MDICA: Maintenance of data integrity in column-oriented database applications

Mª José Suárez-Cabal, Pablo Suárez-Otero, Claudio de la Riva, Javier Tuya

Grupo de Investigación en Ingeniería del Software (Univ. de Oviedo)
http://giis.uniovi.es

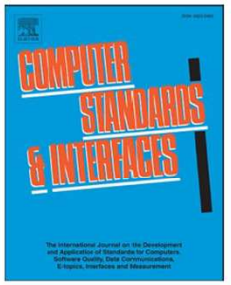
XXVIII Jornadas de Ingeniería del Software y Bases de Datos
A Coruña, 17 de Junio de 2024

Computer Standards & Interfaces 83 (2023) 103642

Contents lists available at ScienceDirect

Computer Standards & Interfaces

journal homepage: www.elsevier.com/locate/csi



MDICA: Maintenance of data integrity in column-oriented database applications

María José Suárez-Cabal *, Pablo Suárez-Otero, Claudio de la Riva, Javier Tuya

Department of Computing, University of Oviedo, Spain

DOI: 10.1016/J.CSI.2022.103642

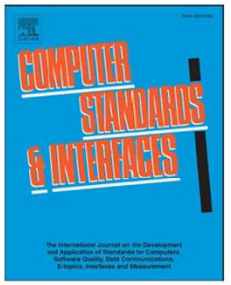
- Enviado: Octubre 2020
- Aceptado y publicado online: Abril 2022
- Publicado Open Access: Enero 2023

Agradecimientos a proyectos:

- TIN2016–76956-C3–1-R financiado por Ministerio de Economía y Competitividad
- PID2019–105455GB-C32 financiado por MCIN/AEI/10.13039/501100011033
- PID2022-137646OB-C32 financiado por MICIU/AEI/10.13039/501100011033

Contents lists available at ScienceDirect

Computer Standards & Interfaces



MD

data

Marí

Departm

Contexto

Bases de datos NoSQL

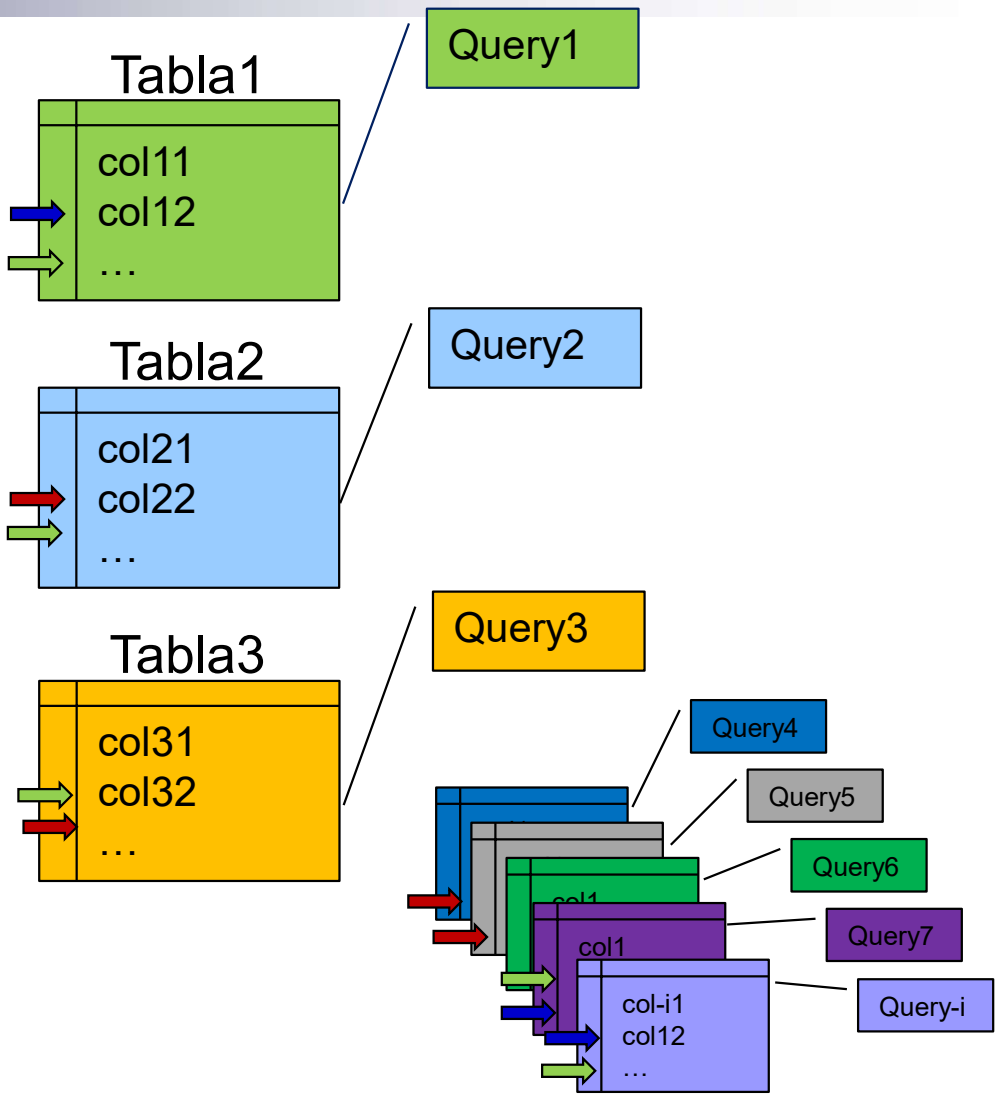
- Clave-valor
- Orientadas a columnas
- Orientadas a documentos
- Orientadas a grafos

