REDBD 2002

Taller sobre Integración Semántica de Fuentes de Datos Distribuidas y Heterogéneas

El Escorial (Madrid), 20 de noviembre de 2002

VII Jornadas de Ingeniería del Software y Bases de Datos

Taller organizado por:
José F. Aldana
Carmen Costilla
Arantza Illarramendi
Esperanza Marcos
Antonio Polo

Estudio Empírico sobre Métodos de Diseño de Almacenes de Datos

Manuel Serrano, Coral Calero, Mario Piattini
Grupo Alarcos
E. S. Informática de Ciudad Real
Universidad de Castilla – La Mancha
{Manuel.Serrano, Coral.Calero, Mario.Piattini}@uclm.es

Resumen

Los almacenes de datos se han convertido en uno de los principales sistemas de información empresarial debido a su utilidad en la toma de decisiones estratégicas. La mayoría de los métodos de diseño de almacenes de datos se basan en modelos multidimensionales utilizando diagramas en estrella. Aunque se supone que el diseño en estrella es más sencillo de comprender por los diseñadores y usuarios del sistema, hasta la fecha no se han realizado estudios que demuestren esta suposición. En el presente artículo, se presenta un estudio empírico que intenta comprobar si los métodos de diseño basados en diagramas de estrella son más fáciles de entender y manejar que los diseños basados en el modelo relacional.

Palabras Clave: Almacenes de datos, Modelo Relacional, Diagramas en estrella

1. Introducción

Hoy en día las empresas pueden almacenar una gran cantidad de información obtenida a muy bajo precio, sin embargo, estos datos no son capaces de proporcionar información (Gardner, 1998). Para solucionar este problema han aparecido los almacenes de datos. Estos almacenes de datos son grandes repositorios de datos que integran datos provenientes de diferentes fuentes y sirven como sistemas de apoyo a la toma de decisiones. Los almacenes de datos se han convertido en una de las principales líneas de investigación y desarrollo de la informática empresarial, por ejemplo, Jarke et al. (2000) pronostican un mercado de 12 millones de dólares americanos para el mercado de los almacenes de datos durante los próximos años. Existen diversas metodologías de diseño para almacenes de datos (Adamson y Venerable, 1998; Hammergren, 1996; Inmon, 1997; Jarke et al., 2000; Kimball et al., 1998) las cuales se basan en los diagramas en estrella (que según los autores, aumentan la comprensión y mejoran el manejo de los almacenes de datos), dejando a un lado al modelo relacional.

En el presente artículo presentamos un estudio estadístico que hemos realizado para comprobar si los diseños en estrella son más comprensibles que los diseños realizados utilizando el modelo relacional. En la siguiente sección se presenta el experimento realizado, en la sección 3 se mostrará la planificación del experimento y en la sección 4 se presentará el análisis de los resultados del experimento. Por último en la última sección se pueden encontrar las conclusiones que se derivan del presente artículo.

2. Experimento controlado

El objetivo de nuestro estudio empírico es determinar si los diseños de almacenes de datos basados en diagramas de estrella son más comprensibles que aquellos realizados basándose en el modelo relacional desde el punto de vista de su uso, observando la influencia del dominio que representan los esquemas. Tras este objetivo subyace nuestra hipótesis de alto nivel, que consiste en evaluar si hay diferencia en el tiempo utilizado en realizar unas consultas sobre esquemas relacionales y en estrella.

3. Planificación del experimento

Nuestro estudio consiste en un experimento controlado realizado a los alumnos de doctorado de la Escuela Superior de Informática de Ciudad Real (UCLM) y dos réplicas del mismo experimento realizadas a los alumnos de tercer curso (matriculados en la asignatura de bases de datos) y los alumnos de doctorado de la Universidad de Pinar del Río (Cuba). Todos nuestros sujetos tenían conocimientos de diseño y uso de bases de datos, además los alumnos de doctorado de Ciudad Real y Pinar del Río tenían conocimientos del diseño y uso de los

almacenes de datos pues habían recibido esta información como parte de sus estudios de doctorado.

El experimento consiste en tres diagramas diseñados según el modelo relacional y otros tres diagramas semánticamente equivalentes a los anteriores usando diagramas en estrella, sobre los que había que realizar unas consultas en SQL. Los diagramas representaban problemas reales y sus dominios y eran lo suficientemente sencillos como para ser comprendidos fácilmente.

El objetivo del experimento era analizar si había diferencia en los tiempos de respuesta entre los diagramas (relacionales y en estrella) semánticamente equivalentes para analizar si alguno resultaba más complejo.

