# CAPÍTULO 1

# INTRODUCCIÓN

Actualmente existe la necesidad de tener una visión analítica y universal de la evolución de situaciones ambientales, sociales, administrativas a través del acceso a bases de información que se alimentan de datos de diferente naturaleza. Sin embargo, muy pocos trabajos hoy en día atacan de manera frontal los problemas de la integración de datos del medio ambiente, en particular datos hidrológicos, considerando las necesidades de los usuarios tanto expertos como casuales.

Tradicionalmente, los sistemas de información construidos para manipular este tipo de datos integran éstos de manera *ad hoc* en bases de datos que no los resguardan de manera segura, ni permiten soportar la explotación y el mantenimiento (actualización) transparente de los mismos. En la administración pública, las diferentes instituciones y dependencias, desarrollan y mantienen sus propios sistemas de información para administrar datos sobre la descripción del comportamiento de los ríos, volumen de aguas y crecimiento de presas con respecto a otras condiciones (meteorología, aspectos sociales, producción). Sin embargo, cada uno observa, anota, manipula y analiza los datos de manera diferente y no proporcionan herramientas orientadas al análisis y explotación de este tipo de información. De esta manera se hace evidente la necesidad de diseñar y construir un soporte para el apoyo al almacenamiento integrado, a la consulta, al análisis, a la visualización y al mantenimiento automático de datos del medio ambiente para tomar decisiones según distintos tipos de necesidades.

La tecnología *Data Warehouse* (DW) parece ofrecer una solución interesante y bien adaptada a estas necesidades. Este trabajo contribuye a mostrar el interés de usar este tipo de tecnología para ofrecer diferentes criterios de observación de información de los principales ríos y presas del país que faciliten su análisis.

## 1.1 Tecnología *Data Warehouse*

Un DW es la colección de una extensa variedad de datos, organizados, integrados, historizados y disponibles para facilitar la toma de decisiones de los usuarios finales [9]. La creación de un DW consiste en 4 pasos (Figura 1.1): diseño, construcción, análisis y mantenimiento.

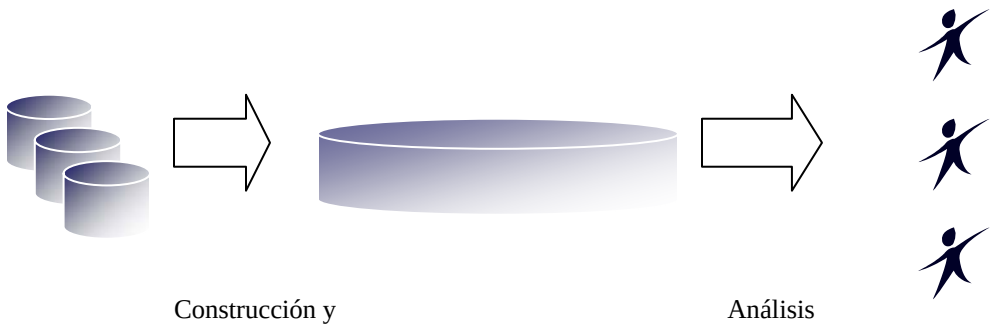


Figura 1.1 Entorno de un DW

### 1.1.1 Diseño

Para el diseño de un DW, en general, se emplea la representación de un modelo multidimensional que se basa en los conceptos de dimensión y medida. Un conjunto de dimensiones ortogonales definen un hiper-cubo como el que se presenta en la Figura 3.

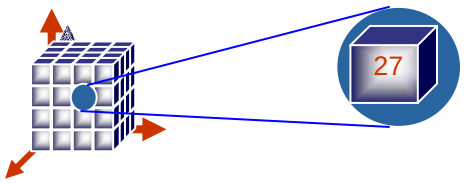


Figura 1.2 Esquema multidimensional (cubo)

Un modelo multidimensional permite definir el esquema multidimensional para diseñar un DW. El esquema multidimensional puede ser implementado por un esquema relacional. Dos tipos de esquemas relacionales pueden implementar un esquema multidimensional: esquema en estrella (*start schema*) y esquema copo de nieve (*snow flake schema*).

El esquema en estrella consta de una tabla principal de hechos donde cada uno de los atributos de ésta corresponde a una tabla de dimensión. Así todas las tablas de dimensión están relacionadas directamente con la tabla de hechos. El esquema de copo de nieve corresponde a la normalización del esquema en estrella. Para ello, se define una tabla de hechos y una tabla por dimensión [4].

### 1.1.2 Construcción

El proceso de construcción lleva la información de las fuentes al DW y se realiza en cuatro fases principales:

1. *Extracción*: consiste en acceder las diversas fuentes y recuperar la información que será integrada en el DW.
2. *Integración*: consiste en transformar los datos recuperados con respecto al esquema del DW. La integración se lleva acabo en dos etapas:

- Homogeneización, transformación de la información en el formato nativo de las fuentes, al formato y modelo de datos del DW.
- Integración, la información recuperada es agregada y organizada con respecto el esquema multidimensional del DW.

3. *Limpieza:* es la corrección en los datos de posibles errores, como datos de tamaño o descripción inconsistentes, falta de datos de entrada o datos que violen las restricciones de integridad del sistema.
4. *Apertura:* revisión de los niveles de agregación y el ordenamiento, así como la construcción de índices y la partición de áreas de almacenamiento.

### 1.1.3 Análisis

Una vez construido el DW se puede realizar un análisis como soporte para la toma de decisiones. El análisis se refiere a la explotación del DW: la forma en que se expresa una consulta analítica, la manera en que los datos serán agregados para ser analizados y la parte de información a la que tendrán acceso los diversos usuarios.

### 1.1.4 Mantenimiento

El mantenimiento de un DW es una función repetitiva cuyo objetivo es refrescar su contenido. Consiste en integrar periódicamente los cambios producidos en las fuentes. Dos puntos son importantes para mantener a un DW: cuándo refrescar y cómo refrescar.

Usualmente los sistemas de DW son refrescados periódicamente (v.g. diariamente o por semana). Las condiciones de refrescado son establecidas por el administrador del DW, dependiendo de las necesidades del usuario, del volumen de los datos, de la frecuencia

con la que cambian, etc. La mayoría de los sistemas de base de datos actuales, proveen servicios de duplicación que soportan técnicas de propagación de datos en forma incremental [1].

## 1.2 Análisis de información de ríos y presas

Los datos relacionados con medidas tomadas de fenómenos naturales que ocurren diariamente en el medio ambiente son muchos y muy diversos, y las necesidades de aplicación para las cuales son calculados son tantas y con objetivos tan diversos que mencionarlas podría llevarnos un trabajo de investigación completo. Sin embargo resulta evidente que los estudios que se llevan a cabo sobre la naturaleza, ya sea de fenómenos relacionados con el agua, viento, etc. siempre son provocados por el afán del hombre de entender la naturaleza, poder registrarla y conocerla para sacar el máximo de provecho para su bienestar y supervivencia. Cuando hablamos de ríos sabemos que existen una gran cantidad de medidas relacionadas con ellos que pueden ser de utilidad para un sin número de aplicaciones que el hombre puede darle, como conocer los ciclos hidrológicos del agua para predecir el comportamiento de un río en una zona específica y determinar posibilidades de asentamientos humanos en esa zona, distritos de riego, por mencionar sólo algunos de las múltiples beneficios que puede traer conocer el comportamiento de una corriente. Para poder obtener información de este tipo es necesario tener acceso a información resumida, histórica y confiable, basada en datos recolectados diariamente de los principales ríos y presas del país que permita soportar un análisis de esta naturaleza.

La información disponible acerca de ríos y presas se encuentra concentrada en una fuente de datos creada por la Comisión Nacional del Agua denominada BANDAS. El BANDAS es la fuente de datos más confiable con que se cuenta con información de esta naturaleza. Posee mediciones de gastos, profundidades, ubicación y nombres de los principales ríos del país a través de periodos de tiempo y que los expertos en el área requieren estudiar para el análisis de los mismos. De igual forma posee información sobre las principales presas del país y las medidas asociadas a las mismas tomadas a través de un periodo de tiempo. El BANDAS es una base de datos dividida en varios discos cada uno de los cuales almacena información de distintas regiones y en algunos existe información de ríos y en otro más de presas. Dicha base de datos se puede acceder a través del SIAS (un sistema de consulta directo de la base), el cual permite conocer medidas específicas, pero no proporciona herramientas de análisis de los datos.