Diseño Query-Driven

- Cada query, una tabla
- Modelo desnormalizado

Modelo conceptual no explícito

- No hay consistencia inter-modelos



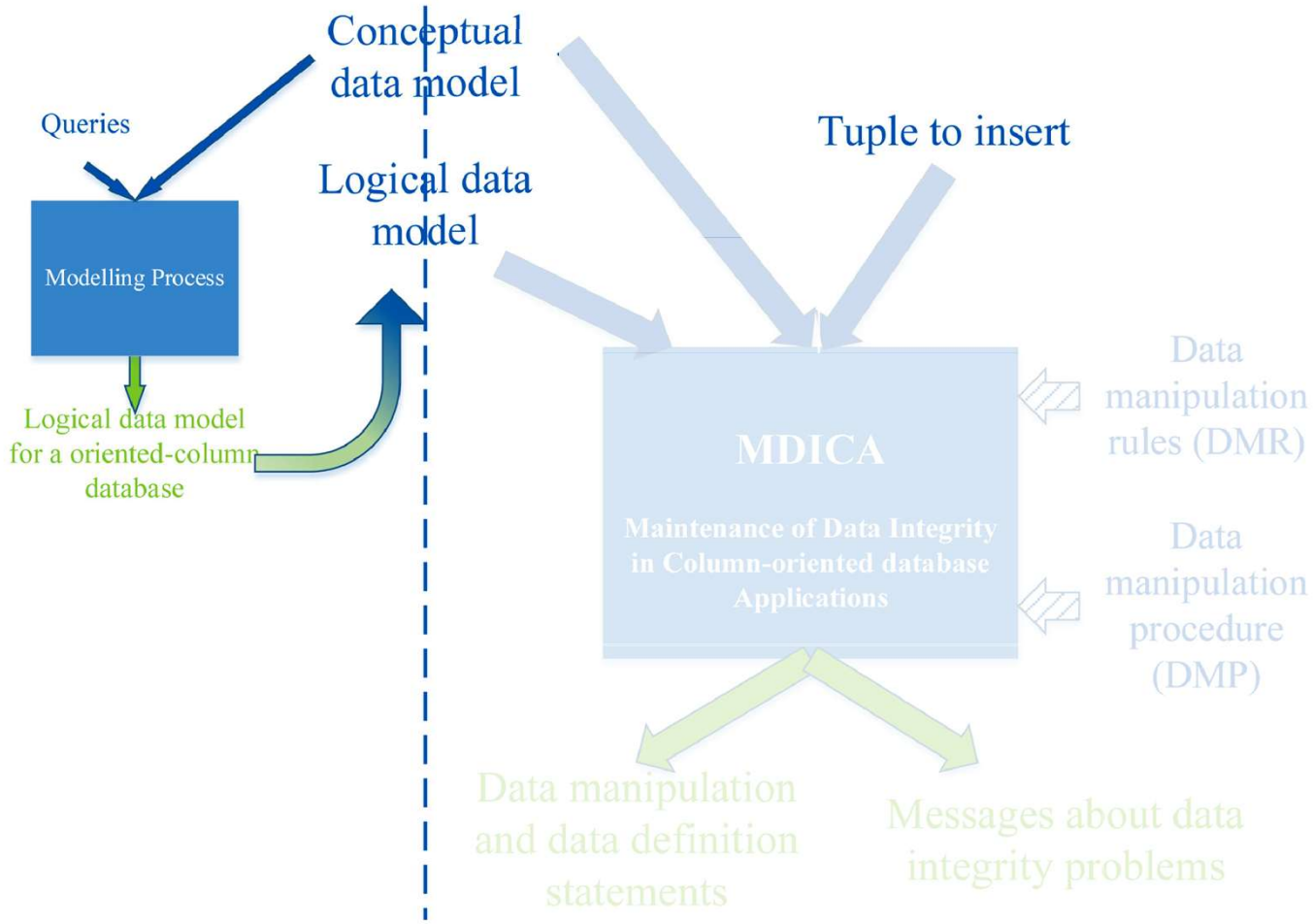
Integridad de los datos

Objetivos

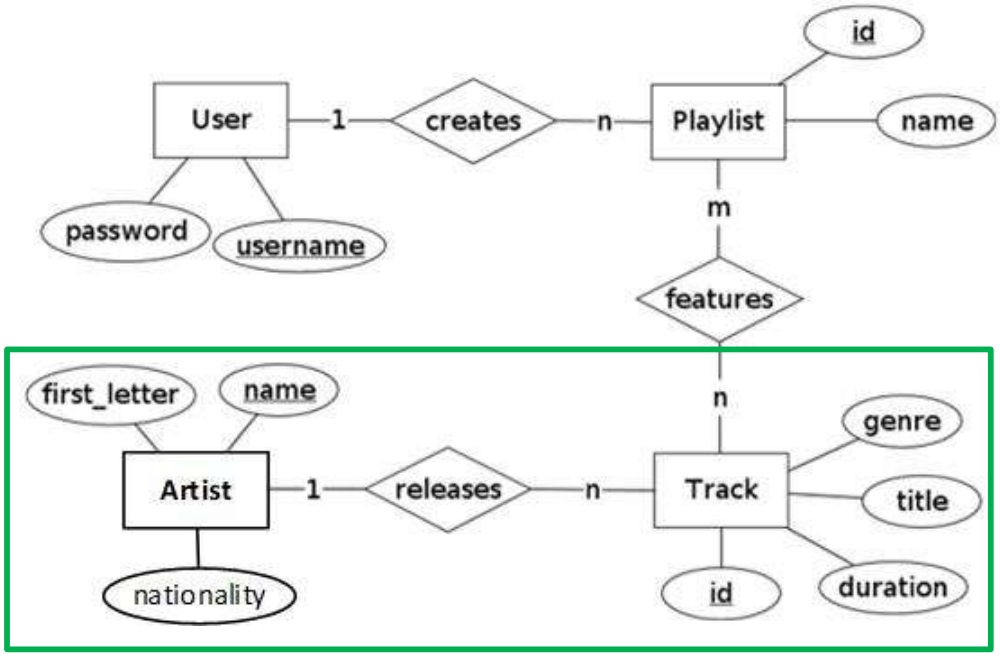
Preservar la integridad de los datos

Advertir de situaciones que puedan ponerla en peligro

Propuesta


Conceptual data model
Queries
Logical data model
Tuple to insert
Modelling Process
Logical data model for a oriented-column database
MDICA
Maintenance of Data Integrity in Column-oriented database Applications
Data manipulation rules (DMR)
Data manipulation procedure (DMP)
Data manipulation and data definition statements
Messages about data integrity problems

Insertar tupla en relación binaria



playlists_by_user (Q1)

user_username	K
playlist_id	C↑
playlist_name	

artists_by_first_letter (Q2)

artist_first_letter	K
artist_name	C↑
artist_nationality	

Track?

tracks_by_artist (Q3)

artist_name	K
track_title	C↑
track_id	C↑
track_duration	
track_genre	
artist_nationality	

tracks_by_genre (Q4)

