● News Feed: Admin should be able to map country, categories, publications and RSS feed  
● Crawl: The system should be able to pull the latest articles from an RSS feed or sitemap of a publication  
● Support for deduplication: The system should not crawl the articles which are already crawled  
● Parser: The system should be able to extract title, body and media links from an HTML page  
● News Update: The user should be able to see news from subscribed publications only  
● Search: The user should be able to search an article based on keywords  
● User Preference: The user should be able to subscribe to the publications and categories of interest

**Non Functional requirements:**

Fast and responsive news retrieval and summarization.

Easily scalable to handle a growing user base.

Reliable news sources with minimal downtime.

End-to-end encryption for user data.

Secure APIs and authentication mechanisms.

Entities:

User:

UserID, Username, Password(cryptographic hash functions are MD5 and SHA1),email,lastLogin

Preferences.

PreferenceId, userId, news\_channel\_name (aajTak, zeeNews)

News\_Source:

channel\_name, source\_id, URL.

News\_Article:

Attributes: article\_id, Title, Content, channel, creation\_date, media\_Id

Media:

media\_id,path,creation\_date,description

Relationships

user, preference (one to many)

Models:

enum NewChannelName {

    Aajtak, ZeeNews

}

APIS

get\_news\_feed(user\_id, user\_screen\_size, count, get\_from)

user\_id: Unique user id of the person who is fetching the timeline.

user\_screen\_size: Attribute which defines the screen size which varies depending on the device on which newsfeed is being shown for example- mobile, tablet, laptop, desktop screen, etc. Thus, the number of posts to be fetched at once and the resolution with which it has to be fetched is determined based on screen size.

count: Number of posts to be fetched at once.

get\_from: The id from and after which the next feed has to be obtained. For example: If the first 20 posts are fetched, for the next get\_timeline() request, get\_from will be 21, and if we are fetching 20 posts each time our count will be 20.

Media content like photos and videos will be obtained from object storage, and the metadata like user, articles,preferences, etc will be obtained from metadata databases.

