Presentación final grupo 11

Calidad de datos e Información

Bruno Ottonelli - 4.954.242-1
Gabriel Rode - 4.535.978-1

Facultad de Ingeniería
Universidad de la República.

Junio 2025

FACULTAD DE INGENIERÍA | UNIVERSIDAD DE LA REPÚBLICA URUGUAY

FASE 1
Data Quality Planning

Entradas y salidas

Entradas y salidas	
Entradas	**Salidas**
	Base de datos integrada (Data at hand) (2.1.5)
	Reporte de análisis de datos (2.2.2)
	Reporte de análisis de requerimientos de usuarios (2.3.2)
	Reporte con problemas de CD (Tabla 2.7)
	Modelo de contexto (2.3.3)

CD: Calidad de datos.

Datos recibidos

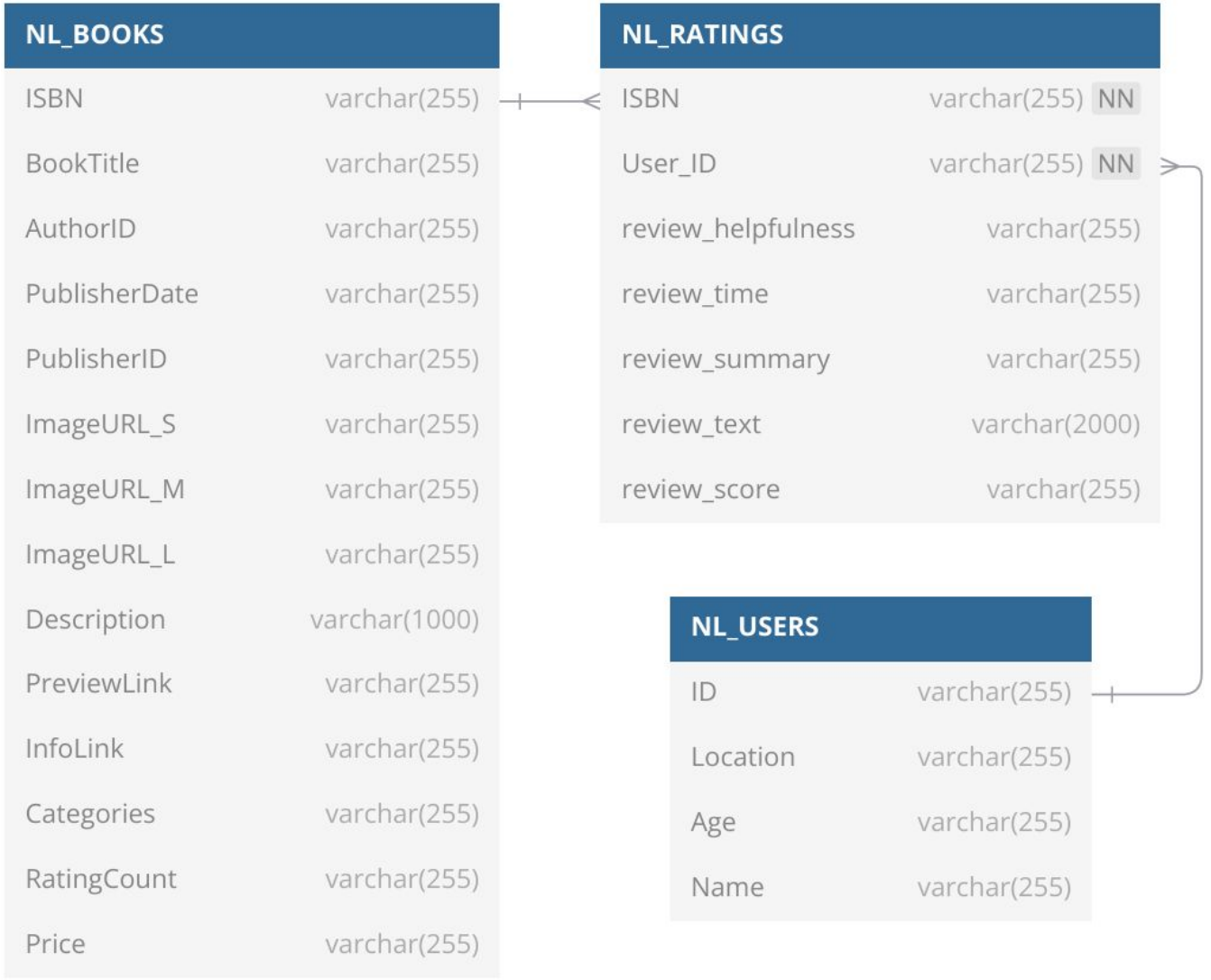
Tablas de la librería L1	
Nombre de tabla	**Atributos**
Books_rating	'Id', 'Title', 'Price', 'User_id', 'profileName', 'review/helpfulness', 'review/score', 'review/time', 'review/summary', 'review/text'
books_data	'Title', 'description', 'authors', 'image', 'previewLink', 'publisher', 'publishedDate', 'infoLink', 'categories', 'ratingsCount'

