Departamento de Ciencias de la Computación
UNIVERSIDAD DE CHILE

Principles of Dataspace Systems

Alon Halevy

Michael Franklin

David Maier

Maíra Marques Samary
Santiago, 21 de octubre de 2010

Agenda

- Motivación
- Desafíos
- Definiciones
- Ejemplos
- Data Integration
- Propiedades
- Respuestas de Consultas
- Introspección
- Respuestas

La web esta se tornando semántica

- Forms (millones)
- Buscadores verticales (centenas)
- Esquemas de anotaciones (facebook, flickr, games)

Motivación

- Demanda en administrar heterogéneas y múltiples fuentes de datos con distintos modelos de datos (DBMS)

Desafíos

- Encontrar fuentes de datos relevantes
- Dar acceso a búsquedas y consultas
- Trazar el origen
- Determinar la validad de la información

Definiciones

- Data Space Management System (DSMS) - abstracción para la administración de datos en ese escenario

EJEMPLOS

Personal Information Management (PIM)

- Ofrecer acceso y manipulación sencillos de toda la información de la computadora de una persona.
  - Hoy esta limitada a consultas de palabras

Personal Information Management (PIM)

- Aplicación de los principios de Dataspaces:
  - Herramienta PIM debe dar acceso a toda la información en la computadora y no solamente la implícita o explicita en el conjunto elegido
  - No se puede asumir que los usuarios van a investir su tiempo en integrar los datos de varias fuentes

Contenidos en la web

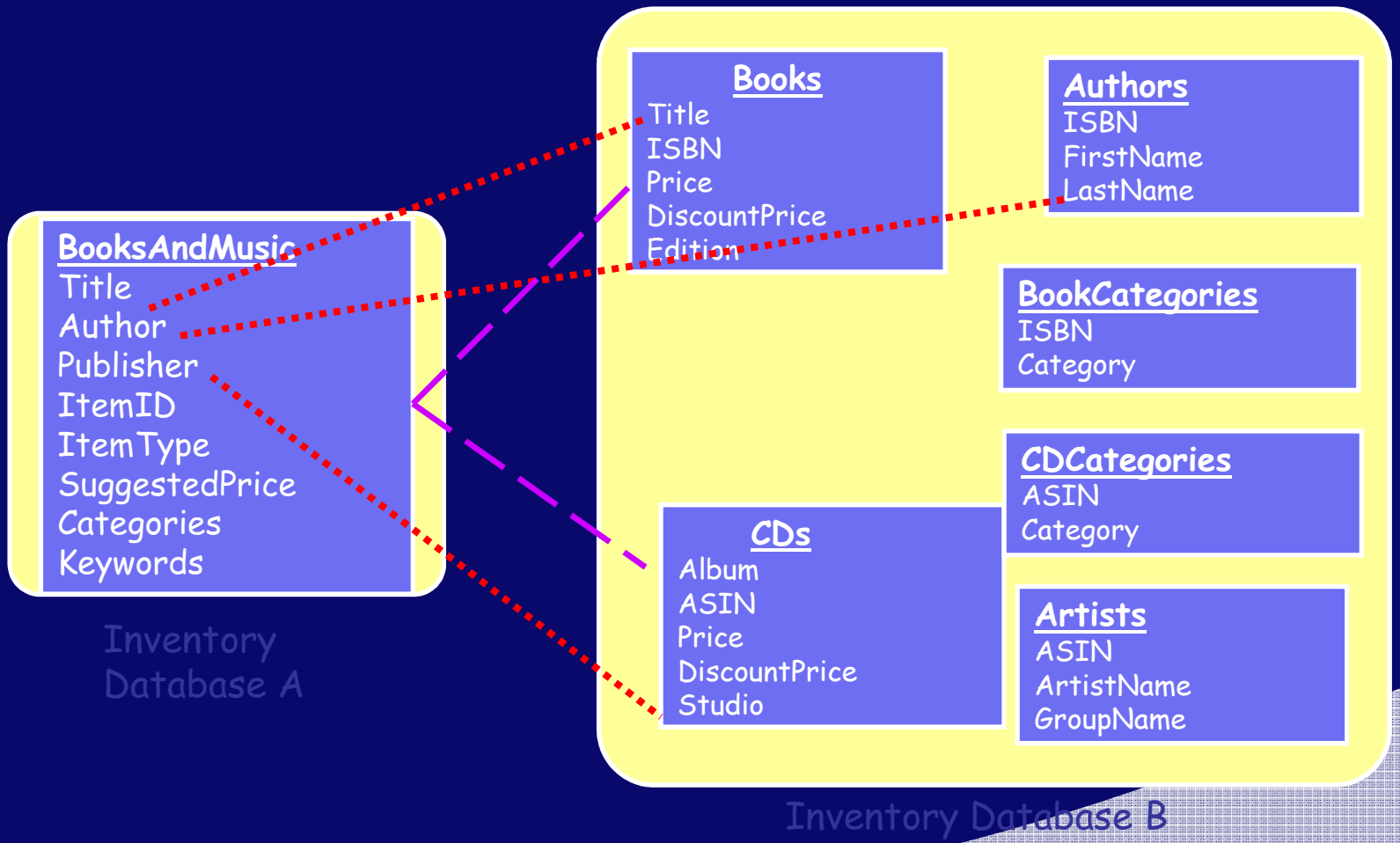
- Contenido no estructurado
  - >20% contenido porno
  - >40% contenido spam

Aplicación de los principios de Dataspaces:

- Necesidad de una herramienta de búsqueda que acepte palabras llave y seleccione fuentes estructuradas relevantes
- Combine las respuestas estructuradas y no estructuradas de una misma manera

DATA INTEGRATION

Data Integration



DataSpace x Data Integration

- Data Integration necesita una integración semántica antes que cualquier servicio sea ofrecido.
- Dataspace seria la evolución de las arquitecturas de integración
- Dataspace entrega funcionalidades básicas sobre todas las fuentes de datos sin importar como es la integración entre ellas.

