



DECSAI

Departamento de Ciencias de la Computación e I.A.

Universidad de Granada



Integración de datos

Fernando Berzal, berzal@acm.org

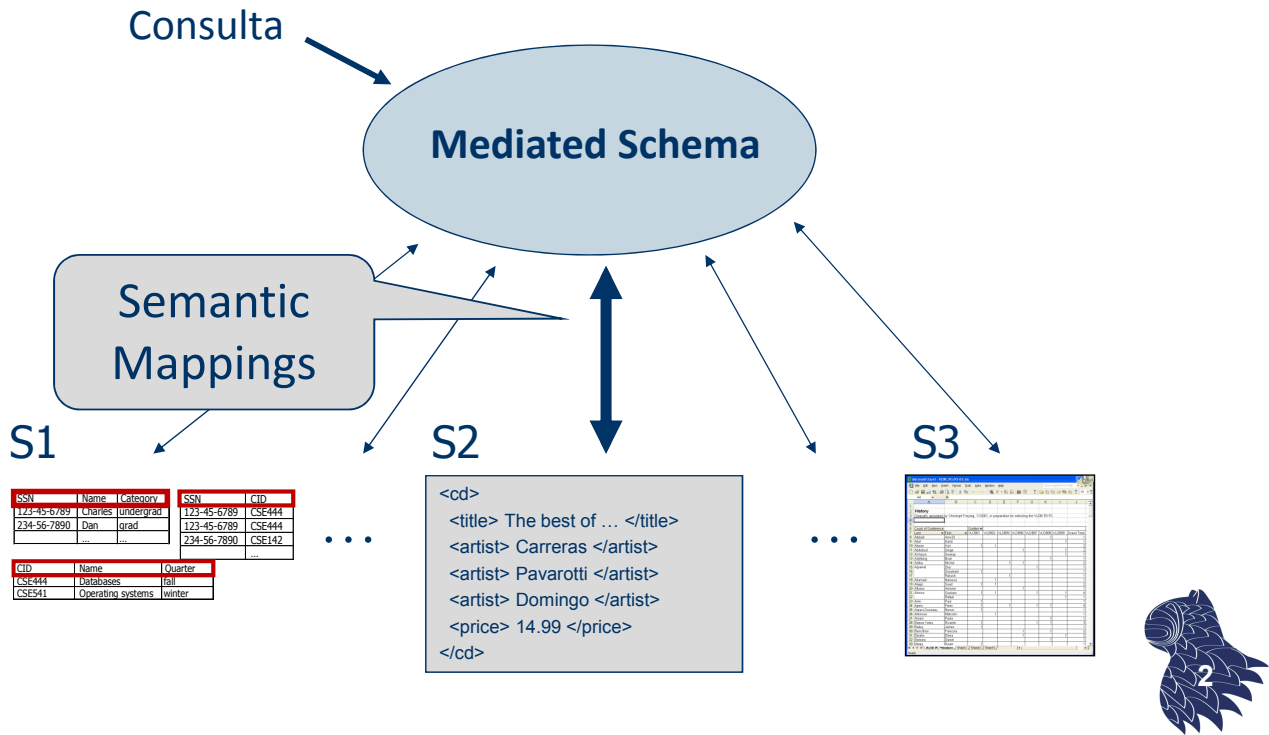
Integración de datos



- Integración de datos
- Descripción de fuentes de datos
 - Heterogeneidad
 - Correspondencias entre esquemas
 - GAV
 - LAV
 - GLAV
- Integración de esquemas [schema matching & mapping]
- Emparejamiento de datos [data matching]
- Wrappers
- Apéndices:
 - Emparejamiento de cadenas [string matching]
 - Procesamiento de consultas



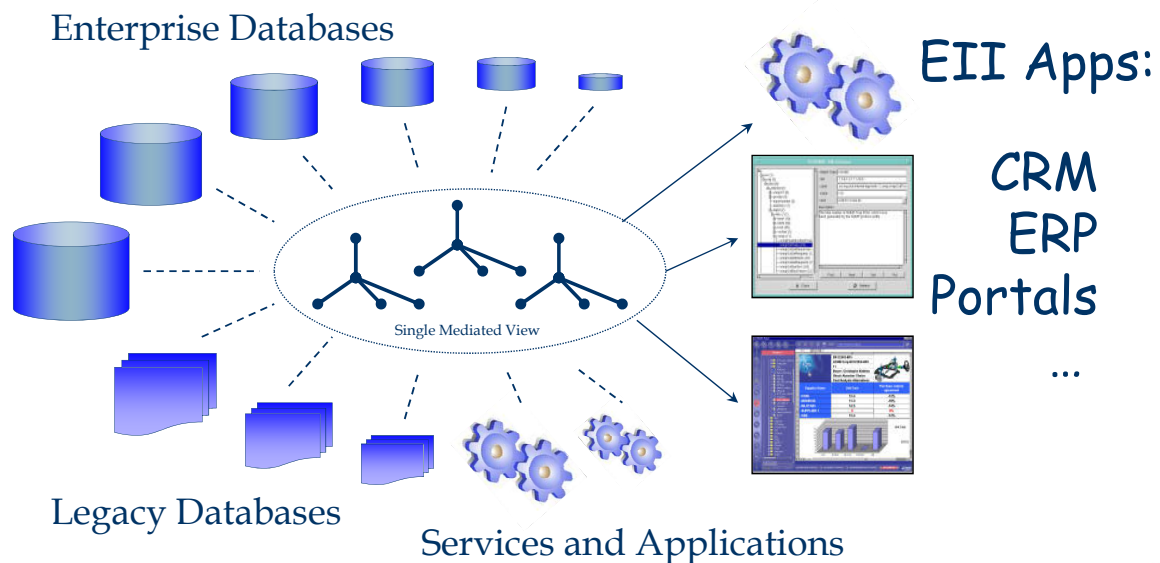
Integración de datos



Integración de datos



Aplicaciones: Empresas



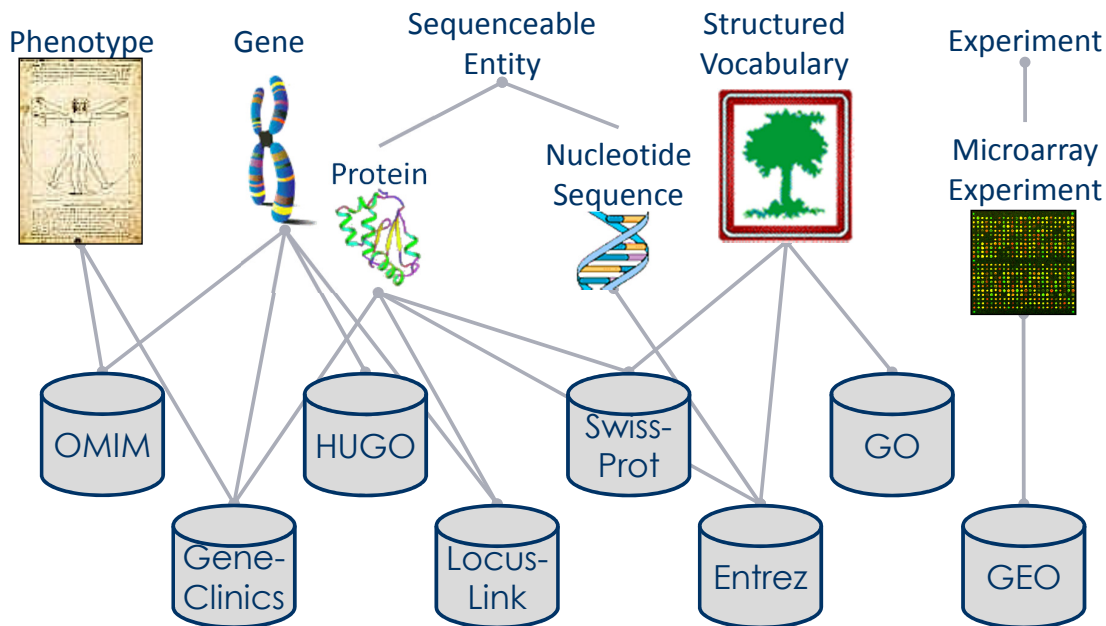
50% del gasto total en IT



Integración de datos



Aplicaciones: Ciencia



Cientos de fuentes de datos biomédicas



Integración de datos



Aplicaciones: Web



Integración de datos

Aplicaciones: Web

The screenshot shows the INEbase website interface. The header includes the INE logo and navigation tabs for 'EI INE', 'Metodos y proyectos', 'Planes', 'Censo electoral', 'INEbase', 'Formación y estadísticas', 'Productos y servicios', and 'Sede electrónica'. The main content area is titled 'Mercado laboral' and lists various statistical operations such as 'Encuesta de población activa', 'Estadística de flujos de la población activa', 'Proyecciones de tasas de actividad', and 'El empleo de las personas con discapacidad'. Each operation includes details like the date or year and a link to 'Últimos datos' or 'Información detallada'.