Tablas de la librería L2	
Nombre de tabla	**Atributos**
Books	'ISBN', 'Book-Title', 'Book-Author', 'Year-Of-Publication', 'Publisher', 'Image-URL-S', 'Image-URL-M', 'Image-URL-L'
Users	'User-ID', 'Location', 'Age'
Ratings	'User-ID', 'ISBN', 'Book-Rating'

Contexto

Componentes de Contexto	
Dominio	D: Libros.
Fuentes de datos	Datos obtenidos de ambas librerías que se fusionarán y los proporcionados por el cliente sobre sus realidades.
Tipos de usuario	U1: Administrador. U2: Publicista digital. U3: Analista de datos.
Tareas	T1: Gestión. T2: Análisis. T3: Consulta.
Reglas de negocio	RN1: Cada libro deberá tener asociado un ISBN, un título, al menos un autor y un editor.
Requerimientos de calidad	RQ1: Frescura de datos: la base debe actualizarse todos los viernes. RQ2: Al menos el 80 % de los usuarios que califican los libros deben ser mayores de 18 años. RQ3: Al menos el 95 % de los libros deben cumplir simultáneamente con los siguientes requisitos: contar con un ISBN, tener el título correctamente escrito y que el nombre del autor incluya al menos un nombre y un apellido. RQ4: Al menos el 60 % de los libros tengan al menos un score mayor o igual a 5. RQ5: La librería pretende tener al menos 500 libros y poseer al menos el 20 % de la lista de los 100 mejores libros de Goodreads. RQ6: Los libros deben contar con fecha de publicación. RQ7: Los libros deben tener editorial. RQ8: Los libros deben tener asignado un valor de score.
Requerimientos del sistema	RS1: Los tiempos de respuesta del sitio Web de la NL no pueden superar los 3 segundos.
Problemas de calidad ya reportados	Ninguno en particular.
Necesidades de filtrado	F1: Libros por fecha (en particular, del año actual). F2: Libros por editorial. F3: Top de libros según su score.

Data at hand


NL_BOOKS
ISBN varchar(255)
BookTitle varchar(255)
AuthorID varchar(255)
PublisherDate varchar(255)
PublisherID varchar(255)
ImageURL_S varchar(255)
ImageURL_M varchar(255)
ImageURL_L varchar(255)
Description varchar(1000)
PreviewLink varchar(255)
InfoLink varchar(255)
Categories varchar(255)
RatingCount varchar(255)
Price varchar(255)
NL_RATINGS
ISBN varchar(255) NN
User_ID varchar(255) NN
review_helpfulness varchar(255)
review_time varchar(255)
review_summary varchar(255)
review_text varchar(2000)
review_score varchar(255)
NL_USERS
ID varchar(255)
Location varchar(255)
Age varchar(255)
Name varchar(255)

