Economía, XXI, 11 (1996), pp. 203-213

Un ambiente para análisis de datos

An environment for data analysis

Marta Sananes y Elizabeth Torres*

Resumen

Se presentan los conceptos de diseño y construcción de la primera versión de un ambiente cómodo para Análisis de Datos, integrando en un solo sistema herramientas útiles de Base de Datos, Hoja de Cálculo y Procesador Estadístico. El gestor interno de Base de Datos está diseñado según el modelo que llamamos Relacional dinámico incorporando además algunos conceptos de diseño por objetos. En la Hoja de Cálculo se introduce el concepto de box para colocar colecciones de datos replicados. El Procesador Estadístico ejecuta programas en un lenguaje particular similar al Pascal que opera sobre conjuntos de datos estructurados como Rangos los cuales pueden definirse tanto en la Hoja de Cálculo como en cualquier Tabla de una Database soportada por el sistema. Globalmente el sistema se diseñó como metáfora de la arquitectura de computadores, integrando una colección de componentes: memoria, procesador, procesadores dedicados, terminal, gestor de archivos, gestor de Bases de Datos, los cuales se comunican mediante un componente transportador o bus. Para una etapa posterior se tiene planeado además integrar este ambiente al ambiente de desarrollo y experimentación con modelos de simulación en lenguaje GLIDER.

1. Introducción

En 1990, cuando se inició el Proyecto *Hoja de Cálculo Estadística*, financiado por el C.D.C.H.T. de la Universidad de Los Andes, nos propusimos diseñar y construir la primera versión de un ambiente cómodo para el Análisis de Datos, integrando en un solo sistema los tipos de herramientas comúnmente usados de Gestor de Base

* Universidad de Los Andes, Instituto de Estadística Aplicada y Computación

de Datos –para organización y registro de datos–, Hoja de Cálculo -para transformaciones de datos y operaciones auxiliares y Procesador Estadístico- para la aplicación de métodos estadísticos. Desde entonces la tendencia integradora se ha consolidado y se ofrecen en el mercado de software diversos productos que la exhiben.

El objetivo del presente trabajo es presentar los conceptos de diseño y construcción de la primera versión de un ambiente cómodo para Análisis de Datos. Para lograr este objetivo se integra, en un solo sistema, herramientas de Base de Datos, una hoja de cálculo y un procesador estadístico. Funcionalmente los componentes se agrupan en los subsistemas: hoja de cálculo (STS Statistical Sheet), gestor de base de datos (STB Statistical Base) y procesador estadístico (STP Statistical Processor). Ellos se relacionan jerárquicamente, siendo STB y STP subordinados a STS, tal como lo ilustra la Figura 1, en donde un pequeño componente de nombre ADAN permanece residente durante el funcionamiento del sistema y sirve para el proceso de arranque del usuario y del conmutador entre los subsistemas. Globalmente el sistema se diseñó como metáfora de la arquitectura de computadores, integrando una colección de componentes:

– Memoria
– Procesador
– Procesador de Fórmulas
– Procesador Estadístico
– Terminal
– Procesador de archivos
– Máquina de Base de Datos

Estos componentes se comunican mediante un componente transportador o bus. En particular mediante el bus se transfieren Rangos entre la Hoja y una Database y entre estos y el Procesador Estadístico.

A continuación se describen los componentes.

2. Memoria

El componente de Memoria contiene y gestiona el almacenamiento de las estructuras de datos que soportan la apariencia de la Hoja.

La estructura lógica de la Matriz Hoja se soporta por una estructura de Matriz u Hoja Virtual segmentada en páginas de tamaño fijo. Mediante funciones se mantiene la correspondencia entre el espacio de coordenadas del usuario y el de la Hoja Virtual.

Los valores contenidos lógicamente en las celdas del usuario pueden almacenarse directamente en las celdas de la Hoja Virtual –valores cortos– o indirectamente mediante referencias a direcciones de almacenamiento en una memoria lógica linealmente estructurada. El caso indirecto incluye el almacenamiento de fórmulas para las cuales se mantiene la fórmula original, la fórmula convertida para interpretación inmediata

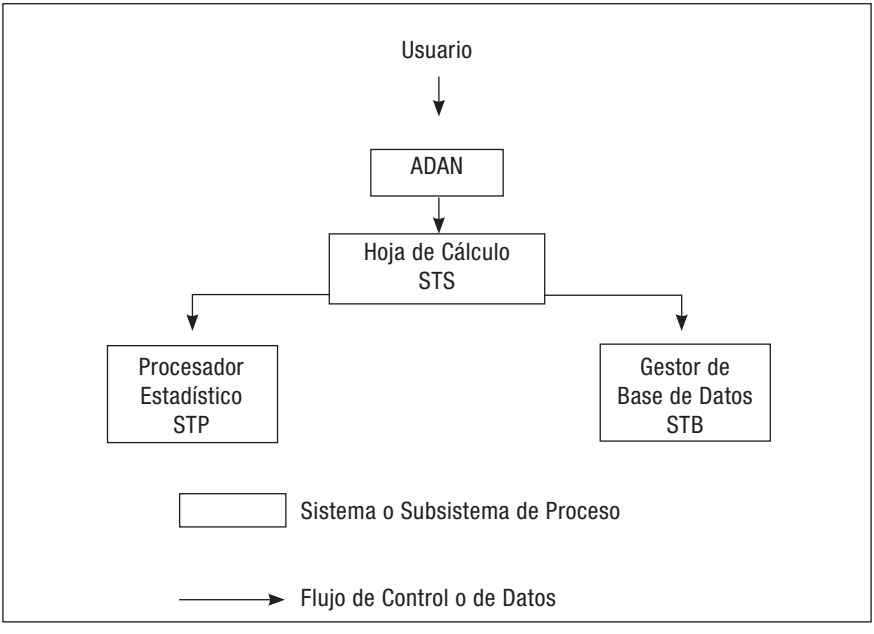

Figura 1. Esquema funcional del sistema

y el valor actual. El caso indirecto también incluye el almacenamiento de los conjuntos o listas de datos que constituyen el contenido de las boxes.

Además de almacenar indirectamente contenidos de celdas, también se usa para almacenar los elementos de los distintos tipos de listas que soportan otros conceptos del sistema. Se mantienen las siguientes listas:

– Rangos: Regiones rectangulares de la Hoja
– Recálculo: Celdas con fórmulas
– Replicaciones: Lista de contenidos de una box
– Atributos: Características particularizadas de columnas de la Hoja

Un Área de Control centraliza las estructuras de control de la pseudo memoria virtual y de las listas.

3. Procesador

Es el componente que recibe los eventos del Terminal-Interfaz de comunicación usuario/sistema- y los interpreta, activando la acción o procedimiento apropiado o señalando error en caso de inconsistencia.

Centraliza el control de los demás componentes. Como reacción a los eventos de selección de operaciones por parte del usuario pueden ocurrir tres tipos de acciones:

– Activación de un procedimiento local que a su vez puede activar a uno o más procedimientos contenidos en procesadores dedicados.
– Activación de un procedimiento contenido en un procesador dedicado.
– Activación de la operación de una de las máquinas subordinadas.

4. Procesador de fórmulas

Contiene todo el tratamiento de fórmulas: análisis sintáctico, validación de consistencia, encadenamiento entre celdas y fórmulas para permitir la aplicación del proceso de recálculo en orden natural[1], traducción a notación postfijo e interpretación.

5. Procesador estadístico

El Procesador Estadístico es una revisión y adaptación del sistema "DELIZ"[2], trabajo que fue presentado ante el IEAC/ULA como Tesis de Maestría.

Esta diseñado como un procesador dedicado que incorpora una biblioteca de procedimientos de cálculo estadístico y un interpretador de programas. Los procedimientos estadísticos se pueden invocar desde los programas escritos en un lenguaje adaptado basado en el Pascal. El interpretador traduce los programas a un lenguaje intermedio-extensión del lenguaje de la máquina PL/0 de Wirth[3] y posteriormente los ejecuta interpretando el programa intermedio.

Hasta ahora la biblioteca estadística está dedicada al análisis de diseño de experimentos y contiene procedimientos para realizar análisis de varianza para diversos tipos de diseños.

Los procedimientos operan sobre conjuntos de datos de estructura matricial que les son transferidos por el sistema desde la Hoja de Cálculo o desde alguna tabla de la Database en uso mediante archivos en almacenamiento secundario para no incurrir en limitaciones por espacio en memoria principal. Por lo tanto, el Procesador Estadístico tiene su propia capacidad de gestión de almacenamiento secundario tanto para acceder como para generar rangos archivados y matrices archivadas.

La capacidad de interacción con el usuario está incluida en la Hoja. Los programas que pautan la ejecución de procesos estadísticos se preparan con un editor de textos que es parte del componente Terminal de comunicación con el usuario. La capacidad de construcción de las matrices de transferencia de datos está en el componente Bus.

6. Terminal

Representa la Interfaz de comunicación entre el usuario y el sistema. Es un componente compuesto a su vez de:

– Gestor de Visualización o Pantalla.
– Gestor de Graficación.
– Gestor de Impresión.
– Gestor de Selección de Acciones o Menú.

Como en toda Hoja de Cálculo, se le presenta al usuario una estructura de árbol de menúes para la selección de operaciones o acciones a ejecutar. El estilo de presentación es el de Lotus 123 versión 2.2.

El componente Menú contiene la estructura arbórea del menú, otras estructuras relacionadas –como la pila de registro de las secuencias de activación– y procedimientos diseñados para la incorporación y operación de menúes.

La estructura lógica de un menú se define en un archivo externo diseñado bajo un concepto parecido al de recurso en el sistema Windows de Microsoft. Un archivo de definición de Menú permite especificar la estructura lógica de un menú, los rótulos de las opciones, el tipo de opción –pasiva o activa– y para cada opción de tipo activa un número entero con el cual se le asocia el procedimiento de su acción en el procesador.

7. Gestor de archivos

El Gestor de Archivos tiene dos responsabilidades:

– Gestionar la persistencia de las Hojas de los usuarios
– Gestionar el respaldo físico de los componentes virtuales de Memoria.

Los componentes de memoria virtual –pseudo memoria y hoja– se inicializan en áreas residentes en memoria principal. La primera vez que una demanda de asignación de espacio hace que se exceda el tamaño de la respectiva área residente, se activa el funcionamiento del proceso de memoria virtual abriendo los archivos de respaldo y copiando las paginas residentes iniciales. A partir de ese momento, cada vez que se detecta la condición de pagina faltante, opera un proceso de desalojo y reemplazo.

8. Máquina gestor de bases de datos

El Gestor de Bases de Datos incorporado internamente como Máquina Subordinada es una revisión y adaptación de uno desarrollado en 1980[4] para el proyecto Modelo Regional de Venezuela (MORVEN) bajo la coordinación de la Arq. Sonia Barrios[5] en el CENDES/UCV.

En el diseño y construcción del nuevo gestor asociado a la Hoja de Cálculo se han actualizado criterios de diseño tanto en los aspectos conceptuales de programación como en los de modelación de datos.

8.1. Criterios De Modelación De Datos

Se ha aplicado en este gestor un modelo básico de organización de datos que llamamos Relacional dinámico. El calificativo de dinámico enfatiza el carácter de extensibilidad por parte del usuario de la definición de los conceptos en cualquier momento de su interacción con el sistema. La figura 2 muestra el meta-esquema para este modelo conceptual, utilizando el estilo de especificación de esquemas de Oracle[6].

El modelo integra los siguientes conceptos:

– Clase de Entidad
– Entidad
– Variable
– Versión

– Componente
– Referencia
– Fuente
– Tablas de Unidades y de Códigos

8.2. Criterios de programación

- *Persistencia:*
  La persistencia de cada Database definida por el usuario está soportada por un conjunto de cinco archivos:

  – Maestro
  – Estructura
  – Diccionario
  – Data
  – Control
  – Gestión de memoria:

  Se ha aplicado un concepto limitado de memoria virtual para la gestión del almacenamiento Maestro que consiste en mantener en memoria principal una porción residente. Esta es hasta ahora la forma simple de realizar un concepto más general, al que llamamos Muestreo Estratégico de Bases de Datos. La idea es que un Gestor de Base de Datos mantenga una muestra relativamente pequeña del contenido de la Database en operación. Esta muestra servirá para evaluar estrategias de consulta aplicándolas a la muestra y escoger la que resulte más eficiente para la resolución de la consulta sobre la Base de Datos completa. Si la muestra es representativa, es válido extrapolar la conclusión de la evaluación.

- *Presentación:*
  La presentación y operación del Gestor se ha diseñado para que se asemeje a la de la Hoja de Cálculo. Se opera mediante menúes estilo Lotus hasta ahora.

Puede hacer transferencias entre Rangos definidos en la Tabla hacia la Hoja de Cálculo y viceversa. También puede definir Rangos –posiblemente toda la tabla– para el Procesador Estadístico.

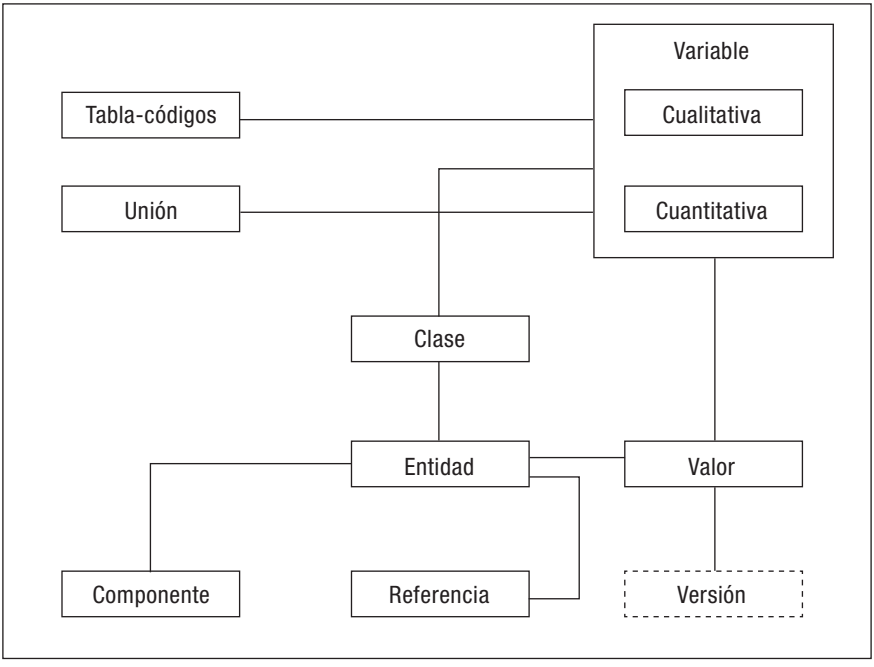

Figura 2. Meta-esquema

9. Bus

Este componente centraliza las estructuras de datos y los procedimientos para realizar las transferencias entre componentes del sistema.

10. Conclusión

En este trabajo se han presentado los conceptos básicos de diseño y construcción de un ambiente que facilite el Análisis de Datos. Se ha integrado, en un único sistema, herramientas útiles de Base de Datos, Hoja de Cálculo y un Procesador Estadístico.

Se ha completado la construcción de un prototipo para plataforma PC/MSDos, programado en Turbo Pascal V6.0., a partir del cual se está construyendo una primera versión del producto.

Esta primera versión a su vez hará de prototipo para la construcción de una versión para PC/Windows reprogramada en Turbo Pascal V7.0 y de otra versión programada en C++ para adaptación a plataformas bajo UNIX/X11/Motif.

11. Notas

1 Soler, Roger. Especificación denotacional y desarrollo de programas. Un ejemplo: La Hoja de Cálculo Departamento de Computación, Facultad de Ciencias, UCV. Caracas, 1985.

2 Torres, Elizabeth. Diseño e implementación de un lenguaje computacional para el Análisis de Experimentos "DELIZ" Tesis de Maestría, IEAC, FACES, ULA, Mérida, 1987.

3 Wirth, Niklaus. Algoritmos + Estructuras de Datos = Programas Ediciones del Castillo S.A., 1980.

4 Kim, Won y Lochovsky, Frederick H., Editores. Object-Oriented Concepts, Databases, and Applications ACM Press, Addison-Wesley Publising Company, 1989.

5 Barrios, Sonia. MORVEN - Hipótesis sobre la configuración espacial de Venezuela, CENDES, UCV, Caracas, 1980.

6 Barker, Richard. Case*Method: Entity Relationship Modelling Addison-Wesley Publishing Company y Oracle Corporation UK Limited, 1990.

12. Referencias

Barker, Richard (1990). *Case Method: Entity Relationship Modelling.* Addison-Wesley Publishing Company y Oracle Corporation UK Limited. Gran Bretaña.

Barrios, Sonia (1980). "MORVEN-Hipótesis sobre la configuración espacial de Venezuela". CENDES. UCV, Caracas.

Kim, Won y Lochovsky, Frederick H. (1989). *Object-Oriented Concepts, Databases, and Applications ACM.* Addison-Wesley Publising Company. Gran Bretaña.

Soler, Roger (1985). "Especificación denotacional y desarrollo de programas. Un ejemplo: La Hoja de Cálculo". Departamento de Computación, Facultad de Ciencias, UCV. Caracas.

Torres, Elizabeth (1987). "Diseño e implementación de un lenguaje computacional para el Análisis de Experimentos DELIZ". Tesis de Maestría. IEAC, FACES, ULA, Mérida.

Wirth, Niklaus (1980). *Algoritmos más Estructuras de Datos, Programas.* Ediciones del Castillo S.A.