

Comprehensive Database Design Documentation for a News Summarizer Application

This documentation is designed to guide industrial training participants through the process of designing and implementing a database for a news summarizer application. The design adheres to industry best practices, ensuring scalability, maintainability, and performance.

Database Design Considerations and Best Practices

1. **Normalization:** The schema is designed to be in Third Normal Form (3NF), which helps in minimizing redundancy and maximizing data integrity:
 - **Primary Key:** Each table has a primary key that uniquely identifies each record.
 - **No Transitive Dependency:** Non-key attributes depend only on the primary key.
2. **Foreign Keys:** Used to establish relationships between tables, maintaining referential integrity and ensuring data consistency.
3. **Data Types and Constraints:** Appropriate data types (e.g., `varchar`, `int`, `datetime`) optimize storage and improve query performance. Constraints ensure adherence to data rules, enhancing quality.
4. **Scalability and Performance:** The relational table structure supports easy indexing and partitioning, facilitating efficient data retrieval as the dataset grows.
5. **Security Considerations:** Sensitive information is separated into different tables and referenced by IDs, supporting better security practices.

Below is a step-by-step guide to creating a database schema considering various entities involved and their relationships

Entities and Their Attributes

1. **News:** Represents individual news articles.
 - **Attributes:** `id`, `category_id`, `author_id`, `editor_id`, `datetime`, `title`, `body`, `link`
2. **Category:** Represents the category of the news (e.g., Sports, Politics).
 - **Attributes:** `id`, `name`
3. **Author:** Represents the author of the news article.
 - **Attributes:** `id`, `name`, `email`
4. **Editor:** Represents the editor who edited the news article.
 - **Attributes:** `id`, `name`, `email`
5. **Image:** Represents images associated with the news article.
 - **Attributes:** `id`, `news_id`, `image_url`

6. **Summary:** Represents the summarized version of the news article, generated by a Language Learning Model (LLM).

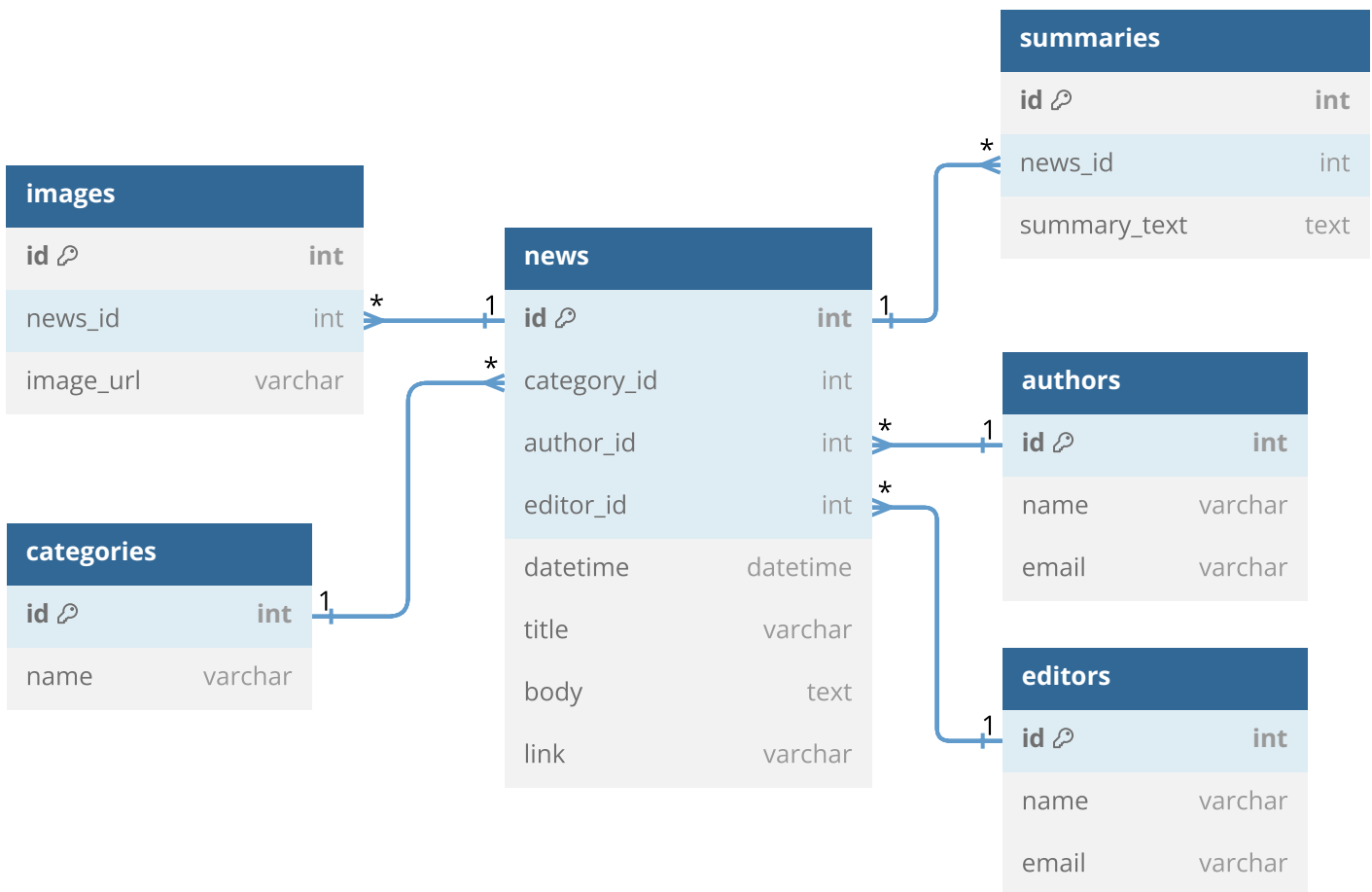
- **Attributes:** id, news_id, summary_text