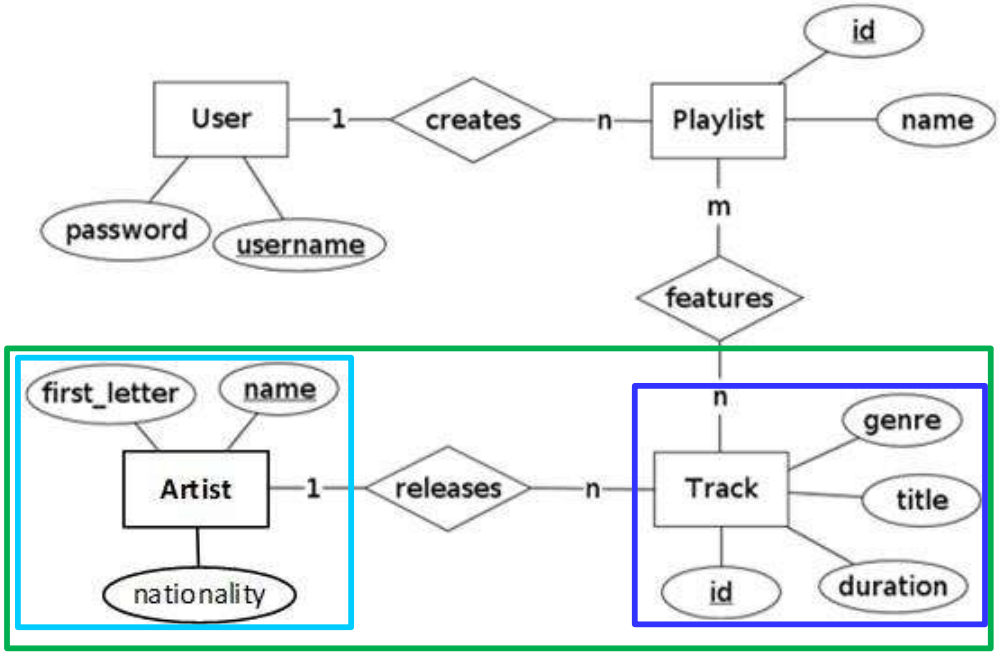
track_genre	K
track_title	C↑
track_id	C↑
track_duration	
artist_name	
artist_nationality	

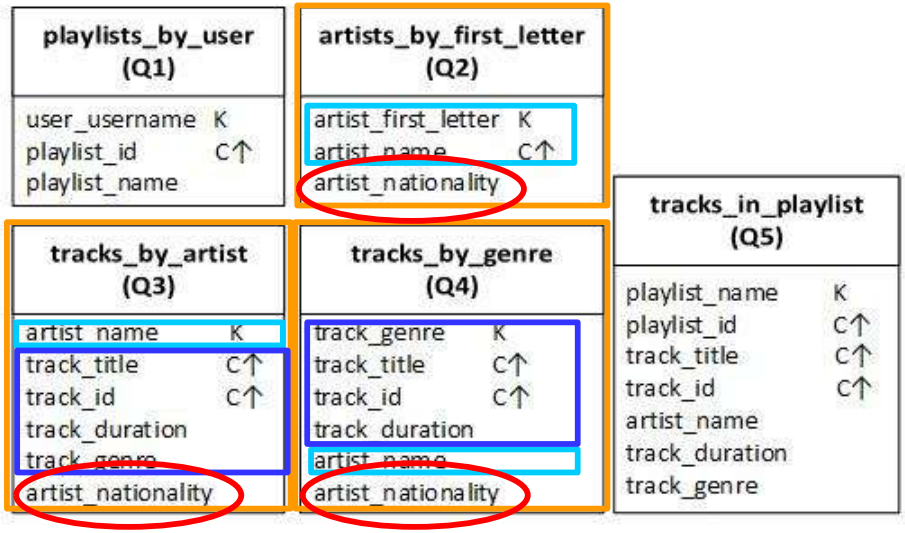
tracks_in_playlist (Q5)

playlist_name	K
playlist_id	C↑
track_title	C↑
track_id	C↑
artist_name	
track_duration	
track_genre	

Tupla a insertar: <(artist.name, “author21”), (artist.first_letter, “a”), (track.id, “id021”), (track.title, “title21”), (track.genre, “genre21”), (track.duration, 21)>

Insertar tupla en relación binaria





Tupla a insertar: <(artist.name, “author21”), (artist.first_letter, “a”), (track.id, “id021”), (track.title, “title21”), (track.genre, “genre21”), (track.duration, 21)>