PROPIEDADES

Propiedades de Dataspaces

- Trabajar con datos y aplicaciones en muchos formatos y debe ser accesible a través de muchos sistemas con distintas interfaces.

- Tiene que ofrecer una manera integrada de búsquedas, consultas, actualizaciones y administración. Al mismo tiempo que el mismo dato tiene que ser accesible y modificable por un interfaz nativa del sistema de hosting

Propiedades de Dataspaces

- Varios niveles de servicio y en algunos casos puede retornar “best-effort” o respuestas aproximadas.
- Debe ofrecer herramientas y caminos para crear una integración mas justa de los datos en el espacio se necesario.

Participantes y relaciones

- El modelo de un Dataspaces es un conjunto de participantes y relaciones
  - Participantes - fuentes de datos individuales (xml, txt, csc, databases)
  - Relaciones – pueden ser dos o mas participantes

RESPUESTAS DE CONSULTAS

Consultas

- Hechas en distintos lenguajes
- Consideran todos los datos relevantes en el Dataspace sin importar el modelo de datos, a menos que sea especificado de otra manera

Respuestas

- Ranqueadas
- Heterogénea
- Fuentes como respuestas
- Iterativa
- Reflectada

Desafíos en términos de modelos de datos y consultas

- Desarrollo de un modelo formal para estudiar las respuestas de las consultas en Dataspaces

- Desarrollo de métodos para responder consultas de múltiples fuentes que no se basen solamente en aplicar un conjunto de semántica correcta

INTROSPECCIÓN

Introspección

- Introspección es una necesidad en Dataspaces
  - Incerteza
  - Inconsistencia
  - Linaje

Incerteza

- El exacto estado del mundo no es conocido.
- El objetivo de una base de datos incierta es representar un conjunto de los posibles estados del mundo (mundos posibles)

Incerteza – A-tupla

(Karina Powers, {345-9934 | 345-9935})

(George Flores, 674-9912)

- Teléfono de Karina es uno de los dos
- Si George tiene un teléfono su numero es conocido

- No son cerradas por los operadores relacionales

Incerteza – X-tupla

(Karina Powers, 345-9934, 345-9935
|(Karina Powers, 345-9935, 345-9934)

- No sabemos cual es el teléfono de Karina y cual es el fax

- No son cerradas por los operadores relacionales

Incerteza – C-tabla

Karina Powers	456-3214	X = 1
Karina Powers	654-1234	X ≠ 1
Karina Powers	456-4444	X ≠ 1

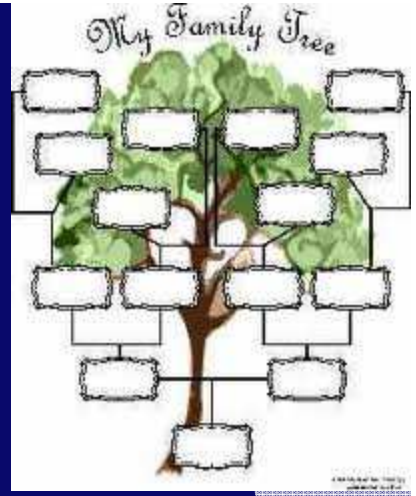
- Existen dos mundos de acuerdo con los valores de x

- Son cerradas por los operadores relacionales
- Chequear si un conjunto de tuplas I es un posible mundo para D es un problema Np-completo

Inconsistencias

- Es la incerteza de cual valor es el correcto
- Bases de datos inconsistente trabajan con situaciones donde hay datos confrontantes
- Considerar los distintos tipos de reparación
- El conjunto de mundos posibles corresponde a las distintas reparaciones

Linaje



- Externa – tuplas inseridas
- Interna – tuplas de respuestas a consultas
  - Árbol de prueba
    - Para consultas con agregación o negación no existe una definición obvia de linaje

Desafíos en términos de incerteza, inconsistencia y linaje

- Desarrollar formalismo que sea capaz de modelar incerteza, inconsistencia y linaje

RESPUESTAS

Respuestas

- ¿Que es una buena respuesta para una consulta?
  - Relevancia para la consulta
  - Certeza de la respuesta
  - Completitud y precisión
  - Tempo de latencia máximo en la respuesta

Desafíos en términos de respuestas

- Definir métricas para comparar la calidad de las respuestas y de los conjuntos de respuestas de los Dataspaces
- Técnicas eficientes de procesamiento de consultas

Desafíos en términos de la atención humana

- Desarrollar métodos para capturar la interacción de las actividades de los en los Dataspaces
- Analizar esas actividades para crear relaciones significativas entre fuentes en los Dataspaces

Thank You!