Relationships

- A **News** article belongs to one **Category**, one **Author**, and one **Editor**.
- A **News** article can have multiple **Images**.
- Each **News** article has one **Summary**.

DB Diagram

👉 [News Summarizer DB Diagram - Click here](#) 👉



Explanation

- **Primary Keys (pk):** Uniquely identifies each record in a table.
- **Foreign Keys (ref):** Establishes a link between two tables.
- **Attributes Types:** varchar for strings, int for integers, datetime for date and time, text for longer text fields.
- **Increment:** Automatically increments the primary key value for new records.

Implementation Strategy

1. **Create the Tables:** Begin by creating the tables in your database management system (DBMS) according to the schema provided.
2. **Populate the Tables:** Insert initial data into `categories` , `authors` , and `editors` first, as these are referenced by the `news` table.
3. **Maintain Data Integrity:** Use transactions to maintain integrity, especially when inserting or updating data across multiple tables.
4. **Indexing:** Apply indexes on frequently queried columns (e.g., `category_id` , `author_id` , `editor_id` in the `news` table) to speed up search operations.
5. **Security Measures:** Implement role-based access controls in the DBMS to restrict who can view or modify certain data.

Synthetic Data Example

categories:

| id | name | description |
|----|----------|-------------------|
| 1 | Politics | Political news |
| 2 | Sports | Sports activities |

authors:

| id | name | email |
|----|------------|-----------------------|
| 1 | John Doe | johndoe@example.com |
| 2 | Jane Smith | janesmith@example.com |

editors:

| id | name | email |
|----|------------|--------------------|
| 1 | Alice Wong | alicew@example.com |
| 2 | Bob Lee | boblee@example.com |

news:

| id | category_id | author_id | editor_id | datetime | title | body |
|----|-------------|-----------|-----------|---------------------|---------------|------|
| 1 | 1 | 1 | 1 | 2023-01-01 10:00:00 | Election 2023 | Deta |
| 2 | 2 | 2 | 2 | 2023-01-02 15:00:00 | Soccer Match | Socc |

images:

| id | news_id | image_url |
|----|---------|-----------------------------|
| 1 | 1 | http://example.com/img1.jpg |
| 2 | 2 | http://example.com/img2.jpg |

summaries:

| id | news_id | summary_text |
|----|---------|--------------------------------|
| 1 | 1 | Short summary of Election 2023 |
| 2 | 2 | Summary of Soccer match |

This comprehensive approach ensures that participants not only understand how to design and implement a database but also appreciate the importance of best practices in database architecture.