ETL Project Proposal

Group members:

Elmaddin Karimov
Batbileg Enkhbat
Mukta Jathar

Data Source:

This dataset is available on Kaggle.com and is a daily record of the top trending YouTube videos.

https://www.kaggle.com/datasnaek/youtube-new#CA_category_id.json

This dataset includes several months of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), with up to 200 listed trending videos per day.

Now includes data from RU, MX, KR, JP and IN regions (Russia, Mexico, South Korea, Japan and India respectively) over the same time period.

Each region's data is in a separate csv file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.

The data also includes a `category_id` field, which varies between regions. To retrieve the categories for a specific video, find it in the associated `JSON`. One such file is included for each of the regions in the dataset.

Proposed Transformation:

- Combine data of different regions (different csv) into one single table, include only the required regions.
- Clean-up the table to include the required columns.
- Use the associated JSON to map the category for each region into the combined table.
- Any other data clean-up and preparation as required.

Proposed Load:

MongoDb to be used to load the extracted and transformed data. Since the dataset would be derived by stitching together data for different regions, in case a need arises in future to accommodate different types/structure of data for different regions, MongoDb will allow that.