Data profiling y Problemas de calidad

Problemas detectados en la Calidad de los Datos	
Campo	**Problema de calidad**
NL_Books.ISBN	P1: Entradas no respetan el formato ISBN ya que algunos están en formato ASID.
NL_Books.PublisherDate	P2: Las fechas tienen distinto formato.
NL_Books.PublisherID	P3: Mismo publisher escrito de forma distinta.
NL_Books.Title	P4: Títulos mal escritos.
NL_Books.AuthorID	P5: Mismo autor escrito de forma distinta.
NL_USERS.Age	P6: Valores de edad poco coherentes (por ejemplo 0).
NL_USERS.Location	P7: Ciudades mal escritas.
NL_RATINGS.review_time	P8: Formato de fecha/hora inconsistente.
NL_RATINGS.review_score	P9: Valores no numéricos en la puntuación.
NL_RATINGS.review_score	P10: Los valores importados entre L1 y L2 manejan distintas escalas (L1 puntúa de 0 a 5 y L2 de 0 a 10).
Base de datos	P11: Gran cantidad de nulos en muchos de los atributos.
NL_Books	P12: No hay campo de rating promedio del libro (podría calcularse).
NL_RATINGS.Helpfullness	P13: Este atributo en realidad debería ser dos atributos diferentes: cantidad de votaciones en esa review y cantidad de votaciones que consideraron útil esa review.
NL_Books.AuthorID	P14: Libros indican autores de forma distinta cuando tienen más de uno (dos libros con autores {A,B} y {B,A} deben considerarse con los mismos autores).
NL_RATINGS.review_score	P15: Podría haber valores fuera del rango 0 a 10.

Nuevas componentes de contexto

Nuevas componentes de Contexto	
Reglas de negocio	RN2: El atributo *ISBN* en *NL_Books* debe ser único a cada libro. RN3: El atributo *Price* en *NL_Books* debe ser un real positivo. RN4: El atributo *Age* en *NL_Users* debe ser un entero positivo. RN5: El atributo *ID* en *NL_Users* debe ser único y no vacío.

Nuevas componentes de Contexto	
Requerimientos de calidad	RQ9: los nombres de las editoriales deben estar estandarizados. RQ10: Las reglas de formato para nombres (autores, libros, editoriales) son: primera letra del nombre propio en mayúsculas y sin punto al final. RQ11: El formato para las fechas será dd/mm/aaaa.

FASE 2
Data Quality Assessment

Entradas y salidas

Entradas y salidas	
Entradas	**Salidas**
Reporte del análisis de requerimientos de usuarios (2.3.2)	Reporte de problemas de CD priorizados (3.1.2)
Reporte del análisis de datos (2.2.2)	Modelo de CD (3.1.4)
Reporte de problemas de CD (Tabla 2.7)	Especificación de la BD de metadatos de CD (3.2.2)
Modelo de Contexto (2.3.3)	Reporte de medición de la CD (3.2.3)
	Modelo de contexto (3.2.4)
	Reporte de evaluación de CD (3.2.3)

CD: Calidad de datos.
BD: Base de datos.

