Data Analysis and Normalization

Steps For sanitization, normalizing and indexing the data

Step 1: Analyzing initial dataset:

Initial Data Set

content_id	c_type	title	headline	description	slug	description	state	duration	video_series	author_1	author_2	tag_1	tag_2	tag_3	thumbnail_1_URL	thumbnail_1_size	thumbnail_1_width	thumbnail_1_height	thumbnail_2_URL	thumbnail_2_size	thumbnail_2_width	thumbnail_2_height	thumbnail_3_URL	thumbnail_3_size	thumbnail_3_width	thumbnail_3_height

Problems with given data sets:

Create anomaly: Creating more than 3 tags for content or adding content with more 2 authors

Update anomaly: If we have to update size of “Compact” Thumbnail, we will have to update entire database with same value.

Step2: Sanitizing to make data atomic and unique

- Atomic data and Removing duplicate columns

content_id	c_type	title	headline	description	slug	description	state	duration	video_series	author_1	author_2	tag_1	tag_2	tag_3	thumbnail_1_URL	thumbnail_1_size	thumbnail_1_width	thumbnail_1_height	thumbnail_2_URL	thumbnail_2_size	thumbnail_2_width	thumbnail_2_height	thumbnail_3_URL	thumbnail_3_size	thumbnail_3_width	thumbnail_3_height

- Unique Keys i.e. removing duplicate rows

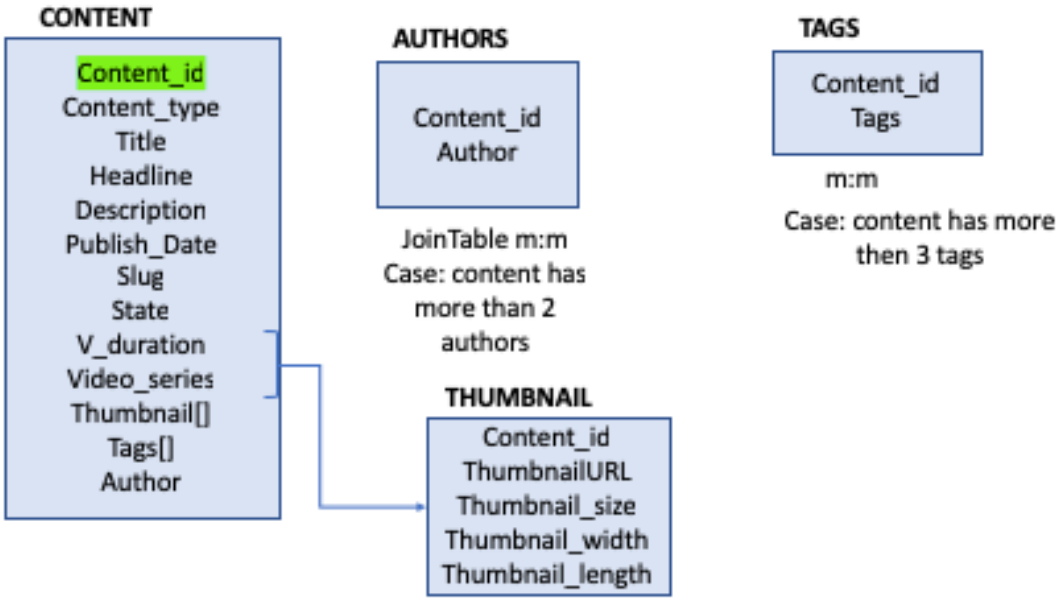
120	f2847dde564526b9eac3a86fd64c18d4	video	M&M's Super Bowl 53 Commercial: "Lock Game" ft. Christina Applegate	Christina Applegate finds herself in a test of wills while shopping.	mms-super-bowl-53-commercial-lock-game-ft-christina-applegate	published	15	none			trailer	super-bowl-xlii	episodei	https://assets1.ignimgs.com/thumbs/2019/01/28/3-f2847dde564526b9eac3a86fd64c18d4-1548715824/frame_0000_compact.jpg	compact	306	172	https://assets1.ignimgs.com/thumbs/2019/01/28/3-f2847dde564526b9eac3a86fd64c18d4-1548715824/frame_0000_medium.jpg	medium	466	262	https://assets1.ignimgs.com/thumbs/2019/01/28/3-f2847dde564526b9eac3a86fd64c18d4-1548715824/frame_0000_large.jpg	large	626	352

For example given row is repeated multiple time

Step 3: Finding dependencies and reducing redundancy

Solution1: Creating separate tables for videos, articles, Tags and thumbnails

Identifying independent sets: Non Key attributes functionally dependent on primary key



Project Structure and Implementation;

Design Pattern: DAO Factory Model

Project implemented using **Data Access Object. DAO** is an object that provides an abstract interface to database. By mapping applications calls to the persistence layer, the DAO provides details without exposing details of the database.

Motivation of DAO pattern: Changing persistence mechanism, service layer doesn't even know where the data comes from, all changes (example change of database mysql or mongo) are done in the DAO layer only.

Project Structure:

- **Consumer:** Contains client that would connect with services. It is starting point for application. Two clients are available:
  - **QueryApp:** consumer for querying database provides services query by id and query by tags
  - **WriterApp:** consumer will read the input file configured in conf.yml and write data into database.
- **Dao:** Contains DAO and DAO Implementations for Author, Content, Tag and Thumbnail. Also contains DAO factory to return objects of respective DAO
- **Exception:** contains code for exception handling
- **Mappings:** Contains configuration files to connect to database, setup up log files, commit size and yml files.
- **Model:** Contains models for mapping databases tables: Author, Content, ContentType, Tag and Thumbnails. Each model implements serializable to
- **Service:** Contains code for services provided by app
- **Util:** Contains two
  - **File utility**: contains functions for reading csv file and other file related functions
  - **SQL utility:** contains SQL queries for insert, updates, select etc.

How to build and run project:

"mvn install" will create backendApi-1.0-SNAPSHOT.jar which have execulatble code.

- To Run Java consumer to create data set:

```
Java -cp backendApi-1.0-SNAPSHOT.jar com.codefoo.consumer.WriterApp
```

- To Run Query consumer:

```
Java -cp backendApi-1.0-SNAPSHOT.jar com.codefoo.consumer. QueryApp
```

Note: Please change conf.yml based on your environment settings**.**

Future work:

- Services to update content, update tag, select by author etc can be added.