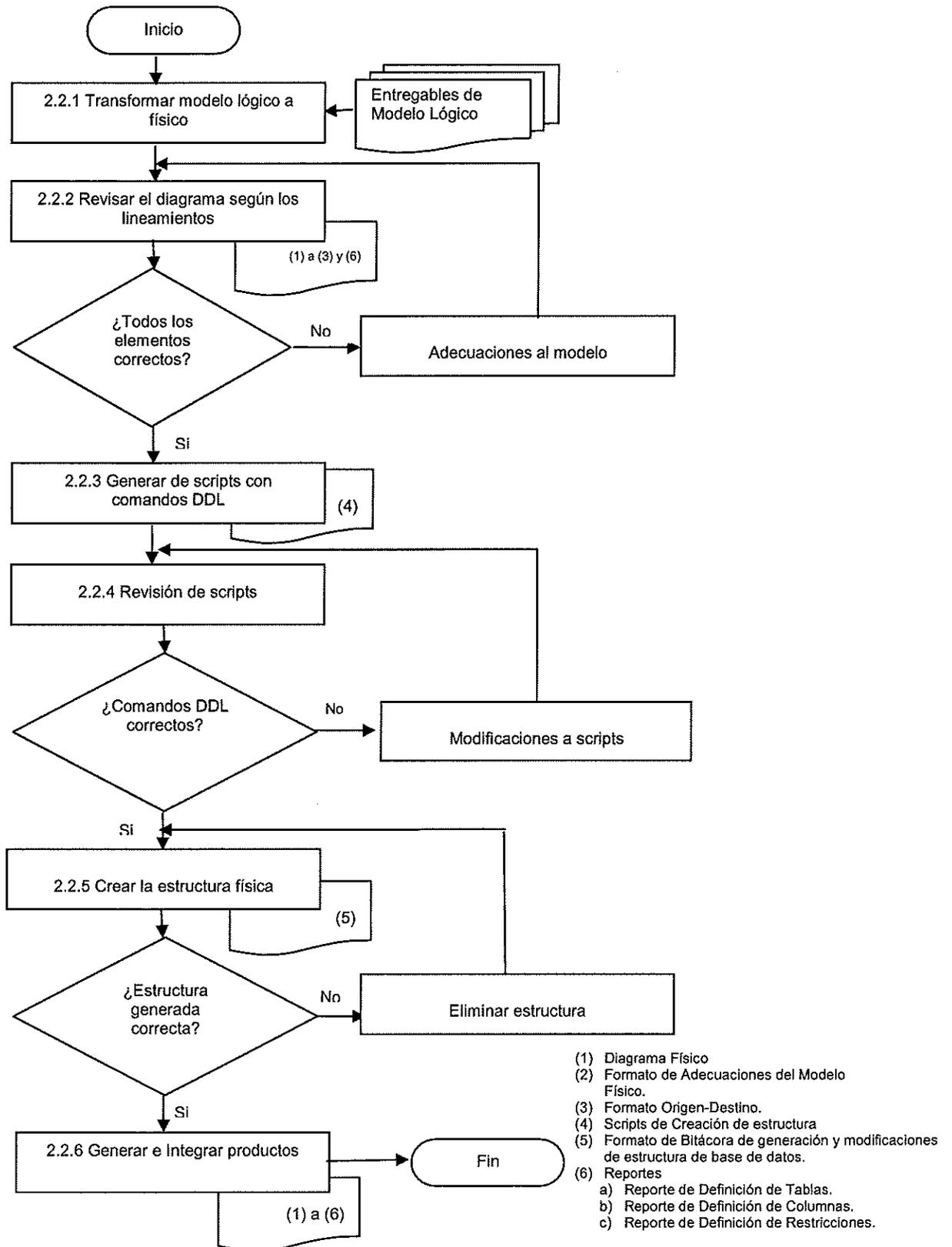


DIAGRAMA DE CARGA DE INFORMACIÓN (PRIMERA PARTE)

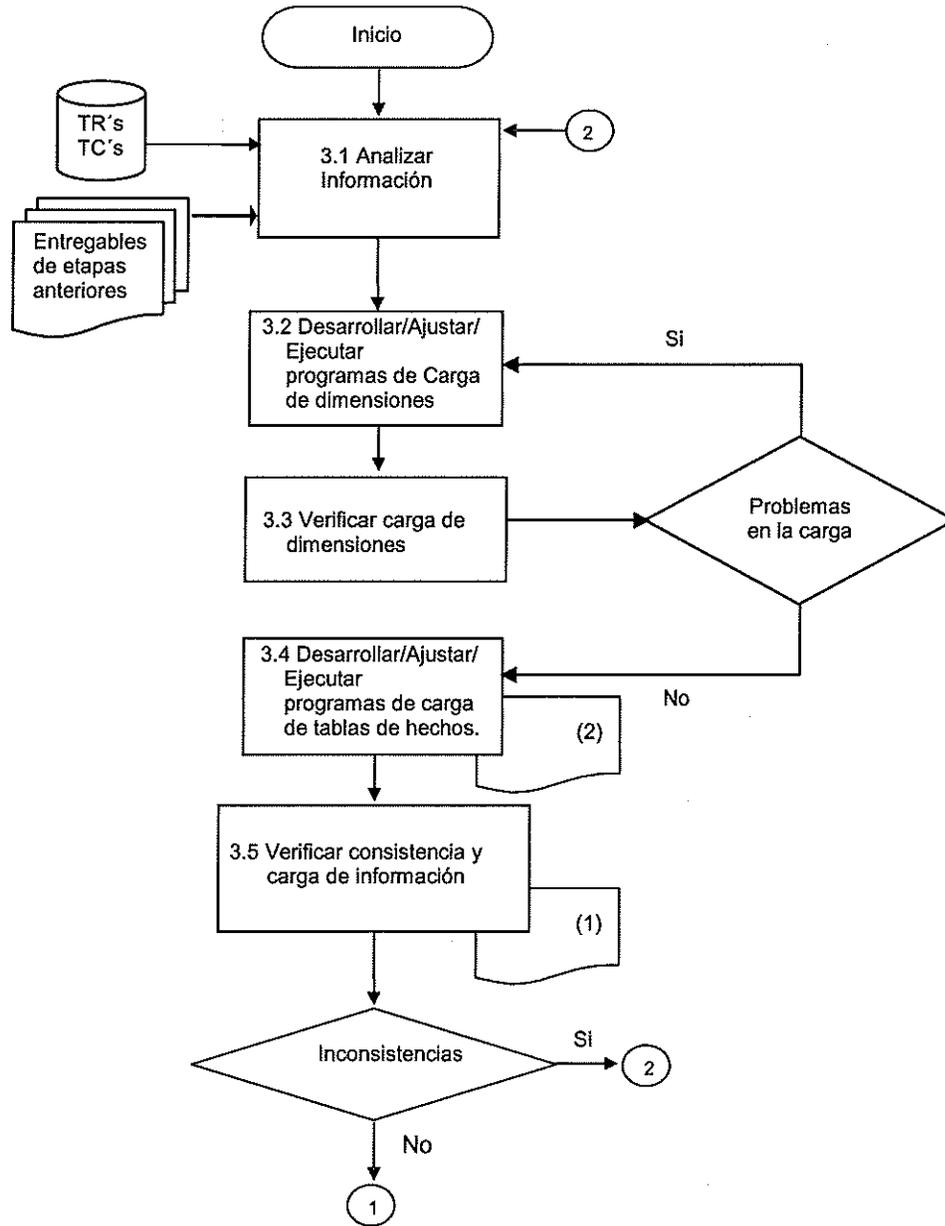
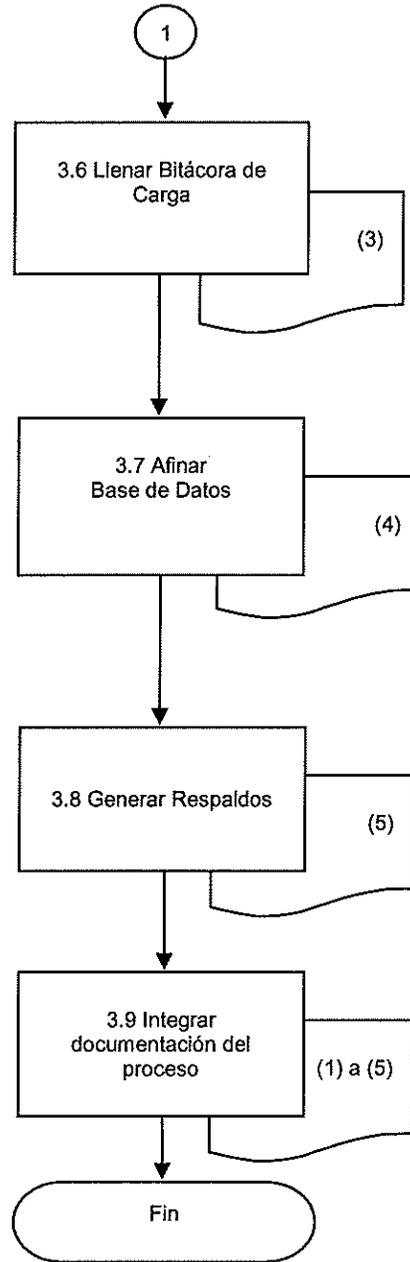


DIAGRAMA DE CARGA DE INFORMACIÓN (SEGUNDA PARTE)



- (1) Formato de Actualización a reglas de Transformación (Opcional).
- (2) Formato de Control de Carga.
- (3) Formato de Bitácora de Carga.
- (4) Formato de Afinación a la Base de Datos.
- (5) Formato de Control de Respaldos.