Recientemente se han desarrollado trabajos relacionados con el BANDAS. Estos trabajos abarcan desde la captura de la información hasta la visualización de la misma vía internet. Los trabajos [20,21,22] presentan distintas formas a través de las cuales los datos capturados por las estaciones hidrométricas son llevados al centro de recepción de la información. El trabajo [23] plantea la forma en que estos datos se reciben, se juntan y son introducidos en el BANDAS. Por último se realizó un trabajo [24] para presentar la información contenida dentro del BANDAS en el internet y que de esta manera los datos puedan ser consultados desde cualquier punto.

## 1.3 Objetivos y metodología

El objetivo general de la tesis es la construcción, implementación y validación de un *Data Warehouse* sobre información de ríos y presas para apoyar el análisis y la toma de decisiones según distintos tipos de necesidades.

### 1.3.1 Metodología

La información de ríos y presas es extensa y muy diversa. Construir un DW implica modelar el tipo de información que será interesante para un contexto aplicativo específico (v.g., gasto máximo, gasto mínimo, profundidad, almacenamiento). Una vez seleccionada dicha información se integra y homogeniza (procesos de construcción de un DW). Efectivamente, son muchas las medidas en las que se puede enfocar la construcción de un DW para la toma de decisiones. Este trabajo considera dos ramas de la información: *los ríos y las presas del país*, y partiendo de ese contexto se toman en cuenta las medidas relacionadas con ambas ramas que sean de interés para el análisis que se quiere soportar.

Para alcanzar el objetivo, se diseñó e implementó un DW con información concentrada dentro del BANDAS. Se estudió la estructura de esta compleja fuente de información, se implementaron extractores de datos adaptados para recuperar información de las fuentes y se construyó un repositorio para su almacenamiento. Finalmente, se construyó un sistema para la expresión y evaluación de consultas analíticas sobre un sistema de bases de datos relacional. El resultado fue el sistema SARP que describimos en la siguiente sección.

### 1.3.2 SARP

La contribución principal de la tesis es SARP un sistema de análisis de datos de ríos y presas del país. SARP construye un conjunto de *Data Warehouses*, lo que se denomina constelación, a partir de datos extraídos del BANDAS. Los *warehouses* implementan sus esquemas multidimensionales sobre un sistema de administración de bases de datos relacional (SGBDR) y están asociados a un motor de consultas analíticas.

El motor de consultas ofrece una interfaz gráfica para la expresión de consultas y se apoya en el lenguaje SQL (*Standard Query Language*) para calcularlas. Cada DW está asociado a un sistema de integración de datos que implementa módulos adaptados a la extracción de datos del BANDAS y su integración con respecto a los esquemas multidimensionales de SARP.

**Constelación de ríos.** La Figura 1.3 ilustra el esquema multidimensional implementado por SARP asociado a cada uno de los DW orientados al análisis de los ríos. El esquema define el cubo de ríos por las dimensiones ZONA, TIEMPO y ESTACIÓN que caracterizan las medidas de acuerdo a la zona, el tiempo y la estación del año. Las medidas a analizar dentro de esta constelación son el gasto medio, gasto máximo y mínimo y la profundidad máxima y mínima del río, donde cada una de ellas representa un *warehouse* dentro de la constelación.

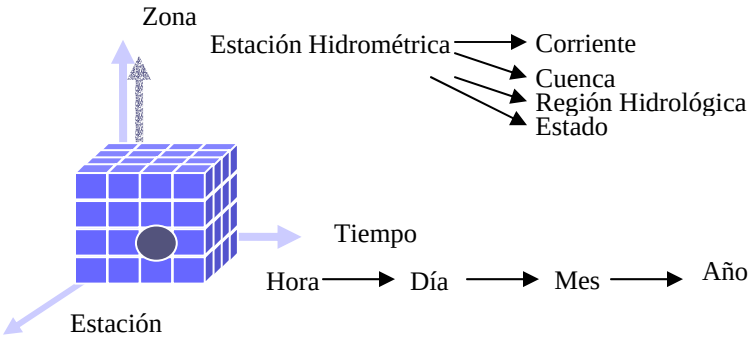


Figura 1.3 Modelo multidimensional para ríos

**Constelación de presas.** La Figura 1.4 ilustra el esquema multidimensional implementado por SARP asociado a cada uno de los DW orientados al análisis de las presas. El esquema define el cubo de ríos por las dimensiones ZONA, TIEMPO y ESTACIÓN que caracterizan las medidas de acuerdo a la zona, el tiempo y la estación del año. Las medidas a analizar son el almacenamiento máximo y mínimo de la presa y su elevación máxima y mínima del nivel de aguas, donde cada una de ellas representa un *warehouse* dentro de la constelación.