DIMENSIONES

Exactitud

Completitud

Unicidad

Consistencia

Exactitud

check_isbn

check_edades

check_price

duplicated_author

date_format

Consistencia

consistencia_ratings

consistencia_fechas

missing_rating_books

Completitud

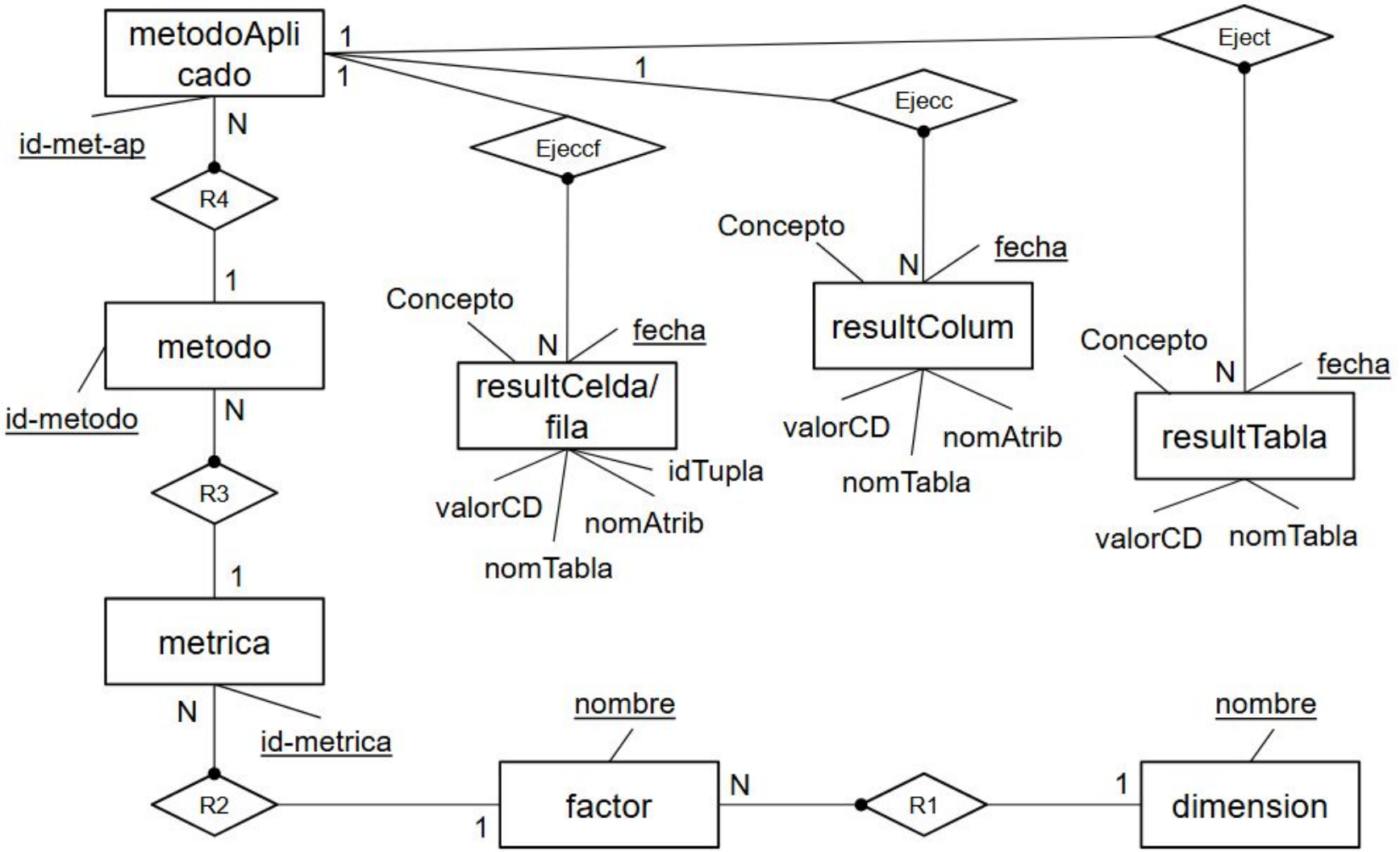
contar_nulls

check_RN1

Unicidad

duplicated_ratio

Base de datos de metadatos de la calidad de datos



Criterios para los umbrales de evaluación

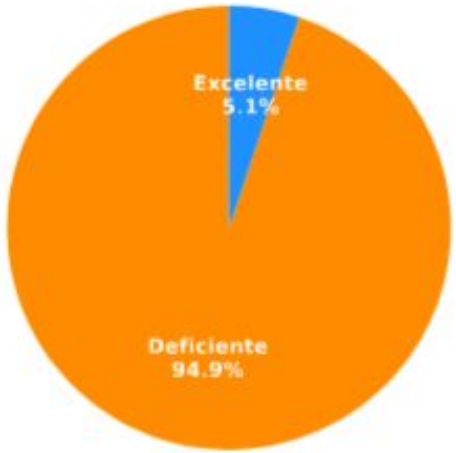
Tabla 3.1: Criterio estándar		
Concepto	**Rango (directo)**	**Rango (inverso)**
Malo	[0, 0.30)	[0.70, 1)
Bueno	[0.30, 0.60)	[0.40, 0.70)
Muy bueno	[0.60, 0.90)	[0.10, 0.40)
Excelente	[0.90, 1]	[0, 0.10)

Tabla 3.2: Criterio estándar estricto		
Concepto	**Directo**	**Inverso**
Deficiente	(0, 0.30]	[0.70, 1]
Malo	(0.30, 0.60]	[0.40, 0.70)
Aceptable	(0.60, 0.75]	[0.25, 0.40)
Bueno	(0.75, 0.90]	[0.10, 0.25)
Excelente	(0.90, 1]	[0, 0.10)