3.1. Hipótesis

H_{01}: No hay diferencia entre los sujetos que usan los dos tipos de esquemas (Relacional y Estrella) con respecto al tiempo

H_{02}: No hay diferencia entre los sujetos que usan los tres tipos de dominios con respecto al tiempo

H_{03}: No hay diferencia entre los distintos tratamientos (Combinación Tipo_Esquema x Dominio) con respecto al tiempo

3.2. Variables

Las variables independientes son las variables sobre las que los efectos deben ser evaluados, en nuestros experimentos estas variables corresponden con el dominio de los esquemas (DOMINIO) y el tipo de diseño utilizado para crearlos (TIPO_ESQUEMA). La variable dependiente es aquella que hemos medido en nuestro experimento, en nuestro caso es el tiempo utilizado para responder a las cuestiones planteadas a los sujetos.

3.3. Diseño

Teniendo en cuenta las hipótesis realizamos un diseño experimental en el que teníamos tres esquemas relacionales y tres esquemas semánticamente equivalentes diseñados mediante diagramas en estrella, por lo que el experimento constaba de seis esquemas. Los sujetos tenían que construir unas consultas SQL y anotar el tiempo (en segundos) que tardaban en realizarlas.

3.4. Objetos usados en el experimento

Los objetos usados fueron, como ya hemos dicho, seis modelos de datos. Para cada uno de los tres dominios existían dos esquemas semánticamente equivalentes diseñados mediante las dos técnicas de diseño. Se proporcionó a los sujetos los seis esquemas en los que se incorporaba una hoja de preguntas para que construyeran unas consultas SQL y anotaran los tiempo utilizados en responder a las cuestiones.

Los experimentos fueron realizados en una sola sesión. Antes de realizar el experimento se hizo una explicación intensiva de que tipo de problemas debían resolver, cómo se debía contestar a las preguntas y que material se estaba proporcionando para la realización del experimento. No obstante, los sujetos no tenían conocimiento de los aspectos que pretendíamos estudiar ni cuales eran las hipótesis que se habían planteado.

3.5. Validez de los resultados

Para conseguir evitar diversas amenazas a la validez del experimento, hemos tomado una serie de medidas:

- Los sujetos de cada experimento tenían una experiencia y unos conocimientos parecidos. Aunque trabajar con estudiantes pueda parecer poro riguroso, existen

estudios que afirman que las diferencias entre alumnos y profesionales son pequeñas y los estudios con estudiantes son viables bajo ciertas condiciones (Hörst et al., 2000).

- Los dominios de los diagramas eran lo suficientemente sencillos y corrientes para que no existiese ningún problema a la hora de entenderlos.
- Para evitar los efectos de aprendizaje los esquemas fueron entregados a cada sujeto en un orden diferente.
- Los sujetos que realizaron el experimento era la primera vez que realizaban un experimento de este tipo, por lo que los efectos de la persistencia están atenuados.
- Los sujetos estaban motivados pues los ejercicios formaban parte de los conocimientos que debían adquirir en su formación.
- No se permitió que los sujetos hablaran entre ellos durante la prueba ni que pudieran copiar los resultados unos de otros.
- Todas las dudas fueron resueltas por la persona que conducía el experimento.

4. Análisis e Interpretación

Antes de proceder al análisis teníamos que establecer un valor para el nivel de significación. Para nuestro experimento hemos fijado un valor $\alpha = 0,1$ pues es una de las formas de aumentar la potencia de nuestras pruebas estadísticas (es decir, la probabilidad de rechazar nuestras hipótesis cuando éstas son falsas). Debido a nuestros objetivos del estudio, a la configuración del experimento (tenemos 3 dominios y 2 tipos de esquema para cada dominio) y a los datos recogidos debemos usar una prueba ANOVA de medidas repetidas univariante (SPSS, 1997).

En las tablas 1, 2 y 3 se muestran los resultados obtenidos de la aplicación de nuestro estadístico a los datos recogidos del experimento. Analizando el valor de significación (Columna Sig en las tablas) con el nivel de $\alpha = 0,1$ vemos que todos los valores son mayores que α y por tanto no podemos rechazar las hipótesis de que no existe diferencia en el tiempo utilizado para responder a las cuestiones con respecto al tipo de diseño (Relacional o estrella, variable TIPO_ESQUEMA), al dominio del diagrama (variable DOMINIO) y a la interacción de ambos factores (TIPO_ESQUEMA x DOMINIO).

Como conclusión del experimento podríamos deducir que no hay diferencia en la compresión de los esquemas debido al método de diseño utilizado y que por tanto da igual diseñar un almacén de datos utilizando el modelo relacional o los diagramas en estrella. Aunque estas conclusiones pueden ser debidas a que los tamaños de los esquemas utilizados en el experimento no eran muy grandes, por lo que deberemos replicar el experimento con objetos más grandes. También podemos observar que el dominio de los esquemas tampoco parece influir en la comprensión de los mismos, lo cual puede estar determinado por el tamaño de los esquemas o por que los dominios eran lo suficientemente conocidos como para no tener repercusión en la complejidad y la comprensión, por ello sería conveniente trabajar con esquemas reales, que puedan determinar si podemos o no aislar la influencia del dominio en un estudio como este.