MODELOS DE DATOS MOLAP

(CUBO)

1. DISEÑO DE DATOS

El diseño de datos define la estructura de datos MOLAP, especificando las características y contenido de cada una de las dimensiones y medidas.

Una vez definida la estructura del cubo se realiza la generación del mismo por medio de herramientas de software especializadas en tal tarea, para su posterior revisión y liberación.

Durante la etapa de diseño se generan documentos cuyas estructuras están predefinidas en la vista detallada de la metodología y que permiten contar con el registro uniforme de los resultados de esta etapa.

DIAGRAMA DE DISEÑO DE DATOS (PRIMERA PARTE)

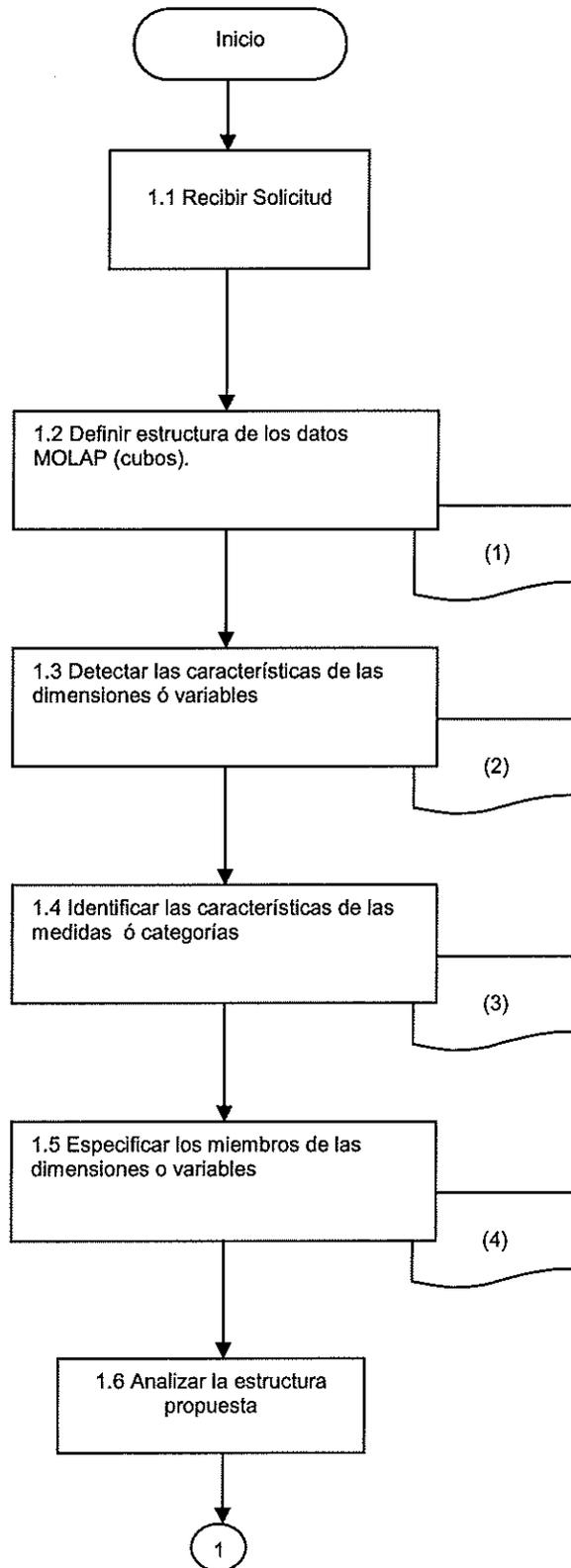
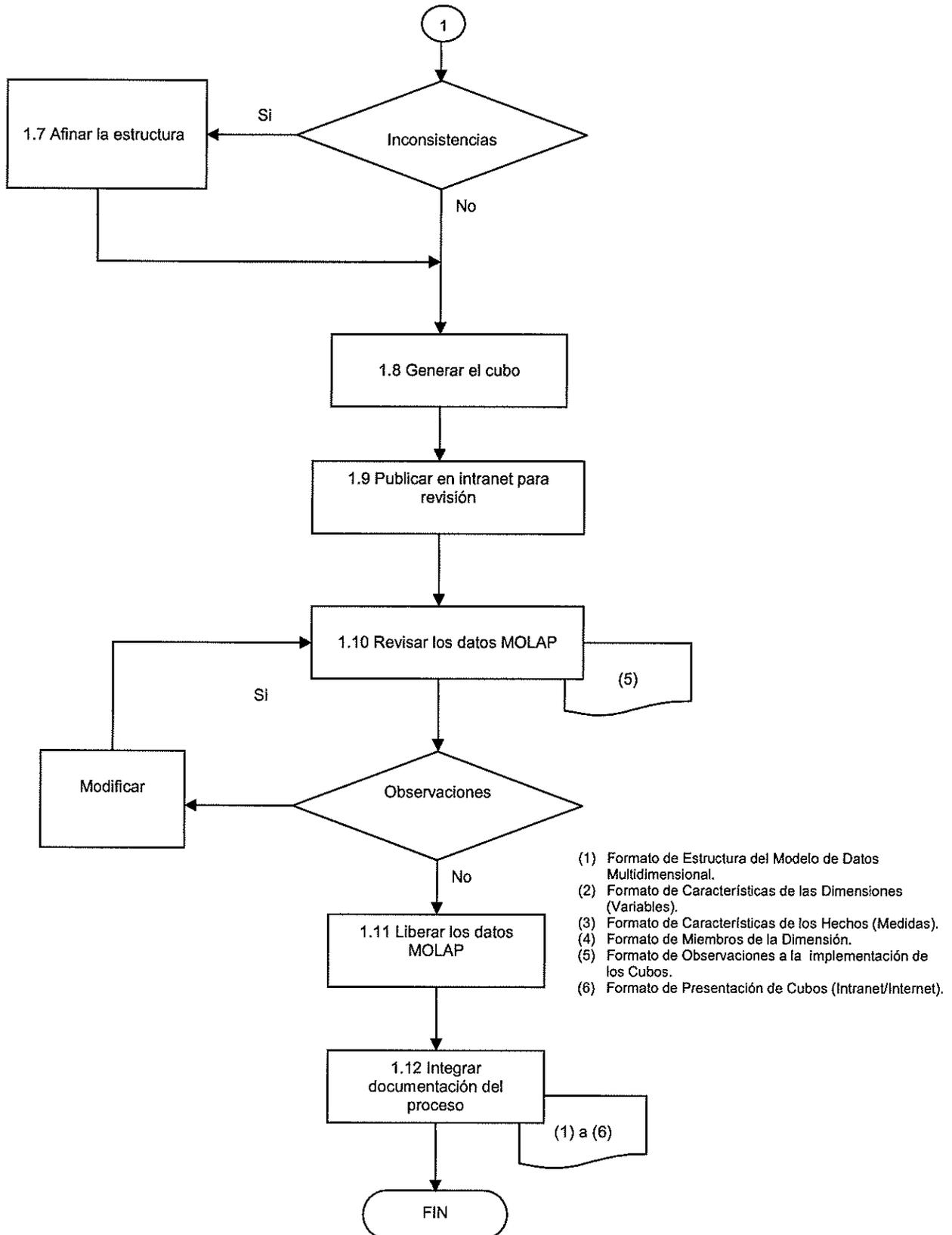