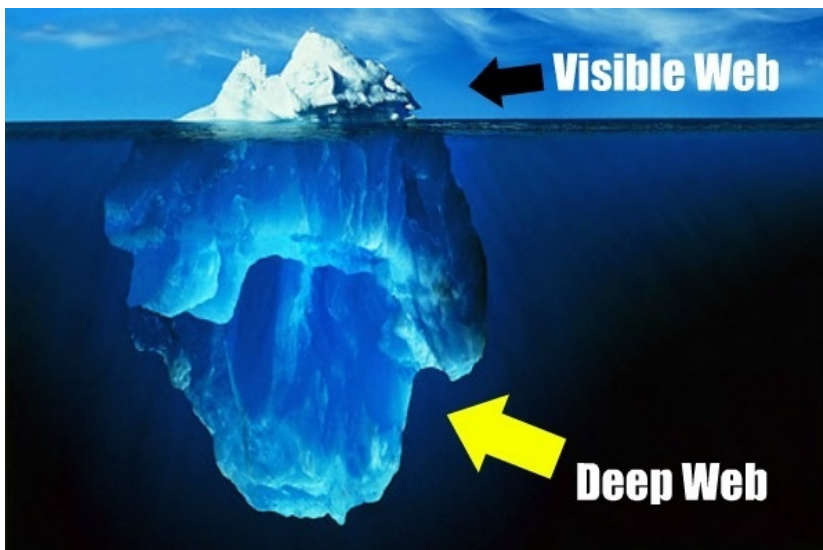
The screenshot shows a web browser displaying a data table from the BDE website. The table is titled '11. ADMINISTRACIONES PÚBLICAS' and '11.4 Pasivos en circulación y deuda según el Protocolo de Déficit Excesivo (PDE). Importes'. It contains multiple columns for different categories of public debt and circulation, with rows of numerical data. The table is presented in a standard web browser layout with a search bar and navigation icons at the top.

Cientos de millones de tablas en la Web...



Integración de datos

Aplicaciones: Deep Web



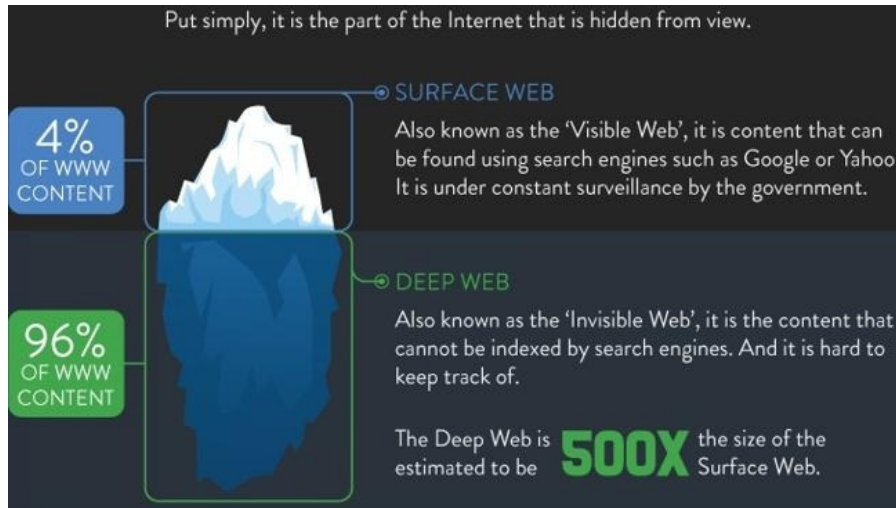
Millones de formularios que dan acceso a fuentes de datos...



Integración de datos



Aplicaciones: Deep Web



Millones de formularios que dan acceso a fuentes de datos...

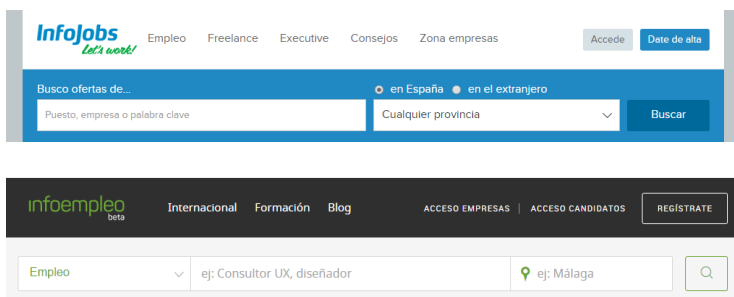
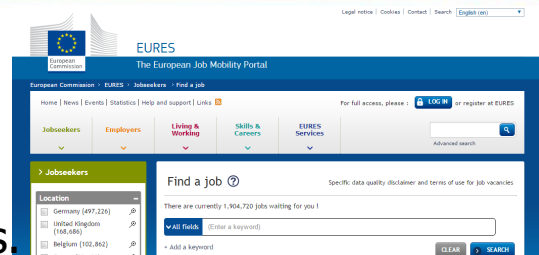


Integración de datos



Aplicaciones: Deep Web

Cada formulario tiene su propia interfaz, por lo que resulta difícil explorar datos de distintas fuentes.



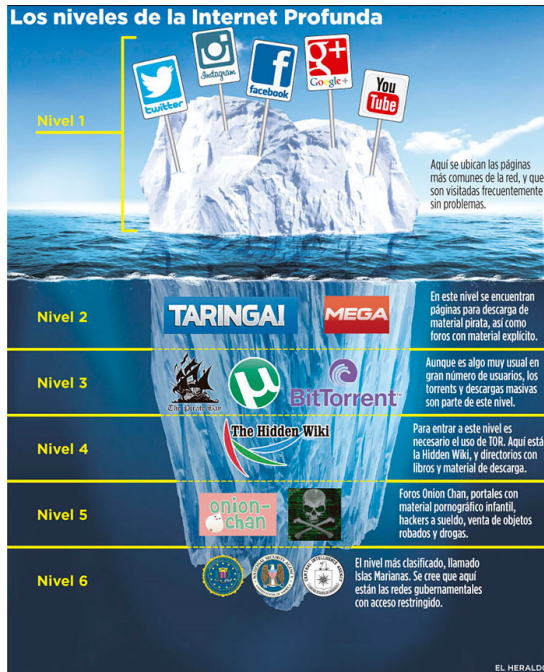
Objetivo (para dominios concretos):
Interfaz única para múltiples fuentes de datos.



Integración de datos



Deep Web vs. Dark Web



sociedad general de autores y editores

The Software Alliance

BSA



Integración de datos



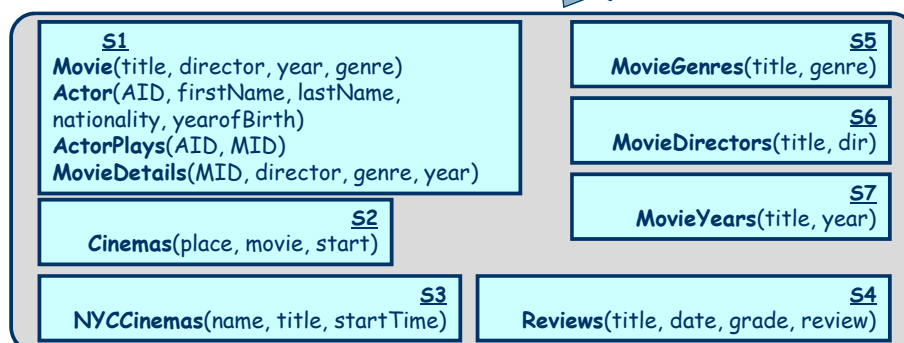
Acceso unificado a múltiples fuentes de datos

Problema común

Heterogeneidad en los esquemas de datos

Movies (title, director, year, genre)
Actors (title, name)
Plays (movie, location, startTime)
Reviews (title, rating, description)

Fuentes de datos



Integración de datos



La integración de datos es difícil...

Razones técnicas

- Múltiples plataformas, no siempre con lenguajes tipo SQL.
- Procesamiento de consultas distribuidas.

Razones lógicas

- Heterogeneidad (en los esquemas y en los datos).

Razones "sociales"

- Identificación de datos relevantes en una empresa.
- **"Data fiefdoms"** [feudos de datos]
 - Convencer a las personas para que colaboren.
 - Implicaciones de privacidad y seguridad.



Integración de datos



Gestión de expectativas en un proyecto de integración de datos

- La integración de datos es **"IA-completa"** (esto es, las soluciones completamente automatizadas son poco probables, puede que imposibles).

Objetivos razonables:

- Reducir el esfuerzo necesario para la creación de una aplicación que requiera la integración de datos.
- Conseguir un rendimiento adecuado del sistema en situaciones con cierto nivel de incertidumbre.



Integración de datos



Arquitectura de un sistema de integración de datos

Alternativas de diseño:

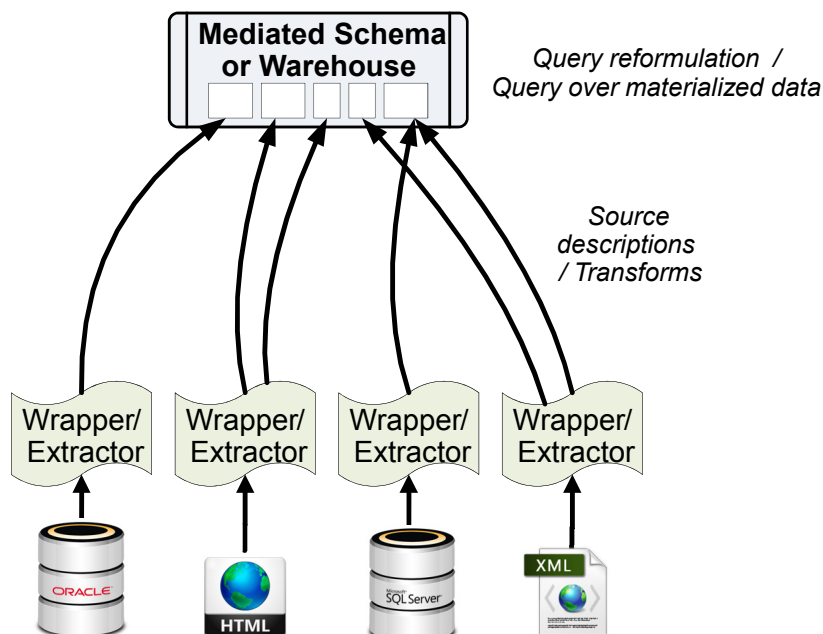
- **Data Warehouse (a.k.a. offline replication)**
(datos almacenados en una base de datos centralizada, independiente de las fuentes de datos)
- **Integración de datos virtual**
(datos en las fuentes de datos, a los que se accede al realizar consultas)



Integración de datos



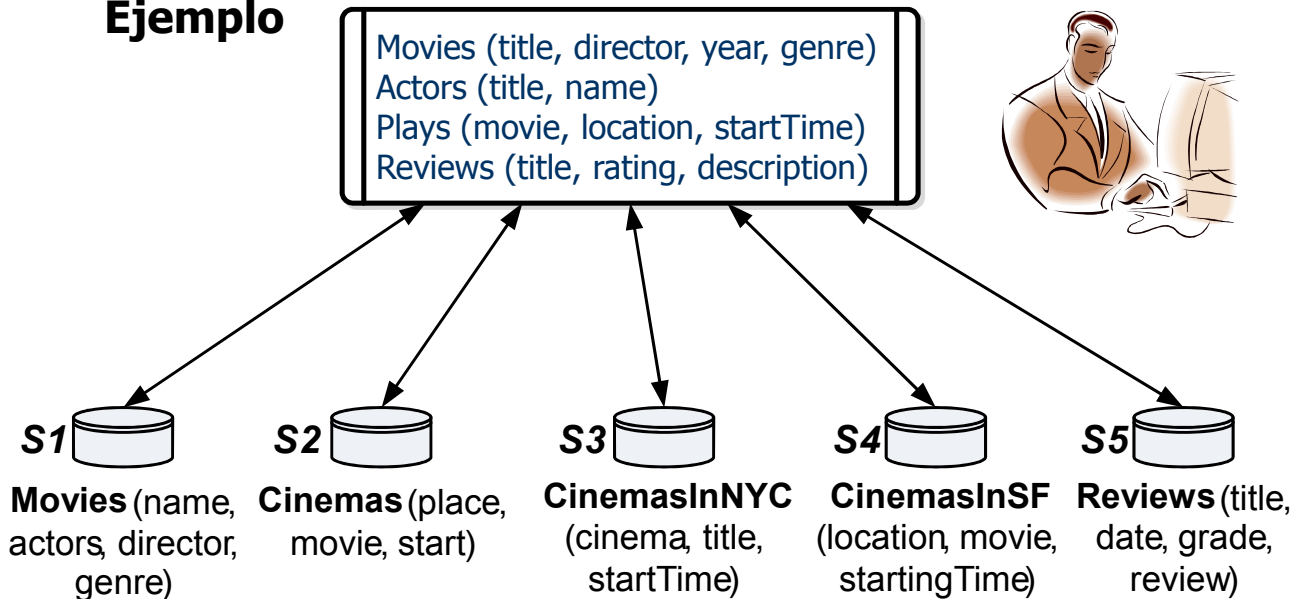
Arquitectura de un sistema de integración de datos



Integración de datos



Ejemplo



Integración de datos



Ejemplo

Comedias de Woody Allen en NY

```
Movies (title, director, year, genre)
Actors (title, name)
Plays (movie, location, startTime)
Reviews (title, rating, description)
```

```
select title, startTime
from Movies, Plays
where Movies.title=Plays.movie
AND location="New York"
AND director="Woody Allen"
```



Integración de datos



Ejemplo

Movies (title, director, year, genre)
Actors (title, name)
Plays (movie, location, startTime)
Reviews (title, rating, description)

select title, startTime
from **Movies, Plays**
where Movies.title=Plays.movie
AND location="New York"
AND director="Woody Allen"

S1	S2	S3	S4	S5
Movies (name, actors, director, genre)	Cinemas (place, movie, start)	Cinemas NYC (cinema, title, startTime)	Cinemas SF (location, movie, startingTime)	Reviews (title, date grade, review)

Fuentes de datos:

- S1 & S3 relevantes.
- S4 & S5 irrelevantes.
- S2 relevante aunque tal vez redundante.



Integración de datos



Wrappers

Enviar consultas a las fuentes de datos
y transformar las respuestas obtenidas.

2. **The Best of the Three Tenors (Audio CD)**
~ by Luciano Pavarotti, Plácido Domingo, Jose Carerras
Avg. Customer Rating: ★★★★★
([Recommend](#); [Why?](#))
Usually ships in 24 hours
List Price: ~~\$48.98~~ [Used & new](#) from \$8.95
Buy new: \$14.99

3. **The Three Tenors In Concert 1994 (Audio CD)**
~ by Jules Massenet, Federico Moreno Torroba, Richard Rodgers
Avg. Customer Rating: ★★★★★
([Recommend](#); [Why?](#))
Usually ships in 24 hours
List Price: ~~\$44.98~~ [Used & new](#) from \$1.79
Buy new: \$10.99 [Club price: \\$8.49](#)

4. **Trombonastics (Audio CD)**
~ by Joseph Alessi
Avg. Customer Rating: ★★★★★
([Rate this item](#))
Usually ships in 24 hours
List Price: ~~\$48.98~~ [Used & new](#) from \$14.23
Buy new: \$14.99

5. **The Three Tenors Christmas (Audio CD)**
~ by Carerras, Domingo, Pavarotti
Avg. Customer Rating: ★★★★★
([Recommend](#); [Why?](#))
Usually ships in 3 to 4 days
List Price: \$13.98 [Used & new](#) from \$1.89
Buy new: \$13.98



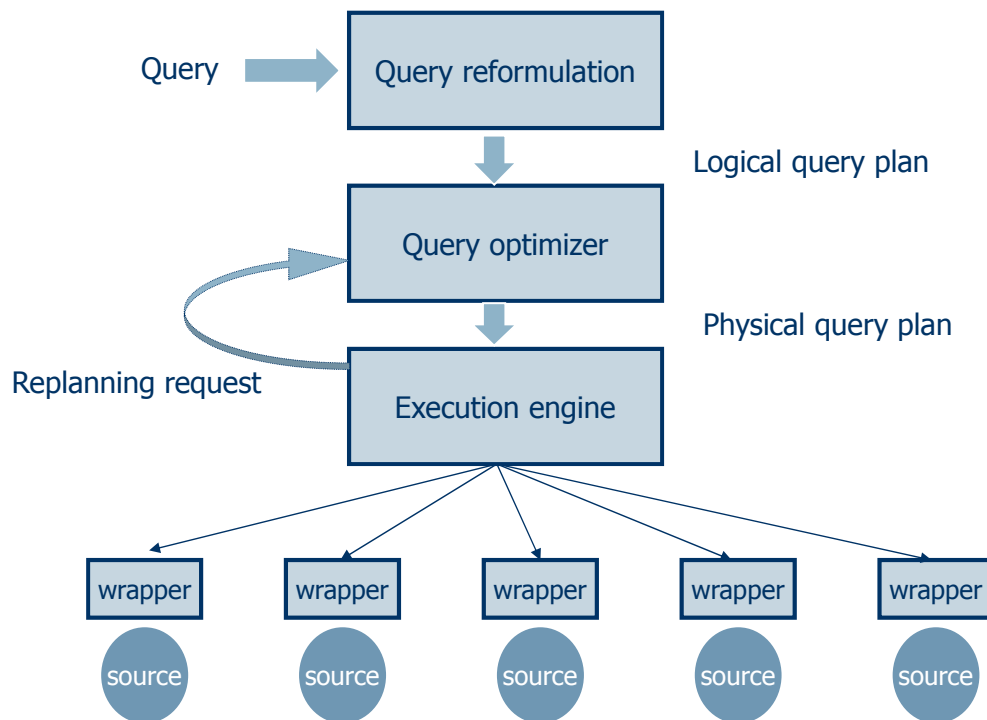
```
<cd>  
<title> The best of ... </title>  
<artist> Carreras </artist>  
<artist> Pavarotti </artist>  
<artist> Domingo </artist>  
<price> 14.99 </price>  
</cd>
```



Integración de datos



Procesamiento de consultas



Fuentes de datos



La descripción de las fuentes de datos permite que un sistema de integración de datos:

- Determine las fuentes relevantes para cada consulta.
- Acceda a las fuentes de datos de forma adecuada.
- Combine datos provenientes de múltiples fuentes.
- Supere las limitaciones de fuentes específicas.
- Identifique la forma más eficiente de procesar consultas.



Fuentes de datos



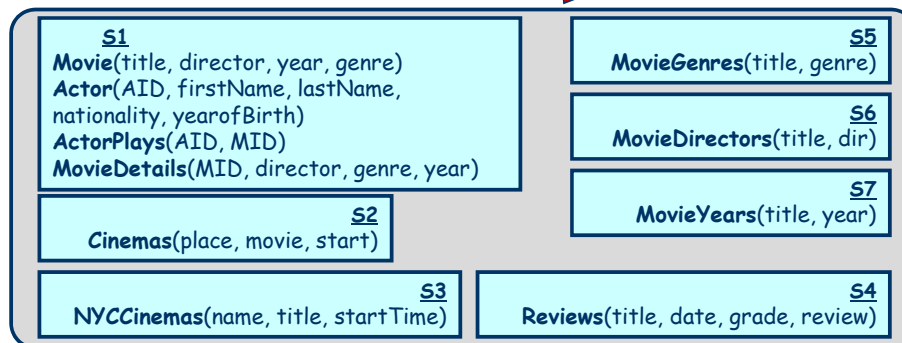
Problema

Cómo describir la relación entre el esquema integrado y las fuentes de datos

Esquema integrado

Movies (title, director, year, genre)
Actors (title, name)
Plays (movie, location, startTime)
Reviews (title, rating, description)

Fuentes de datos



22

Fuentes de datos



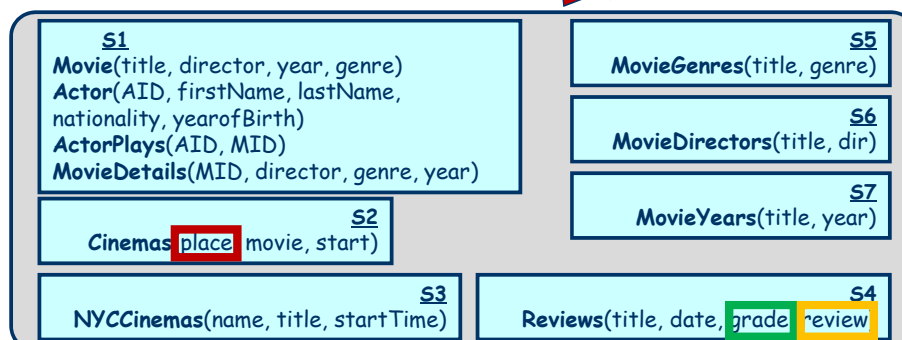
Heterogeneidad

Nombres de atributos y tablas

Esquema integrado

Movies (title, director, year, genre)
Actors (title, name)
Plays (movie, location, startTime)
Reviews (title, rating, description)

Fuentes de datos



23

Fuentes de datos

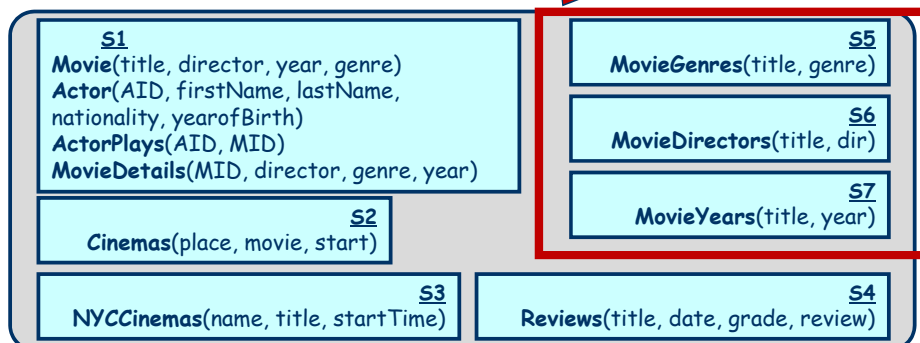


Heterogeneidad
Organización
tabular diferente

Esquema integrado

Movies (title, director, year, genre)
Actors (title, name)
Plays (movie, location, startTime)
Reviews (title, rating, description)

Fuentes de datos



24

Fuentes de datos

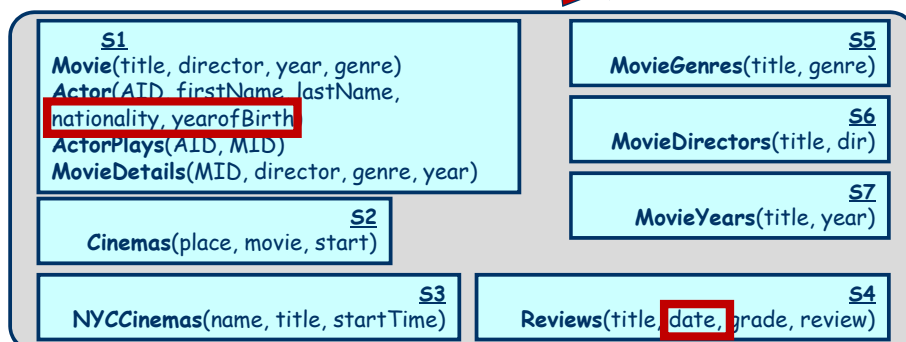


Heterogeneidad
Distinto
grado de detalle

Esquema integrado

Movies (title, director, year, genre)
Actors (title, name)
Plays (movie, location, startTime)
Reviews (title, rating, description)

Fuentes de datos



25

Fuentes de datos



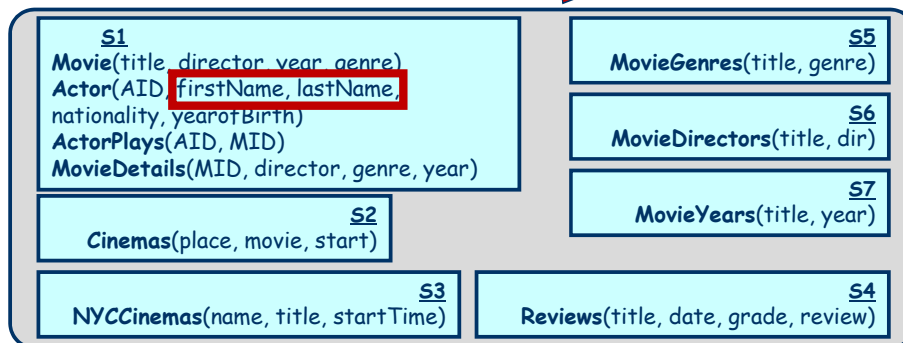
Heterogeneidad

Distinta
representación
de los datos

Esquema integrado

Movies (title, director, year, genre)
Actors (title, **name**)
Plays (movie, location, startTime)
Reviews (title, rating, description)

Fuentes de datos

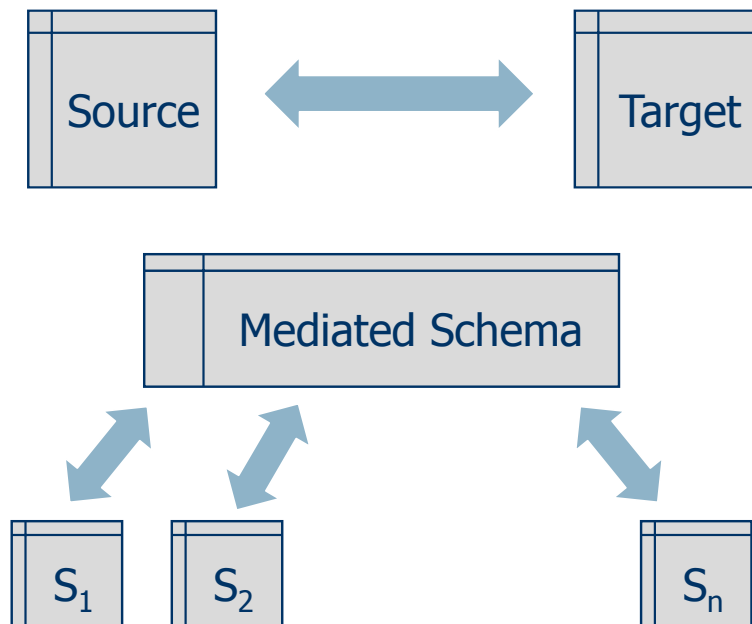


26

Fuentes de datos



Correspondencias entre esquemas



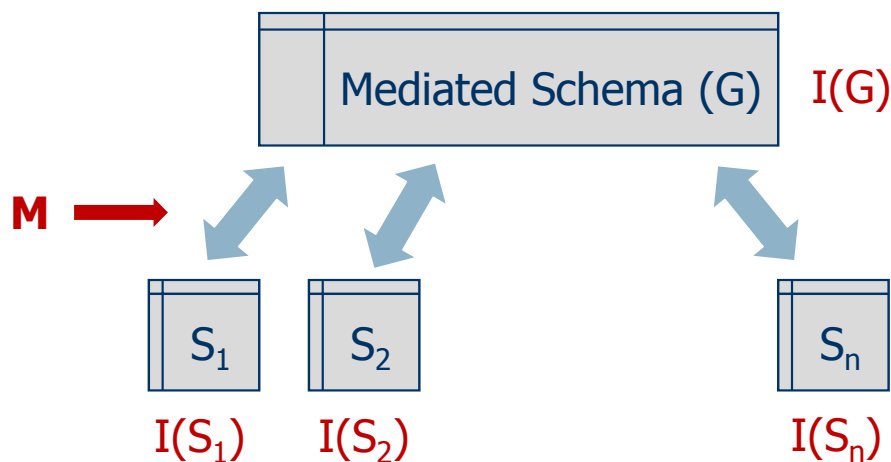
27

Fuentes de datos



Correspondencias entre esquemas

¿Qué instancias del esquema integrado son consistentes con las instancias actuales de las fuentes de datos?



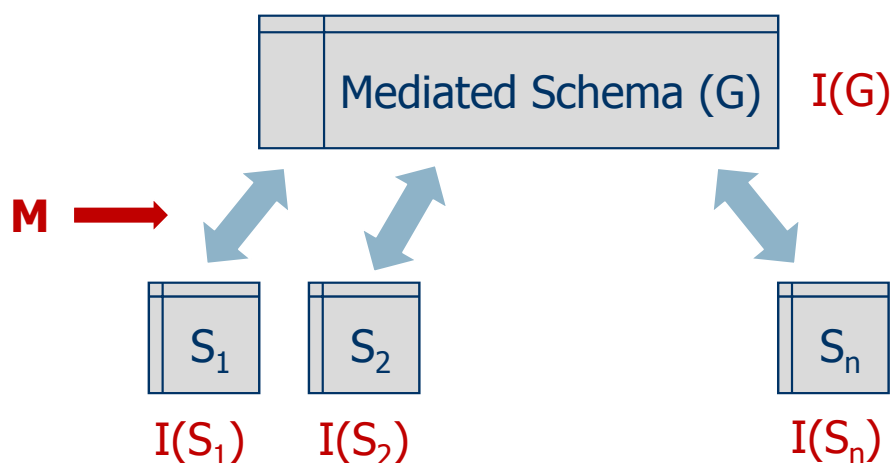
Fuentes de datos



Correspondencias entre esquemas

Formalmente,

$$M_R \subseteq I(G) \times I(S_1) \times \dots \times I(S_n)$$



Fuentes de datos



Correspondencias entre esquemas

Ejemplo

Fuente de datos (Director, Title, Year) con tuplas
{(Allen, Manhattan, 1979), (Coppola, Godfather, 1972)}

Esquema integrado (Title, Year):

- Proyección simple de la fuente de datos.
- Única instancia posible:
{(Manhattan, 1979), (Godfather, 1972)}



Fuentes de datos



Correspondencias entre esquemas

Ejemplo

Fuente de datos (Title, Year) con tuplas
{(Manhattan, 1979), (GodFather, 1972)}

Esquema integrado (Director, Title, Year):

- Una instancia posible:
{(Allen, Manhattan, 1979), (Coppola, GodFather, 1972)}
- Otra instancia posible:
{(Halevy, Manhattan, 1979), (Stonebraker, GodFather, 1972)}

Ante determinadas consultas (e.g. años de las películas),
se obtienen las respuestas correctas, ante otras no...

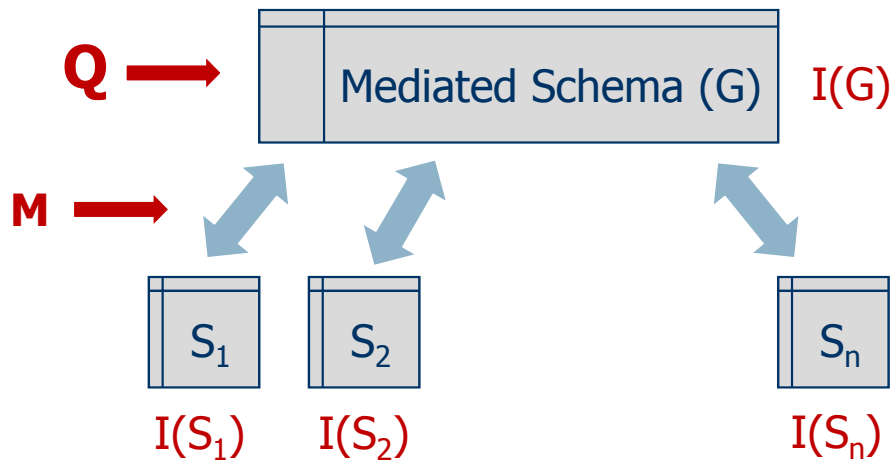


Fuentes de datos



Correspondencias entre esquemas

La respuesta a una consulta Q es certera si toda instancia del esquema integrado es consistente con las instancias de las fuentes y la correspondencia M .



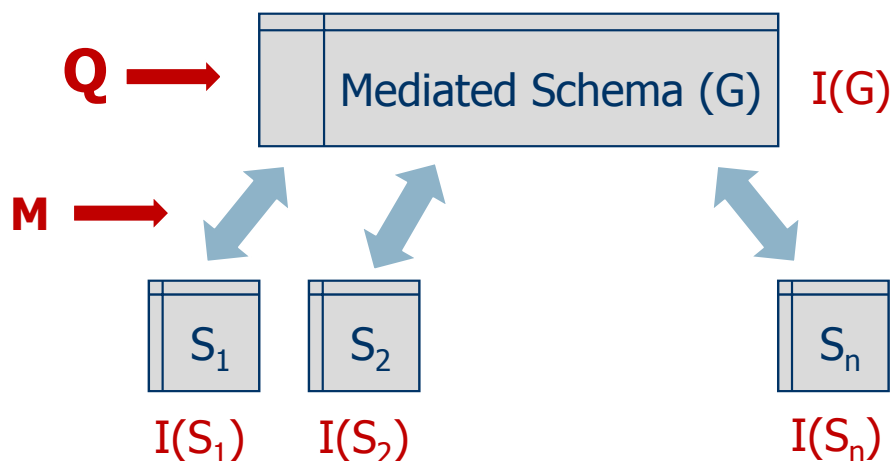
Fuentes de datos



Correspondencias entre esquemas

Formalmente,

$$t \in Q(s_1, \dots, s_n) \text{ iff } t \in Q(g) \text{ for } \forall g, \text{ s.t. } (g, s_1, \dots, s_n) \in M_R$$



Fuentes de datos



Características deseables de un lenguaje de descripción de fuentes de datos

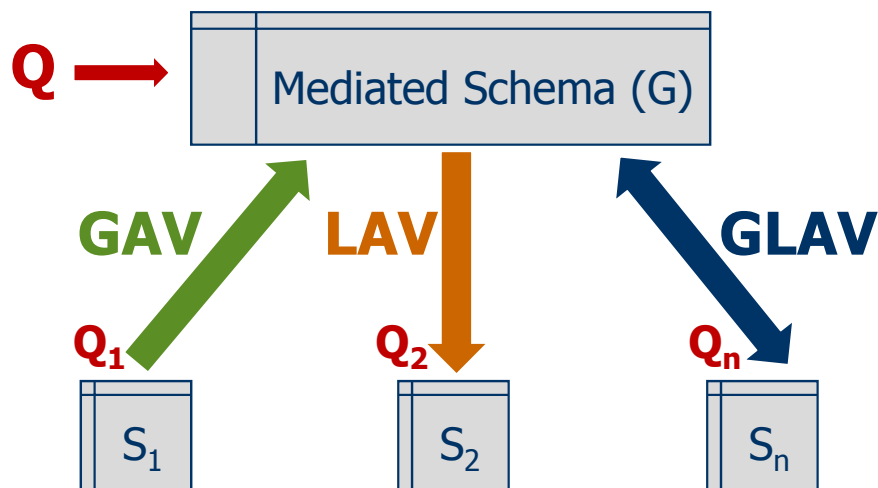
- **Flexibilidad**: Capacidad de expresar relaciones entre esquemas reales.
- **Reformulación eficiente** (complejidad computacional de la reformulación y ejecución de consultas).
- **Facilidad de actualización** (que resulte sencillo añadir y modificar fuentes de datos).



Fuentes de datos



Descripción de fuentes de datos



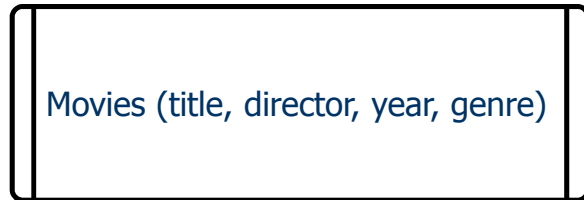
Fuentes de datos



GAV [Global-as-View]



Esquema integrado definido como un conjunto de vistas sobre las fuentes de datos.



S1
Movie (MID, title)
Actor (AID, firstName, lastName, nationality, yearofBirth)
ActorPlays (AID, MID)
MovieDetails (MID, director, genre, year)

Movies (title, director, year, genre) \supseteq
S1.Movie (MID, title) \bowtie
S1.MovieDetail (MID, director, genre, year)



GAV [Global-as-View]



Esquema integrado definido como un conjunto de vistas sobre las fuentes de datos.

Formalmente, un conjunto de expresiones de la forma

$$G_i(\bar{X}) \supseteq Q(\bar{S}) \quad \text{o bien} \quad G_i(\bar{X}) = Q(\bar{S})$$

Hipótesis de mundo abierto

Hipótesis de mundo cerrado

donde G_i es una relación del esquema integrado y Q una consulta sobre las fuentes de datos.



GAV [Global-as-View]



Ejemplo

Movies (title, director, year, genre)

Movies (title, director, year, genre) \supseteq

S1.Movie (MID, title) \bowtie

S1.MovieDetail (MID, director, genre, year)

Movies (title, director, year, genre) \supseteq

S4.MovieGenres (title, genre) \bowtie

S5.MovieDirectors (title, director) \bowtie

S6.MovieYears (title, year)



GAV [Global-as-View]



Ejemplo

Plays (movie, location, startTime)

S2

Cinemas (place, movie, start)

S3

NYCinemas (name, title, startTime)

Plays (movie, location, startTime) \supseteq

S2.Cinemas (movie, place, start)

Plays (movie, location, startTime) \supseteq

S3.NYCinemas (title, name, startTime)





Reformulación de consultas

Dada una consulta Q sobre el esquema integrado G,
determinar la mejor consulta posible
sobre las fuentes de datos S:

Q (title, location, startTime) :-
Movies (title, director, year, "comedy"),
Plays (title, location, st), st \geq 8 p.m.

Movies (title, director, year, genre) \supseteq
S1.Movie (MID, title) \bowtie
S1.MovieDetail (MID, director, genre, year)

Plays (movie, location, startTime) \supseteq
S2.Cinemas (movie, place, start)



Reformulación de consultas

Alternativa 1

Q (title, location, start) :-
Movies (title, director, year, "comedy"),
Plays (title, location, start), start \geq 8 p.m.



Q' (title, location, start) :-
S1.Movie (MID, title),
S1.MovieDetail (MID, director, "comedy", year),
S2.Cinemas (title, location, start), start \geq 8 p.m.



GAV [Global-as-View]



Reformulación de consultas

Alternativa 2

Q (title, location, start) :-

Movies (title, director, year, "comedy"),

Plays (title, location, start), start \geq 8 p.m.



Q' (title, location, start) :-

S1.Movie (MID, title),

S1.MovieDetail (MID, director, "comedy", year),

S3.NYCinemas(location,title,start), start \geq 8 p.m.



GAV [Global-as-View]



Semántica de GAV

$(g, s_1, \dots, s_n) \in M_R$ si

$$G_i(\bar{X}) \supseteq Q(\bar{S})$$

Hipótesis de mundo abierto

La extensión de G_i en g es un superconjunto de evaluar Q_i sobre las fuentes de datos.

$$G_i(\bar{X}) = Q(\bar{S})$$

Hipótesis de mundo cerrado

La extensión de G_i en g es igual al resultado de evaluar Q_i sobre las fuentes de datos.





Ejemplo problemático

Movies (title, director, year, genre)
Actors (title, name)
Plays (movie, location, startTime)
Reviews (title, rating, description)

S8
ActorDirector (actor, director)

Actors (NULL, actor) \supseteq
S8.ActorDirector (actor, director)

Movies (NULL, director, NULL, NULL) \supseteq
S8.ActorDirector (actor, director)



Ejemplo problemático

Actors (NULL, actor) \supseteq
S8.ActorDirector (actor, director)

Movies (NULL, director, NULL, NULL) \supseteq
S8.ActorDirector (actor, director)

Dadas las tuplas de S8 {(Keaton, Allen), (Pacino, Coppola)},
aparecerían las siguientes tuplas en el esquema integrado:

Actors { (NULL, Keaton), (NULL, Pacino) }
Movies { (NULL, Allen, NULL, NULL),
(NULL, Coppola, NULL, NULL) }





Ejemplo problemático

Actors (NULL, actor) \supseteq

S8.ActorDirector (actor, director)

Movies (NULL, director, NULL, NULL) \supseteq

S8.ActorDirector (actor, director)

No se pueden resolver consultas del tipo:



Q(actor, director) :-

Actors (**title**, actor),

Movies (**title**, director, genre, year).

LAV [Local-as-View] resolverá el problema...



Resumen

- El esquema integrado se define como un conjunto de vistas sobre las fuentes de datos.
- La reformulación de consultas es conceptualmente sencilla (reformulación en tiempo polinómico).
- GAV fuerza que todo se vea desde la perspectiva del esquema integrado: no puede capturar determinadas organizaciones de las fuentes de datos.



LAV [Local-as-View]



Fuentes de datos definidas como vistas sobre el esquema integrado.

S5
MovieGenres (title, genre)

S6
MovieDirectors (title, director)

S7
MovieYears (title, year)

Movies (title, director, year, genre)
Actors (title, name)
Plays (movie, location, startTime)
Reviews (title, rating, description)

S5.MovieGenres (title, genre)
 \subseteq Movies (title, director, year, genre)

S6.MovieDirectors (title, director)
 \subseteq Movies (title, director, year, genre)

S5.MovieYears (title, year)
 \subseteq Movies (title, director, year, genre)



LAV [Local-as-View]



Fuentes de datos definidas como vistas sobre el esquema integrado.

S8
ActorDirector (actor, director)

Movies (title, director, year, genre)
Actors (title, name)
Plays (movie, location, startTime)
Reviews (title, rating, description)

S8.ActorDirector (actor, director) \subseteq
Movies (title, director, year, genre)
 \bowtie Actors(title, actor)



LAV [Local-as-View]



Fuentes de datos definidas como vistas sobre el esquema integrado.

Formalmente, un conjunto de expresiones de la forma

$$S_i(\bar{X}) \subseteq Q_i(G) \text{ o bien } S_i(\bar{X}) = Q_i(G)$$

Hipótesis de mundo abierto

Hipótesis de mundo cerrado

donde S_i es una relación en una fuente de datos y $Q_i(G)$ una consulta sobre el esquema integrado.



LAV [Local-as-View]



Semántica de LAV

$(g, s_1, \dots, s_n) \in M_R$ si

$$S_i(\bar{X}) \subseteq Q_i(G)$$

Hipótesis de mundo abierto

El resultado de Q_i sobre g en un superconjunto de s_i

$$S_i(\bar{X}) = Q_i(G)$$

Hipótesis de mundo cerrado

El resultado de Q_i sobre g es igual a s_i



LAV [Local-as-View]



A diferencia de GAV, las definiciones LAV implican la existencia de un conjunto de posibles bases de datos para el esquema integrado.

Ejemplo $S8.ActorDirector(actor, director) \subseteq$
 $Movies(title, director, year, genre)$
 $\bowtie Actors(title, actor)$

$S8: \{ (Keaton, Allen) \}$

Dos posibles bases de datos:

- $Movie: \{ ("manhattan", allen, 1979, comedy) \}$
 $Actor: \{ ("manhattan", keaton) \}$
- $Movie: \{ ("foobar", allen, 1979, comedy) \}$
 $Actor: \{ ("foobar", keaton) \}$



LAV [Local-as-View]



Ya que las fuentes de datos pueden ser incompletas, otras tuplas pueden aparecer en la instancia del esquema integrado.

Ejemplo $S8.ActorDirector(actor, director) \subseteq$
 $Movies(title, director, year, genre)$
 $\bowtie Actors(title, actor)$

$S8: \{ (Keaton, Allen) \}$

Tuplas adicionales (de otras fuentes):

- $Movie: \{ (manhattan, allen, 1981, comedy),$
 $(the\ godfather, coppola, 1972, drama) \}$
- $Actor: \{ (manhattan, keaton),$
 $(the\ godfather, keaton) \}$



LAV [Local-as-View]



Realización de consultas

$S8.ActorDirector(actor, director) \subseteq$
Movies (title, director, year, genre),
Actors(title, actor).

$S8: \{ (Keaton, Allen) \}$

$Q(actor, director) :-$
Movies (title, director, year, genre),
Actors (title, actor).



$Q'(actor, director) :-$
 $S8.ActorDirector(actor, director).$



54

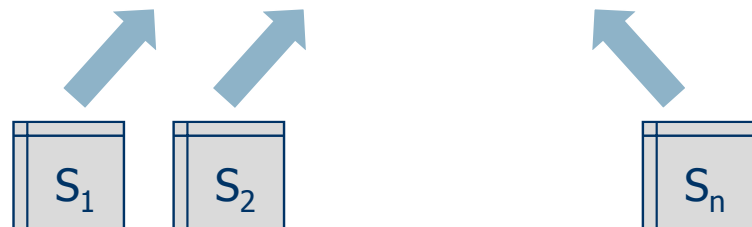
LAV [Local-as-View]



Realización de consultas



Dadas las tuplas de las fuentes de datos (expresadas como vistas sobre un esquema integrado [sin tuplas]),



se realizan consultas sobre el esquema integrado.

Problema:

Responder consultas a partir de vistas
(problema ya estudiado en bases de datos).



55



Realización de consultas

Responder consultas a partir de vistas

$$S_i(\overline{X}) \subseteq Q_i(G)$$

Sea Q una consulta sobre G :

Encontrar todas las respuestas a Q es un problema co-NP-duro sobre el tamaño del conjunto de datos si las consultas incluyen uniones o negaciones.



Resumen

- Las fuentes de datos se definen como vistas sobre el esquema integrado.
- Reformulación de consultas = Consultas sobre vistas
 - Suele funcionar bien en la práctica.
 - Garantiza encontrar todas las respuestas (bajo determinadas condiciones).
- LAV permite manejar información incompleta (GAV sólo admite que una única instancia del esquema integrado sea consistente con las fuentes).



LAV [Local-as-View]



Limitación de LAV

Movies (title, director, year, genre)

S1
Movie (MID, title)
Actor (AID, firstName, lastName, nationality, yearofBirth)
ActorPlays (AID, MID)
MovieDetails
(MID, director, genre, year)

Si una clave (e.g. MID) es interna a una fuente de datos, LAV no permite utilizarla...



GLAV [Global-and-Local-as-View]



Combinando GAV y LAV...

Conjunto de expresiones de la forma

$$Q^S(\bar{X}) \subseteq Q^G(\bar{X}) \quad \text{o bien} \quad Q^S(\bar{X}) = Q^G(\bar{X})$$

donde Q^G es una consulta sobre el esquema integrado
y Q^S es una consulta sobre las fuentes de datos.

S1.Movie(MID,title), S1.MovieDetail(MID,dir,genre,year)
 \subseteq Movies(title,dir,genre,year)



GLAV [Global-and-Local-as-View]

Reformulación de consultas en GLAV

$$Q^S(\bar{X}) \subseteq Q^G(\bar{X})$$

Dada una consulta Q:

- Q': Reescribimos Q usando las vistas $Q_1^G \dots Q_n^G$
- Q'': Reemplazamos Q_i^G por Q_i^S en Q'
- Aplicamos Q'' sobre las fuentes de datos ($Q_1^S \dots Q_n^S$)



GLAV [Global-and-Local-as-View]

Reformulación de consultas en GLAV

$S1.Movie(MID, title), S1.MovieDetail(MID, dir, genre, year)$
 $\subseteq Movies(title, dir, genre, year), year \geq 1970$

$Q(title, year) :- Movies(title, director, 'comedy', year).$

$Q'(title, year) :- Movies(title, _, 'comedy', year), year \geq 1970.$

$Q''(title, year) :-$

$S1.Movie(MID, title),$
 $S1.MovieDetail(MID, _, 'comedy', year).$



Patrones de acceso a los datos

En ocasiones, el acceso a las fuentes de datos sólo se puede realizar de formas específicas.

Ejemplos

- Formularios web (HTTP GET/POST).
- Servicios web (SOAP).
- Tipos de consultas limitados para poder controlar la carga de un sistema.

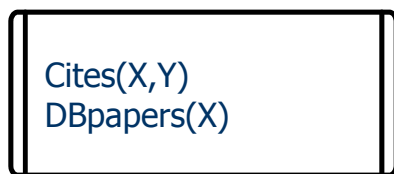
Una posible solución:

Modelar las limitaciones etiquetando las relaciones de las fuentes de datos.



Patrones de acceso a los datos

Ejemplo: b (bound) vs. f (free)





$$S1: CitationDB^{bf}(X, Y) \subseteq Cites(X, Y)$$

$$S2: CitingPapers^f(X) \subseteq Cites(X, Y)$$

$$S3: DBSource^f(X) \subseteq DBpapers(X)$$

Para que un plan de consulta sea ejecutable, toda variable ligada (b) debe tener un valor.

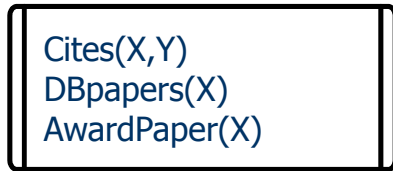
Q(X) :- Cites(X,001).

- Q'(X) :- CitationDB(X,001). 
- Q'(X) :- CitingPapers(X), CitationDB(X,001). 



Patrones de acceso a los datos

Ejemplo: b (bound) vs. f (free)



$$S1: CitationDB^{bf}(X, Y) \subseteq Cites(X, Y)$$

$$S2: CitingPapers^f(X) \subseteq Cites(X, Y)$$

$$S3: DBSource^f(X) \subseteq DBpapers(X)$$

$$S4: AwardDB^b(X) \subseteq AwardPaper(X)$$

$Q(X) :- AwardPaper(X).$

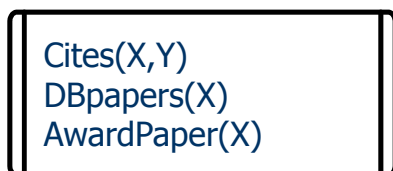
- $Q'(X) :- DBSource(X), AwardDB(X).$
- $Q'(X) :- DBSource(Y), CitationDB(Y, X), AwardDB(X).$
- $Q'(X) :- DBSource(Y), CitationDB(Y, X1), \dots$
... $CitationDB(Xn, X), AwardDB(X).$

No se terminaría nunca de reescribir la consulta...



Patrones de acceso a los datos

Ejemplo: b (bound) vs. f (free)



$$S1: CitationDB^{bf}(X, Y) \subseteq Cites(X, Y)$$

$$S2: CitingPapers^f(X) \subseteq Cites(X, Y)$$

$$S3: DBSource^f(X) \subseteq DBpapers(X)$$

$$S4: AwardDB^b(X) \subseteq AwardPaper(X)$$

... salvo que recurramos a consultas recursivas (Datalog):

$papers(X) :- DBSource(X).$

$papers(X) :- papers(Y), CitationDB(Y, X).$

$Q'(X) :- papers(X), AwardDB(X).$

Siempre se puede hacer 😊



Heterogeneidad en los datos



Un problema práctico muy habitual:
Cuando los datos provienen de distintas fuentes,
rara vez coinciden a la perfección
(lo que impide realizar reuniones, por ejemplo).

Ejemplos

- Diferencias de escala: °C vs. °F
- Precios con y sin impuestos.
- Calificaciones numéricas (7.7) vs. grados (notable).
- Granularidad: (nombre, apellido1, apellido2) vs. (nombre).
- Uso de abreviaturas: {Calle, C/}, {Doctor, Dr.}...



Heterogeneidad en los datos



Correspondencias con transformaciones

Forecast (city, day, (temp-32)*5/9, humidity)
⊆ Weather(city, temp, humidity, day)

CDStore (cd, price)
⊆ CDPrices(cd, state, price*(1+rate)),
LocalTaxes(state, rate).



Heterogeneidad en los datos



Reconciliación de referencias

Identificación de las múltiples formas mediante las que se hace referencia a una misma entidad en el mundo real.

IBM vs. International Business Machines

MSFT vs. Microsoft vs. Microsoft Corp. vs. Microsoft Corporation

F. Berzal vs. Fernando Berzal vs. Fernando Berzal Galiano

Berkeley, CA. vs. Berkeley, Calif. vs. Berkeley, California

España vs. Reino de España

¿Cómo?

Creación de tablas de concordancia
(pares de valores equivalentes)

→ **Técnicas de emparejamiento [matching]**



Emparejamiento de cadenas



Formalización del problema

Dados dos conjuntos de cadenas X e Y,
encontrar todos los pares (x,y) , con $x \in X$ e $y \in Y$,
que hacen referencia a la misma entidad en el mundo real.

Cada par (x,y) identificado será un emparejamiento [match].

<u>Set X</u>	<u>Set Y</u>	<u>Matches</u>
$x_1 = \text{Dave Smith}$	$y_1 = \text{David D. Smith}$	(x_1, y_1)
$x_2 = \text{Joe Wilson}$	$y_2 = \text{Daniel W. Smith}$	(x_3, y_2)
$x_3 = \text{Dan Smith}$		



Emparejamiento de cadenas



Desafíos prácticos

- **Precisión [accuracy]:**

Las cadenas que debemos emparejar no siempre son iguales (typos, errores de OCR, formatos diferentes, abreviaturas y omisiones, apodos, cambios de orden...).

- **Escalabilidad [scalability]:**

Emparejar cada cadena con todas las demás no es práctico, $O(n^2)$, por lo que deberemos reducir el número de comprobaciones necesario.



Medidas de similitud



Las cadenas que deseáramos emparejar no siempre aparecen de la misma forma:

- Errores mecanográficos

David vs. Davod

- Errores de OCR

datos vs. dalos

- Abreviaturas (en ocasiones, no estándar) y omisiones

Calle Real vs. C/ Real vs. C./ Real vs. Cl. Real vs. Call. Real

- Diferentes nombres y apodos

José vs. Jose vs. Pepe

- Cambios de orden en subcadenas

ETSIIT, Universidad de Granada
vs. Universidad de Granada, ETSIIT



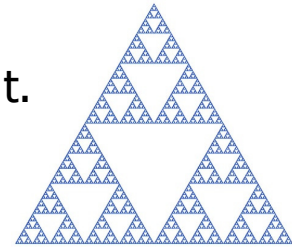
Medidas de similitud



Solución

Definir una medida de similitud $s(x,y) \in [0,1]$

- Cuanto mayor sea la similitud $s(x,y)$, mayor es la probabilidad de que x e y casen.
- Normalmente, x e y emparejan si $s(x,y) \geq t$.



NOTA

También se pueden utilizar funciones de coste o métricas de distancia: cuanto menor sea su valor, mayor es la similitud.

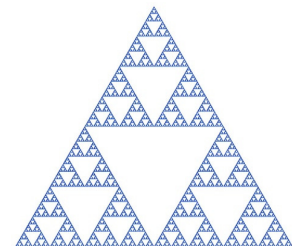


Medidas de similitud



Distintas formas de medir la similitud entre cadenas:

- **Medidas basadas en secuencias:**
Distancia de edición, Needleman-Wunch, affine gap, Smith-Waterman, Jaro, Jaro-Winkler...
- **Medidas basadas en conjuntos:**
solapamiento, Jaccard, TF/IDF...
- **Medidas híbridas** (e.g. Monge-Elkan)
- **Medidas fonéticas** (e.g. Soundex)



Bibliografía recomendada



- Hai Doan, Alon Halevy & Zachary Ives:
Principles of Data Integration
Morgan Kaufmann, 1st edition, 2012.
ISBN 0124160441
<http://research.cs.wisc.edu/dibook/>



Chapter 1: Introduction

Chapter 3: Describing Data Sources

Chapter 4: String Matching