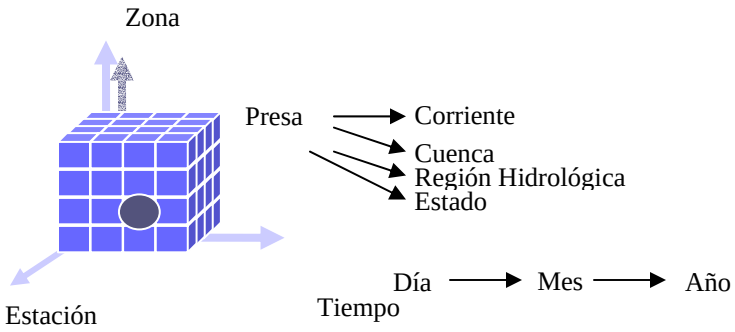


Figura 1.4 Modelo multidimensional para presas

**Extracción de datos.** El BANDAS se encuentra estructurado en un sistema de base de datos relacional. Debido a que el sistema nativo bajo el cual se encuentra es Visual Fox Pro, se utilizó un lenguaje integrado en el mismo ambiente de trabajo, denominado Visual Basic para la extracción de datos que se almacenan en el DW. El proceso de extracción de SARP se implementó a través de extractores adaptados a la estructura del BANDAS. Cada extractor incluye un tipo de conexión a la base que procesa los datos y los transforma para ser integrados en el DW.

**Análisis de los datos.** Para explotar el contenido del DW, SARP implementa un motor de ejecución de consultas analíticas basadas en operadores de *Online Analisis Processing*

(OLAP): *drill-down, roll-up, slice_and_dice* [1]. Por ejemplo, el usuario desea conocer el gasto máximo de agua de un río en los últimos diez años para conocer si es posible instalar una planta eléctrica en esa zona, de acuerdo a la cantidad de agua que pasa por esa zona. Otro ejemplo aplicativo es que el usuario desee conocer la cantidad de agua almacenada en una presa en un período de tiempo para obtener un patrón de conducta de acuerdo a su comportamiento.

## 1.4 Organización del documento

En este capítulo se presenta un breve preámbulo del contexto de realización de esta tesis, de sus objetivos y resultados principales. El resto del documento está organizado de la siguiente manera.

- El capítulo 2 define los conceptos de base asociados a la tecnología *Data Warehouse*. Se presenta la arquitectura general de un DW y sus funciones principales. Se define el concepto de modelo y esquema multidimensional en los que se basa el diseño de un DW. Enseguida se describe el proceso de construcción de un DW a partir de fuentes heterogéneas, se señala la dificultad de la integración (homogeneización, transformación) de datos. Luego se definen los operadores de análisis OLAP y finalmente se describe el problema de mantenimiento del DW.

- El capítulo 3, describe a SARP un sistema de consulta analítico para ríos y presas. Primero ilustra la arquitectura general de SARP y describe los esquemas multidimensionales que implementa y la estructura de las fuentes de los que se alimenta. Enseguida describe las funciones principales de SARP: construcción,

análisis y mantenimiento. En particular, describe la estrategia que implementa para el procesamiento de consultas OLAP y el tipo de consultas que puede ejecutar sobre los esquemas del DW.

- El capítulo 4 describe la implementación de SARP. Primero enumera las herramientas usadas y describe la manera en que la arquitectura general de SARP fue implementada en un contexto relacional. Se describe la arquitectura general de los módulos del sistema y las estrategias usadas para implementar sus funciones principales. Finalmente, el capítulo describe el uso y configuración de SARP y discute sus limitaciones y perspectivas de implementación.

- El capítulo 5, concluye el trabajo, subraya los resultados alcanzados, sus limitaciones y su contribución. Finalmente, enumera y discute las perspectivas que se identifican para la continuación del trabajo.