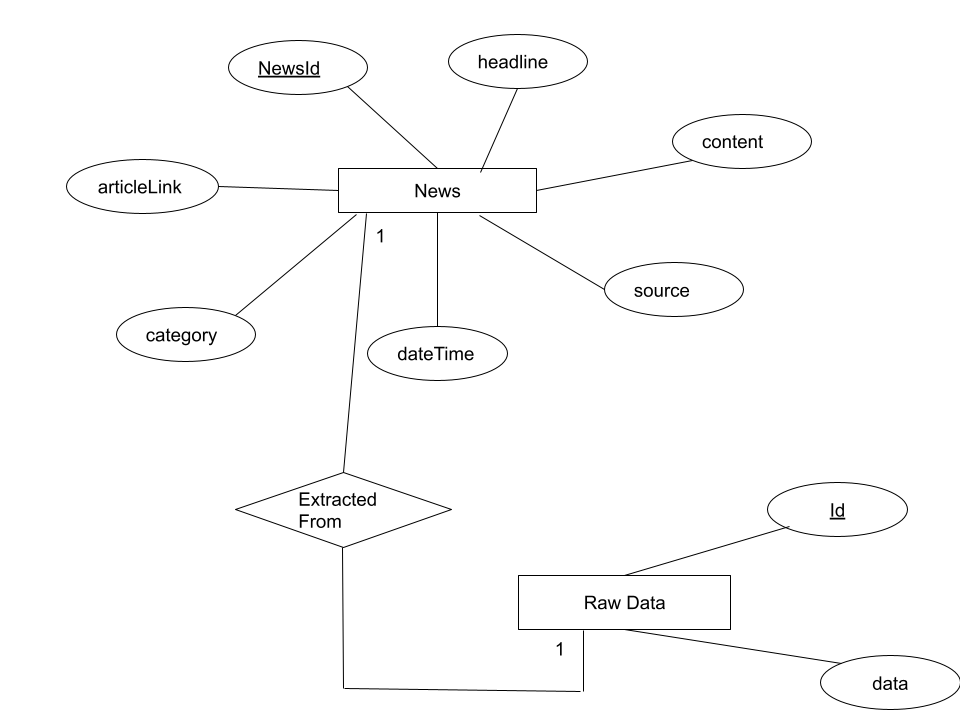
**Database Design and Implementation Phase 1**

**Project: Web Scraper For Latest News**

**Entity Relationship Diagram:**



**Data Dictionary:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Field Name** | **Data type** | **Constraint** | **Description** |
|  |  |  |  |
| **For News Table** |  |  |  |
| NewsID | Number | Primary key | Identity |
| Headline | varchar | Not null | News Headline |
| Content | varchar | Not null | News Details |
| Source | varchar | Not null | Source of news |
| DateTime | DateTime | Not null | Date/Time of news |
| Category | varchar | - | News Category |
| Article\_link | varchar | Not null | Website Link from which news is extracted |
|  |  |  |  |
| **For RawData Table** |  |  |  |
| Id | Number | Primary Key | Identity |
| Data | varchar | Not null | Html content from which news is extracted |

**MongoDB**

MongoDB is a cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with optional schemas.

MongoDB database stores its data in collections. A collection holds one or more documents, which corresponds to a record or a row in a relational database table, and each document has one or more fields, which corresponds to a column in a relational database table.

**Why MongoDB?**

- Data models are flexible and scalable

- NoSQL queries are faster

- Easy to use

- Supports non-structured data as it is required to store elements of html code.

**DDL for MongoDB in Python:**

1. Create Database:

myclient = pymongo.MongoClient(*connectionString*)

mydb = myclient[*DBname*]

1. Create Collection:

mycol = mydb[*collectionName*]

1. Drop Collection:

mycol.drop()

1. Truncate Collection:

mycol.delete\_many( { } )

**DQL for MongoDB in Python:**

1. Query all document

mycol.find():

1. Find a document:

myquery = { "category": "sports" }

mydoc = mycol.find(myquery)

**DML for MongoDB in Python:**

1. Insert a document:

mydoc = { "headline": "England lead Sri Lanka by 185", "category":"sports", "content" : "", "articleLink" : "" }

id = mycol.insert\_one(mydoc)

1. Update a document:

myquery = { "category": "sports" }

newvalues = { "$set": { "category": "tech" } }

mycol.update\_one(myquery, newvalues)

1. Delete a document:

myquery = { "category": "tech" }

mycol.delete\_one(myquery)

**Implementation:**

1. Requirement: To fetch news of last 12 hrs

Query:

myquery = { "time" : { "$gt" : "datetime.datetime.now().hour - 12" } }

mydoc = list(mycol.find(myquery))

for doc in mydoc:

print(“\n”, mydoc[“headline”])

1. Requirement: To delete news older than 36 hrs

Query:

myquery = { "time" : { "$lt" : "datetime.datetime.now().hour - 36" }}

mycol.delete\_many(myquery)