Tabla 3.3: Criterio binario	
Concepto	**Valor**
Deficiente	0
Excelente	1

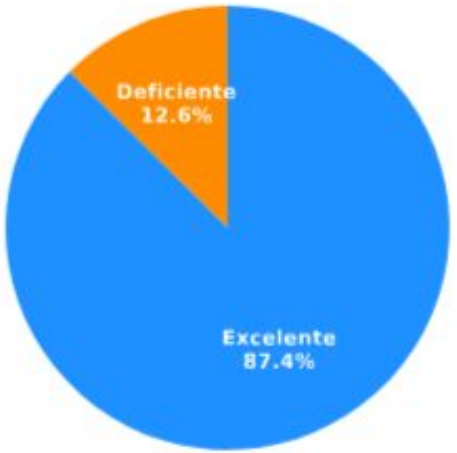
MA_consistencia_ratings
NL_BOOKS.ISBN

MA_check_edades
NL_USERS.Age



MA_consistencia_fechas
NL_BOOKS.ISBN

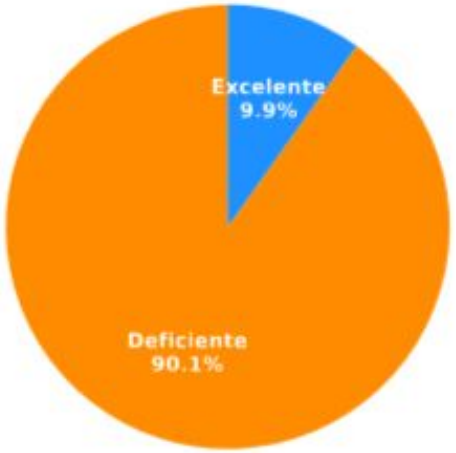
MA_missing_rating_books
NL_BOOKS.ISBN



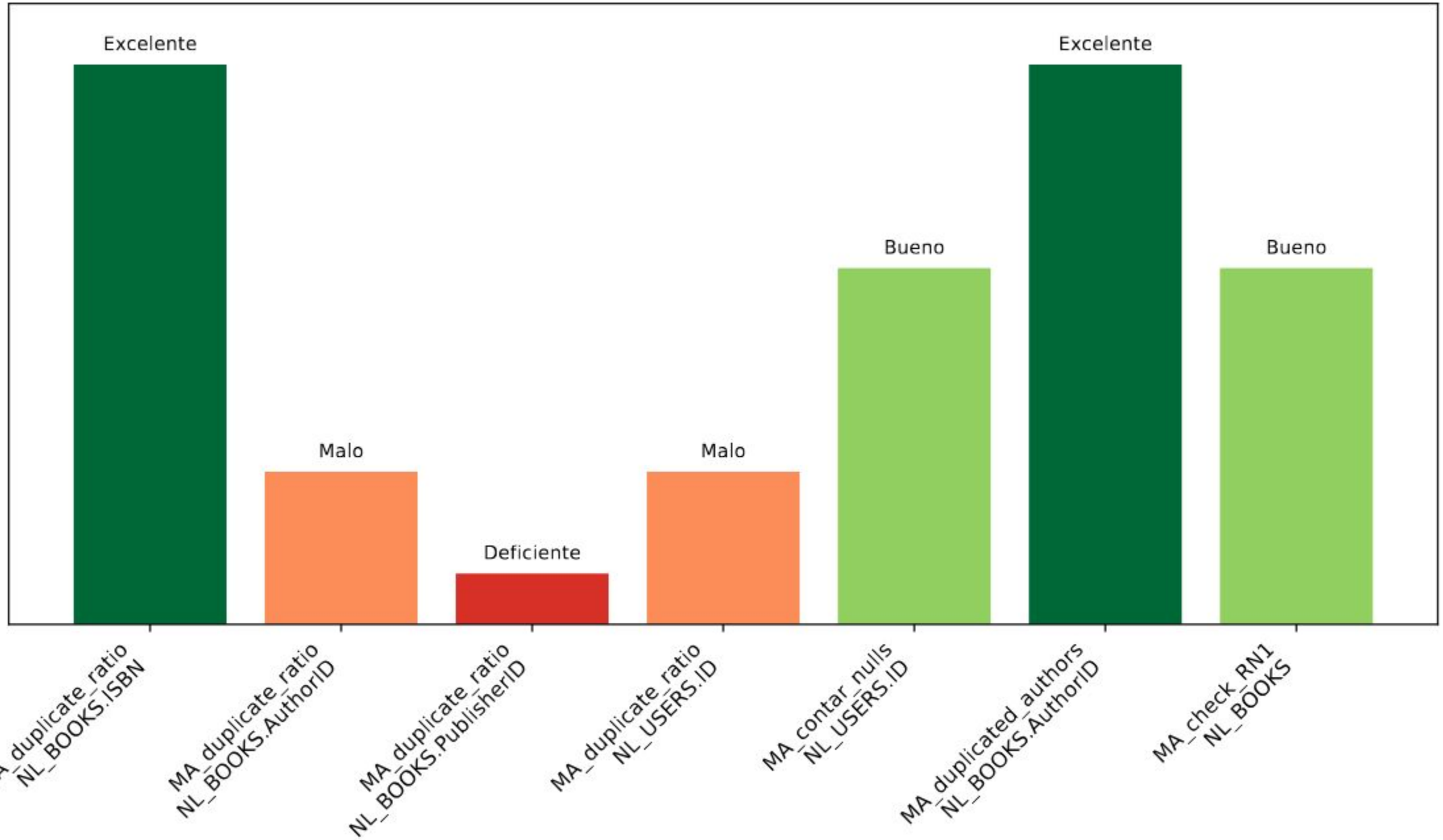
