BITS, CIENCIA Y SOCIEDAD
columnas desde el dcc

Abril 1, 2010

Codd: ¿Cómo darle un buen diseño a los datos?

Categoría: Sin categoría — Tags: bases de datos, edgar codd, Pablo Barceló — dccuchile - 2:32 pm

Por Pablo Barceló, profesor del Depto. de Ciencias de la Computación, FCFM, de la Universidad de Chile

En mi última columna, publicada hace dos meses, me referí a los algoritmos probabilistas y anuncié extenderme sobre estos en mi próxima columna. Pero como en este tiempo ha corrido algo más de agua bajo el puente, las circunstancias me motivaron a cambiar el tema en pos de algo más contingente.

De toda la maraña de información semiprocesada, y sólo ligeramente comprendida, que he recibido en este mes pos-terremoto, hay una que -casi en un proceso darwiniano- ha logrado captar mi interés como científico dedicado a los sistemas de datos: el problema de los duplicados (nombres repetidos) en la lista de fallecidos/desaparecidos por el tsunami. En particular, el gobierno reconoció dos problemas con la lista; por un lado, la existencia de personas con el mismo RUT mencionadas más de una vez y, por el otro, la existencia de personas distintas que compartían el mismo RUT.

La verdad es que no es mi intención (ni tampoco mi misión) tratar de escudriñar las razones profundas de por qué esta lista tenía errores, ni tampoco desnudar falencias de los equipos informáticos de este gobierno ni del anterior. Ciertamente el problema -sus causas y soluciones- es mucho más complejo que lo que un par de conceptos teóricos puede resolver. Sin duda esto apunta a un tema técnico/teórico muy interesante, y que es ubicuo en el estudio actual de los sistemas de información: la integridad de los datos, es decir, su veracidad y exactitud.

Imagine por un momento que usted quiere construir un modelo de datos. Para ello hay dos cosas fundamentales que debe diseñar. Primero, la manera en que los datos van a ser almacenados. Y segundo, un lenguaje para extraer la información que está implícita en esos datos. Piense, por ejemplo, en el modelo estándar de almacenamiento de información actual. En éste, los datos se almacenan en tablas y para consultarlos se utiliza algún tipo de lenguaje comercial como el SQL. Este modelo corresponde al modelo relacional de información ideado a principios de los '70 por el padre de las bases de datos, Edgar Codd.

Diseñados el esquema de los datos y el lenguaje de consulta, uno podría darse por satisfecho y echar la máquina a andar. Pero precisamente es lo que no hizo Codd: siendo un personaje visionario y genial, ya en esa época se dio cuenta que la integridad de los datos era fundamental. Y que ninguno de los dos componentes anteriores podía lidiar con este tema. Para mitigar este defecto Codd ideó una nueva componente del modelo: las restricciones de integridad. Estas permiten entregarle "semántica" a la información almacenada o, en otros términos, asegurar que la información almacenada satisfaga ciertas restricciones impuestas según lo que se quiere modelar (como tal las restricciones de integridad son metadatos, pues proveen información sobre los datos mismos). Por ejemplo: una restricción de integridad muy simple y razonable en una lista de personas sería algo así como "no pueden existir dos personas que tengan el mismo RUT". ¿Le suena familiar?

Por supuesto, los problemas de integridad de información que existen en la actualidad son mucho mayores que cualquiera que Codd haya podido imaginar (por visionario que fuera). Principalmente porque los volúmenes de datos que manejamos hoy en día no son comparables a nada antes visto en la humanidad. En particular, problemas como la consistencia, incompletitud e inexactitud de la información son pan de cada día y debemos lidiar con ellos provistos de técnicas mucho más avanzadas que sólo imponiendo restricciones de integridad. Además, tan sólo identificar cuáles son las restricciones de integridad correctas para un dominio específico es cada vez más complejo y muchas veces requiere la participación de expertos en tal dominio.

Pero no podemos culpar de esto a Codd. El murió en 2003 en la paz de su jubilación en Florida, feliz de haber contribuido de manera relevante al desarrollo de los sistemas de datos. Y probablemente no demasiado preocupado de los problemas que las nuevas fuentes de datos nos causan. Por todas sus contribuciones Codd recibió el premio Turing en 1981. Estas no se limitaban al diseño del modelo de datos que mencionamos arriba, sino también a muchas de sus extensiones más futuristas. Por ejemplo, la introducción de valores nulos para representar información incompleta o inexacta (tema que hoy acapara las conferencias de bases de datos), el diseño de formas normales para el correcto almacenamiento de información y una versión inicial de los sistemas OLAP (de procesamiento analítico online).

permalink trackback
Comentarios (0)
« Older Posts

Archivos

Otros Blogueros