DIAGRAMA DE DISEÑO DE DATOS (SEGUNDA PARTE)



GLOSARIO

GLOSARIO

DWH

(Data Warehouse) Almacén de Datos

Dimensión

Es cada uno de los ejes en un modelo multidimensional y contiene el detalle de los valores que se están contabilizando (catálogos).

Hecho

Parte de un modelo de datos multidimensional que contiene valores de tipo cuantitativo que se obtienen generalmente por aplicación de una función estadística También conocido como medida o categoría.

Modelo de datos relacional (Nivel detalle)

Se entiende por información a nivel registro la información que se recaba de cada una de las unidades de observación y a la que se le aplican procesos de validación y de agrupamiento para su publicación.

Modelo de datos MOLAP (Cubo)

Es un conjunto de datos que se construye a partir de un subconjunto de un almacén de datos (tablas de hechos y dimensiones) y se organiza y resume en una estructura multidimensional definida por un conjunto de dimensiones y medidas. Es OLAP que accede directamente a bases de datos multidimensionales (Multidimensional Online Analytical Processing).

Modelo de datos ROLAP (Estrella)

Es un modelo de datos utilizado con frecuencia en un almacén de datos, que deriva su nombre del hecho que su diagrama forma una estrella, con puntos radiales desde el centro. El centro de la estrella consiste de una o más tablas de hechos, y las puntas de la estrella son las tablas de dimensiones. Es una forma de OLAP que realiza análisis multidimensional de datos almacenados en bases de datos relacionales (Relational Online Analytical Processing).