Fuente		Suma de cuadrados	gl	Media cuadrática	F	Sig.	Potencia
Intersección	Hipótesis	25530167,58	1	25530167,58	108,12	0,06	0,90
	Error	236120,30	1	236120,30			
TIPO_ESQUEMA	Hipótesis	10730,09	1	10730,09	0,01	0,95	0,10
	Error	1680826,46	1	1680826,46			
DOMINIO	Hipótesis	14558,91	2	7279,45	0,44	0,69	0,14
	Error	33043,27	2	16521,63			
TIPO_ESQ * DOMINIO	Hipótesis	8045,24	2	4022,62	2,73	0,27	0,32
	Error	2945,24	2	1472,62			

Tabla 1. Resultados del ANOVA para los alumnos de doctorado de la E.S. Informática

Fuente		Suma de cuadrados	gl	Media cuadrática	F	Sig.	Potencia
Intersección	Hipótesis	4605724,80	1	4605724,80	162,29	0,05	0,95
	Error	28378,83	1	28378,83			
TIPO_ESQUEMA	Hipótesis	8557,55	1	8557,55	0,09	0,82	0,10
	Error	98157,55	1	98157,55			
DOMINIO	Hipótesis	4643,79	2	2321,90	0,19	0,84	0,12
	Error	24705,57	2	12352,78			
TIPO_ESQ * DOMINIO	Hipótesis	9252,44	2	4626,22	0,33	0,75	0,13
	Error	27950,91	2	13975,45			

Tabla 2. Resultados del ANOVA para los alumnos de tercer curso de la E.S. Informática

Fuente		Suma de cuadrados	gl	Media cuadrática	F	Sig.	Potencia
Intersección	Hipótesis	227055223,97	1	227055223,97	30,04	0,11	0,61
	Error	7557702,31	1	7557702,31			
TIPO_ESQUEMA	Hipótesis	101117,21	1	101117,21	0,02	0,90	0,10
	Error	4183548,98	1	4183548,98			
DOMINIO	Hipótesis	4633505,47	2	2316752,73	4,29	0,19	0,41
	Error	1080115,07	2	540057,53			
TIPO_ESQ * DOMINIO	Hipótesis	1278292,95	2	639146,47	2,14	0,32	0,27
	Error	597794,23	2	298897,12			

Tabla 3. Resultados del ANOVA para los alumnos de doctorado de la Univ de Pinar del Río

5. Conclusiones

Los almacenes de datos son una de las principales tendencias empresariales en los sistemas de información pues son sistemas que ayudan en la toma de decisiones estratégicas.

Se han propuesto diversos métodos de diseño de almacenes de datos basados en los diagramas de estrella, ya que estos diagramas aumentan la eficacia y la comprensión de los esquemas de los almacenes de datos. Aunque esta afirmación es ampliamente aceptada, no se ha demostrado empíricamente que sea cierta, por lo que hemos realizado un experimento para poder comprobarlo. Como conclusión de nuestro estudio podemos decir que no parece haber diferencias entre los dos tipos de métodos de diseño y por lo tanto no podemos afirmar que los diseños en estrella sean más comprensibles.

Esta conclusión puede deberse a que los objetos utilizados en el experimento no eran muy grandes y tendremos que replicar el estudio con objetos mayores, datos reales y expertos y observar si se obtienen los resultados esperados.

Agradecimientos

Esta investigación es parte del proyecto CALDEA (TIC 2000-0024-P4-02) financiado por la Subdirección General de Proyectos de Investigación, Ministerio de Ciencia y Tecnología.

Referencias

Adamson, C. y Venerable, M. (1998) *Data Warehouse Design Solutions.* John Wiley & Sons.

Gardner, S.R. (1998). Building the data warehouse, *Communications of the ACM*, 41(9). pp. 52-60.

Hammergren, T. (1996). *Data Warehousing Building the Corporate Knowledge Base* International Thomson Computer Press, Milford.

Hörst, M., Regnell, B. y Wohlin, C. (2000). Using Students as Subjects – A Comparative Study of Students & Profesionals in Lead-Time Impact Assessment. *4th Conference on Empirical Assessment & Evaluation in Software Engineering, EASE*, Keele University, UK.

Inmon, W. H. (1997). *Building the Data Warehouse.* John Wiley and Sons, 2nd edn.

Jarke, M., Lenzerini, M., Vassilou, Y. and Vassiliadis, P. (2000). *Fundamentals of Data Warehouses.* Springer.

Kimball, R., Reeves, L., Ross, M. and Thornthwaite, W. (1998). *The Data Warehouse Lifecycle Toolkit.* John Wiley and Sons.

SPSS (1997), SPSS Advanced Statistics 7.5, SPSS, Inc.

Van Solingen, R. y Berhghout, E. (1999). *The Goal/Question/Metric Method,* McGraw-Hill.