MA_check_ISBN
NL_BOOKS.ISBN

MA_date_format
NL_RATINGS.review_time

MA_check_price
NL_BOOKS.Price



MA_date_format
NL_BOOKS.PublisherDate

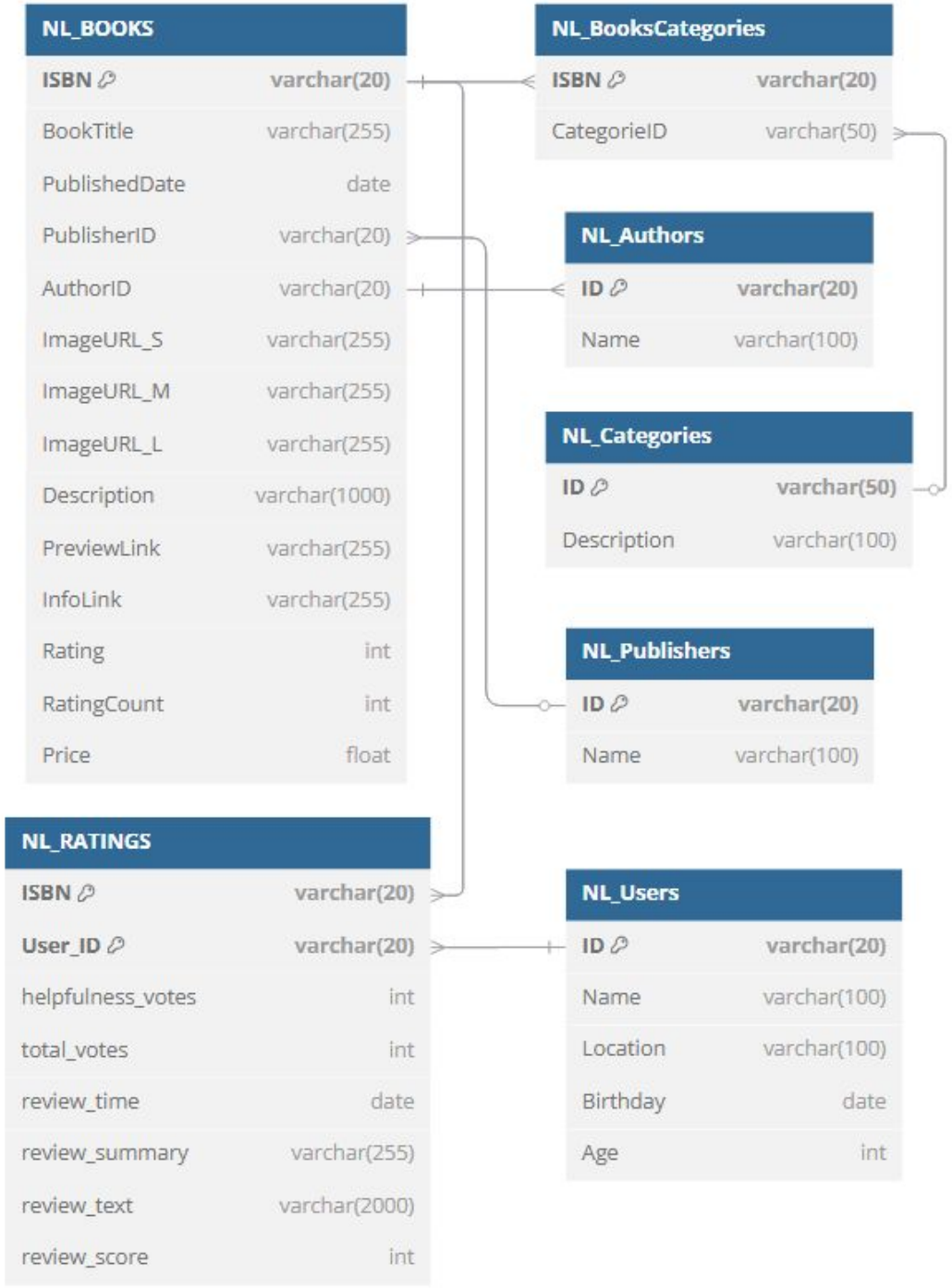

Excelente
Malo
Deficiente
Malo
Bueno
Excelente
Bueno
MA_duplicate_ratio NL_BOOKS.ISBN
MA_duplicate_ratio NL_BOOKS.AuthorID
MA_duplicate_ratio NL_BOOKS.PublisherID
MA_duplicate_ratio NL_USERS.ID
MA_contar_nulls NL_USERS.ID
MA_duplicated_authors NL_BOOKS.AuthorID
MA_check_RN1 NL_BOOKS

FASE 3
Data Quality Improvement

ANÁLISIS DE CAUSAS

- Dos bases de datos distintas
- Mal diseño de BD originales
- Errores de tipeo

PLAN DE MEJORAS



Muchas gracias

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY