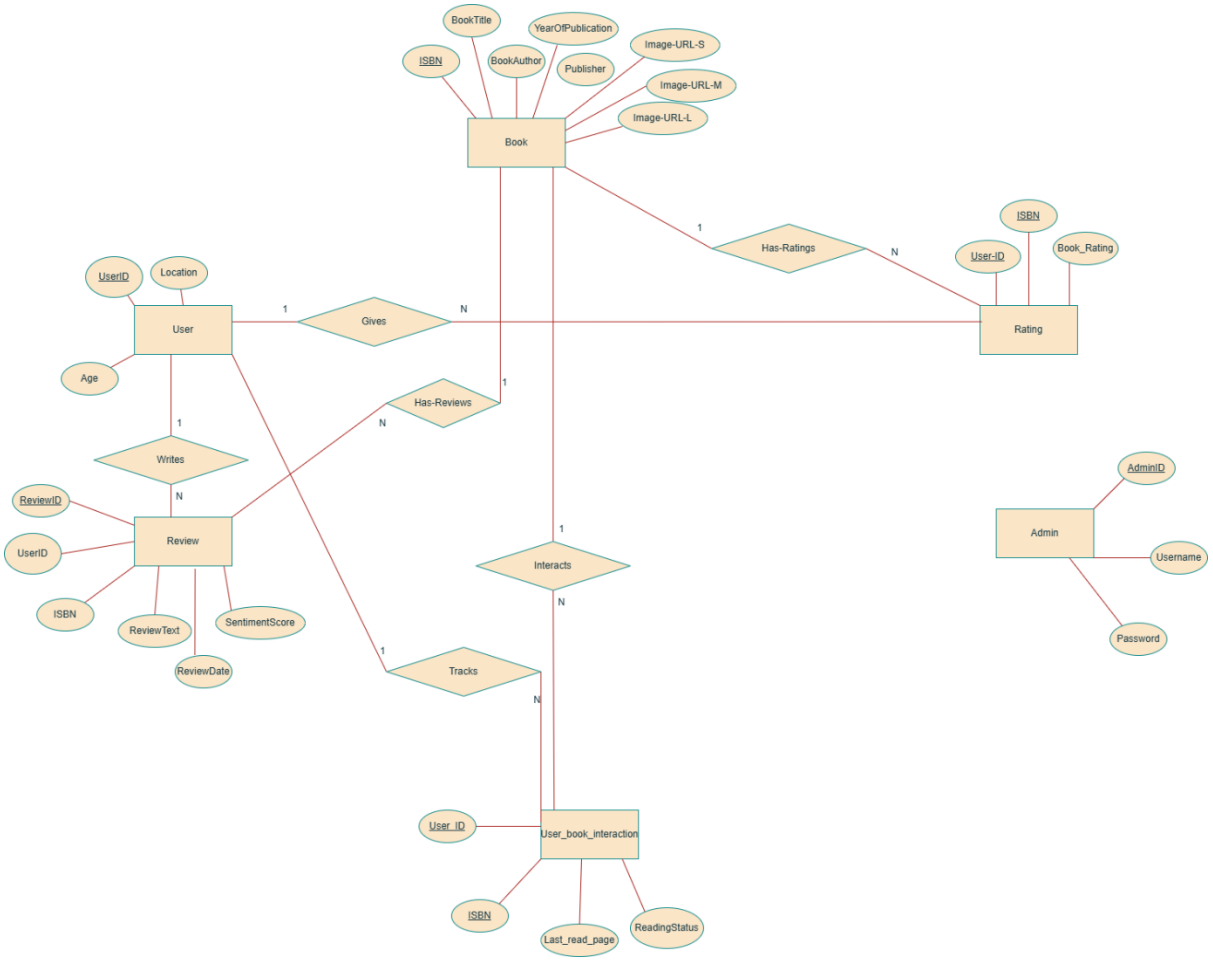
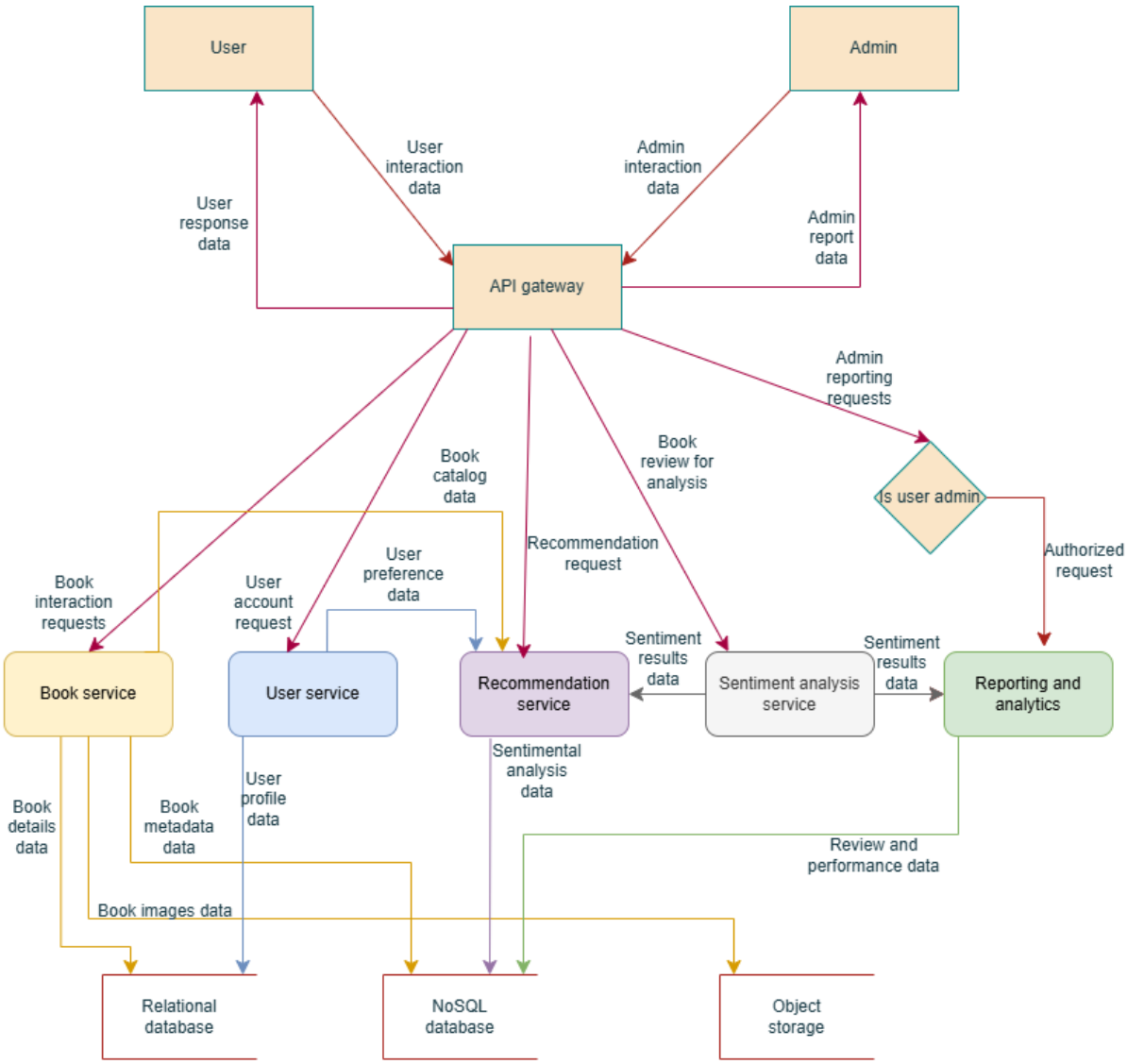
ER Diagram:


BookTitle
YearOfPublication
ISBN
BookAuthor
Publisher
Image-URL-S
Image-URL-M
Image-URL-L
Book
1
Has-Ratings
N
ISBN
User-ID
Book_Rating
Rating
UserID
Location
User
1
Gives
N
Age
1
Has-Reviews
N
1
Writes
N
ReviewID
UserID
Review
ISBN
ReviewText
SentimentScore
ReviewDate
1
Interacts
N
AdminID
Admin
Username
Password
1
Tracks
N
User_ID
User_book_interaction
ISBN
Last_read_page
ReadingStatus

Dataflow diagram



Dataset analysis

1. **Books Dataset**
    - **ISBN**: Unique identifier for books.
    - **Book-Title**: Title of the book.
    - **Book-Author**: Name of the author.
    - **Year-Of-Publication**: Year the book was published (important for identifying outdated/invalid entries).
    - **Publisher**: Name of the publisher.

2. **Users Dataset**
    - **User-ID**: Unique user identifier.
    - **Location**: Location of the user.
    - **Age**: Age of the user (used for demographic analysis).
3. **Ratings Dataset**
    - **User-ID**: ID of the user who gave the rating.
    - **ISBN**: ID of the book being rated.
    - **Book-Rating**: Rating value from the user (0–10 scale).

Relationships Between Data

- **Users ↔ Ratings**: Users give ratings to books.
- **Books ↔ Ratings**: Each book receives multiple ratings from users.
- ISBN is the common key that connects Books and Ratings datasets.
- User-ID is the common key that connects Users and Ratings datasets.

Data Issues Identified

1. **Missing Values**
    - Book-Author and Publisher in the Books dataset had missing values, filled with "Unknown".
    - Age in Users had missing or unrealistic values, filled with median.
2. **Outliers Detected**
    a. Year-Of-Publication in Books:
        - 4,465 outliers detected (e.g., 0, 9999, very old years).
        - Replaced with median year.
    b. Age in Users:
        - 59,040 outliers found (below 5 or above 100).
        - Replaced with median age.
    c. Book-Rating in Ratings:
        - 0 outliers found.
        - No action was needed.
3. **Duplicate Records**
    - Duplicate User-IDs were found.
4. **Inconsistencies**
    - Some years were strings or invalid.

- Rating values outside the expected range (e.g., >10) were considered inconsistent and dropped.