Insertar tupla en relación binaria

Tupla a insertar: <(artist.name, “author21”), (artist.first_letter, “a”),
(track.id, “id021”), (track.title, “title21”), (track.genre, “genre21”),
(track.duration, 21)>

Warning(ATA): Absence of target tables for entity Track

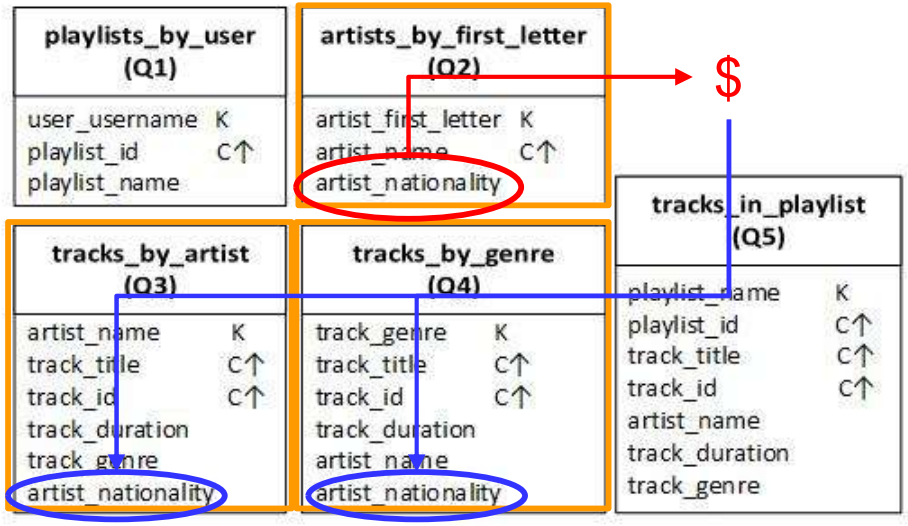
Information(ADC-S): Absence of data for column *artist_nationality*
Select *artist_nationality* from table *artists_by_first_letter*

$ = SELECT artist_nationality FROM artists_by_first_letter WHERE artist_name= “author21” and artist_first_letter=“a”

INSERT INTO artists_by_first_letter (first_letter, artist_name, artist_nationality) VALUES (“a”, “author21”,$)

INSERT INTO tracks_by_artist (artist_name, track_title, track_id, track_genre, track_duration, artist_nationality) VALUES (“author21”, “title21”, “id21”, “genre21”, 21, $)

INSERT INTO tracks_by_genre (track_genre, track_title, track_id, track_duration, artist_name, artist_nationality) VALUES (“genre21”, “title21”, “id21”, 21, “author21”, $)



Experimentos para validación

	Modelo conceptual		Modelo lógico
Caso de estudio	**Item**	**#item**	**#tablas**
Digital Library Portal	Entidad	4	-
	1:n	3	4
	n:m	3	3
			2*
Hotel Reservations	Entidad	7	2
	1:n	3	-
	n:m	3	2
			5*
Digital Music Store	Entidad	4	1
	1:n	2	3
	n:m	1	-
			1*

* Múltiples ítems

A. Chebotko, A. Kashlev, S. Lu.
A big data modeling methodology for apache Cassandra.
IEEE Int. Conf. on Big Data, 2015, pp. 238–245

J. Carpenter, E. Hewitt.
Cassandra: The Definitive Guide, 2nd ed. O'Reilly Media, Inc., 2016.

The Playlist tutorial
https://docs.datastax.com/en/archived/playlist/doc/java/p laylistPreface.html

Experimentos: diseño de casos

	Modelo conceptual		Modelo lógico
Caso de estudio	**Item**	**#item**	**#tablas**
Digital Library Portal	Entidad	4	-
	1:n	3	4
	n:m	3	3
			2*
Hotel Reservations	Entidad	7	2
	1:n	3	
	n:m	3	2
			5*
Digital Music Store	Entidad	4	1
	1:n	2	3
	n:m	1	
			1*

* Múltiples ítems

¿Dónde insertar una tupla?

- Entidades
- Relaciones
- Múltiples ítems

¿Qué información insertar?

- Todos los atributos
- Sólo claves
- -PK
- -Atributo

Experimentos: diseño de casos

	Modelo conceptual		Modelo lógico	Diseño casos
Caso de estudio	**Item**	**#item**	**#tablas**	**#casos**
Digital Library Portal	Entidad	4	-	25
	1:n	3	4	32
	n:m	3	3	31
			2*	30
				118
Hotel Reservations	Entidad	7	2	35
	1:n	3		26
	n:m	3	2	21
			5*	36
				118
Digital Music Store	Entidad	4	1	19
	1:n	2	3	15
	n:m	1		8
			1*	11
				53
Total		**30**	**23**	**289**

* Múltiples ítems

Resultados: inserción de tuplas

	Inserciones válidas		Inserciones con mensajes de error						Total
Caso de estudio			ATT		AKA		AKC		
	#	%	#	%	#	%	#	%	#
Digital Library Portal	54	45.8	37	31.4	26	22.0	1	0.8	118
Hotel Reservations	55	46,6	32	27.1	29	24.6	2	1.7	118
Digital Music Store	21	39.6	17	32.1	13	24.5	2	3.8	53
Total	**130**	**45.0**	**86**	**29.8**	**68**	**23.5**	**5**	**1.7**	**289**

Claves en columnas

Claves en atributos

Tablas objetivo

Resultados: operaciones

Caso de estudio	#inser.	INSERT			SELECT			CREATE & COPY			Total	
		#	%	Avg	#	%	Avg	#	%	Avg	#	Avg
Digital Library Portal	54	144	75.0	2.7	36	18.8	0.7	12	6.3	0.2	192	3.6
Hotel Reservations	55	129	70.9	2.4	51	28.0	0.9	2	1.1	0.0	182	3.3
Digital Music Store	21	59	62.1	2.2	24	25.3	1.1	12	12.6	0.6	95	4.5
Total	**130**	**332**	**70.8**	**2.6**	**111**	**23.7**	**0.9**	**26**	**5.5**	**0.2**	**469**	**3.6**

Insert

- Mayor % de operaciones

Select

- Obtener datos de tablas para completar tupla

Create & Copy

- Nuevas tablas y copia de datos
- Ocasional en casos reales

Resultados: mensajes

Caso de estudio	Mensajes informativos						Mensajes de advertencia											
	ADC-S		ADC-C		Total		AWC		ATA		TNW-C		TNW-K		ADC		Total	
	#	Avg	#	Avg	#	Avg	#	Avg	#	Avg	#	Avg	#	Avg	#	Avg	#	Avg
Digital Library Portal (54 inserciones)	60	1.1	12	0.2	72	1.3	306	5.7	141	2.6	6	0.1	0	0.0	7	0.1	460	8.5
Hotel Reservations (55 inserciones)	71	1.3	2	0.0	73	1.3	32	0.6	154	2.8	53	1.0	55	1.0	13	0.2	307	5.6
Digital Music Store (21 inserciones)	38	1.8	12	0.6	50	2.4	25	1.2	39	1.9	0	0.0	0	0.0	1	0.0	65	3.1
Total (130 inserciones)	169	1.3	26	0.2	195	1.5	363	2.8	334	2.6	59	0.5	55	0.4	21	0.2	832	6.4

Mensajes informativos

- Ausencia de valores que podrían extraerse (SELECT, Create & Copy)

Mensajes de advertencia

- Posible pérdida de información
- Ausencia de datos para columnas
- Discrepancias entre modelos, tablas no modeladas adecuadamente

Conclusiones

Mantener integridad de datos al insertar tuplas:

- Determinación de tablas afectadas
- Propuesta de operaciones
- Mensajes

Ayuda a desarrolladores:

- Ahorro de tiempo
- Reducción de errores

Trabajos futuros

Mantener integridad de datos:

- Evolución del modelo conceptual
- Cambios a nivel lógico:
  - Inserción de datos en una tabla
  - Evolución del esquema

Inferencia del modelo conceptual a partir del modelo lógico

Agradecimientos a proyectos:
- TIN2016–76956-C3–1-R financiado por Ministerio de Economía y Competitividad
- PID2019–105455GB-C32 financiado por MCIN/AEI/10.13039/501100011033
- PID2022-137646OB-C32 financiado por MICIU/AEI/10.13039/501100011033

MDICA: Maintenance of data integrity in column-oriented database applications

Gracias por su atención
Preguntas

Mª José Suárez-Cabal

XXVIII Jornadas de Ingeniería del Software y Bases de Datos
A Coruña, 17 de Junio de 2024