Kimball Lifecycle Proposal Template

CIS 9440 - Data Warehousing for Analytics

Final Project Milestone 3

Group Number - 12

Student(s) – Chau Hoang and Tanay Mukherjee

This Proposal is the beginning of your semester-long Final Project. The goal of the project is to develop a working Data Warehouse using a commercial database management system. Your project will use data from at least 2 sources, dimensionally model the data inside your Data Warehouse, and connect to a Business Intelligence application to produce valuable, actionable insights.

For motivation on project ideas, search for public datasets that interests you and your group. Then, think about how these datasets (maybe combined with other datasets) could address a problem or opportunity. Below are just a few (of many) public data sources:

1. Kaggle
2. NYC Open Data
3. Opendata.gov
4. Gapminder
5. Zillow
6. NOAA Climate Data
7. Google's public datasets

Data Warehouse Project Title:

YouTube videos statistical analysis for 10 different countries and comparison of the free content platforms against paid platforms (OTT) like Netflix.

Motivation for Project idea:

We want to analyze YouTube trending videos statistics in 10 different countries to learn more about different trends and topics that people from different parts of the world are interested in at the same time frame. Now do comparison with Netflix data by countries, category and time period.

Description of the issues or opportunities the project will address:

The dataset gives us a lot of avenues to analyze the videos people watch across the globe.

1. What category of videos really trend on a daily basis? Is there a connection between multiple Geos?

2. Does the number of views have a relationship with total interaction (like, dislikes or comments)?

3. What kind of videos have a high frequency of restriction enabled on likes or comments?

4. What are the top topics for viewership - entertainment, politics, sports, etc?

5. An opportunity to keep the date format consistent across all countries for which we analyse.

6. Next, we redo these above analysis for Netflix data and compare it with YouTube result by joining the data to see if we can manage to get some nice insights about people's choice on free video entertainment platform like YouTube v/s OTT platforms like Netflix.

Business Justification:

High-level Business Initiative:

To see any correlation between top trending videos in countries with high YouTube viewership and what are the preferences on Netflix for those countries.

BI Sponsors and Stakeholders (who will own this project?)

Analytics team of a consulting firm who are trying to do some market research on visual content across various countries and then do profiling based on category of the content.

What's the Business Value?

Can this exercise be used to extend a possible collaboration with Ad agencies to figure out what kind of videos might bring more traction to their Product Ads for future campaign strategy on trending YouTube videos. How can OTT platforms like Netflix make benefit out of this?

How long will this take? How much will this cost?

This project will take 2 months. The estimated cost for this project is $250.

Technical Justification:

This dataset includes several months (and counting) 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.

EDIT: 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 file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.

Netflix data is a single file with data that includes show type, director, release year, genre, cast, rating, duration, country.

Which data sources do we already have for this project?

Dataset 1: Kaggle (https://www.kaggle.com/datasnaek/youtube-new)

Dataset 2: Kaggle (https://www.kaggle.com/shivamb/netflix-shows)

What new data sources do we need (if any)

The source data has a lot of info from multiple geographies (10 to be precise), and we can use it to do all sorts of combinations to find any relationship in trends or patterns of user viewership. We also added Netflix viewership data that can be used to further enrich our analysis about content preferences across geographies.

Is the data we have conformed, consistent, and current? (data quality)

Yes, the data is consistent but the category ID is unique for each country so that might need some engineering to properly categorize them. The data for all 10 countries follow the same sequence for data fields so that will be easy to do some analysis. It follows the RDBMS structure with 'video_id' acting as the primary key. As we start digging deep we can see if there is a need for further normalization but it is clearly following up to 3NF. A more detailed version of data modeling is attached at the end.

What technical skills will we need to complete this project?

Data cleaning, data transformation, data analysis.

Will we need any new types of technologies?

We will need Python, Google BigQuery and Tableau

Key Performance Indicators (KPI's) your Data Warehouse will display:

1. Number of likes per total views
2. Number of dislikes per total views
3. Number of comments per total views
4. Number of views by category
5. Top channels with most trending videos (by different granularity)
6. Above 4 KPIs further broken down by countries
7. % of restriction in user engagement by category, country or channels
8. Total number of trending videos removed by country
9. Total number of trending videos removed by category
10. Count of most common tags for each category used YouTube channel owners
11. Type of shows on Netflix (TV show or movie)
12. TV Rating
13. Duration of show/time

(We will continue to add more KPIs worth examining as we proceed with the project execution)

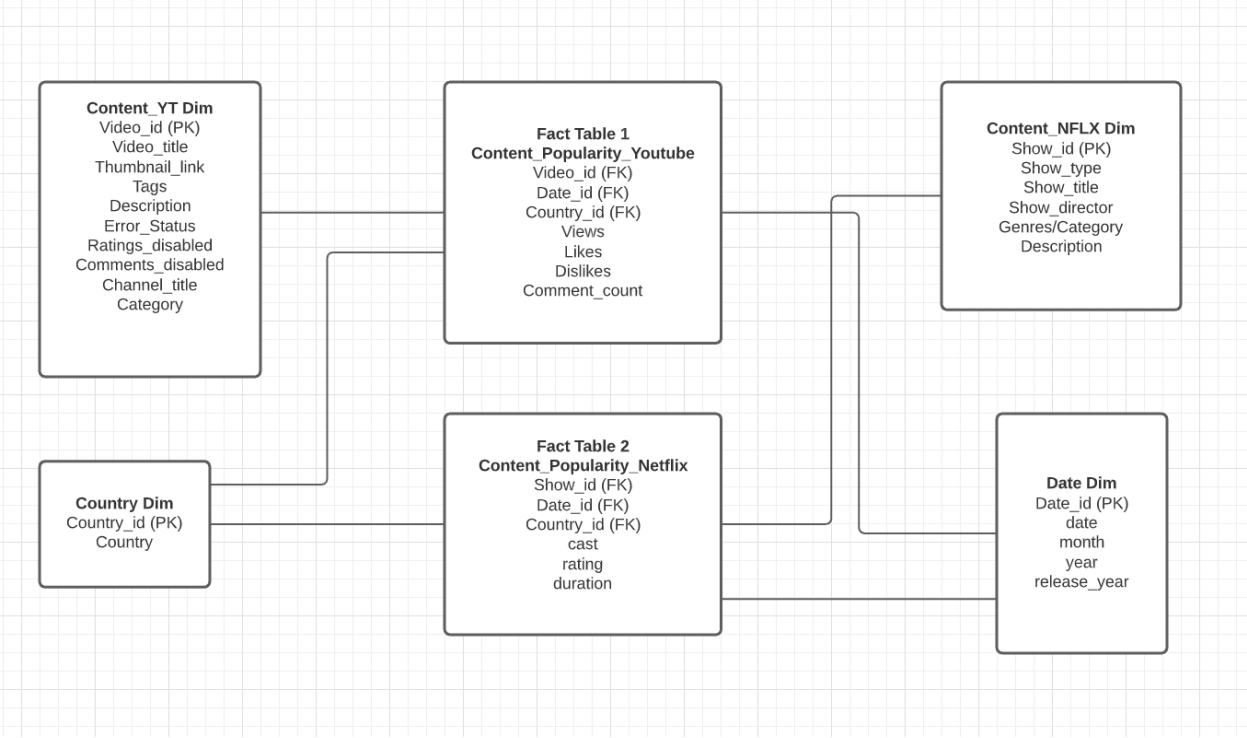
Short Description:

Audiences/ stakeholders for this project, can use these above KPIs to understand the content that are popular amongst viewers in different countries. They also get to observe the patterns (similarities/ differences) over time, common with audience watching a particular show on Netflix v/s what they consume on free content platforms like YouTube.

It also helps producers for particular shows on Netflix to partner with channel owners on YouTube for promotion of their upcoming shows.

YouTube channel owners can analyze the commonly used tags by other top content producers to understand what the best way is to maximize their engagement and interaction with audiences.

Dimensional model:



BUS matrix:

	Dimensions					
Business Processes (Facts)	Content_YT	Content_NFLX	Date	Country		
Content_Popularity_Youtube	X		X	X		
Content_Popularity_Netflix		X	X	X		

Documentation of ETL Process:

As a team we cleaned the data on python. We had two datasets in Netflix and YouTube. The YouTube data was a compilation of many files each representing a country.

We will be attaching the jupyter notebook that includes all our work. But, below are the key tasks we undertook:

1. Create a country dimension from the Netflix data. It has many rows with some having countries with multiple items per cell – indicating launch of a show simultaneously across various countries. We cleaned it and created a dimension that every country mentioned is recorded once.

2. Created a date dimension from the feature – 'date added' to find the year, month, day and days of the week for all Netflix shows and the associated date id is what we will use as the primary key for the dimension and later as a foreign key in the associated fact table. This date dimension is not final yet as we will revise it when we get to the YouTube data.

3. Post this we create a content_netflix dimesnsion to report on all kinds of dimensions necessary for later analysis.

4. Then a fact table for Netflix based on the above 3 dimensions, and also add all the measure quantifiable metrics for analysis. The details are there in the dimensional model above.

Before we start on the YouTube data, things to be highlighted are multiple files for 10 different countries and we came up with a solution to merge all of them. Also, initially we only had category id the details of which were not included in the files but in the corresponding json files, and we extracted that info.

5. Once the data was ready, we created the date dimension on the 'trending date' when the content became popular as YouTube is an open platform and then clean is as we did in point 2 above and then merge them together and remove duplicates. Now, we have the date dimension which is commonly associated with both fact tables.

6. Next is to create the content dimension for YouTube which is all the dimensions we will need for measuring any quantity in future.

7. Also, the content dimension has no country column, so we used the name of the file to get the details of country code and then used the inbuild pycountry package to find the resultant country name. Now, this will work fine with the country dimension but we check if the names of the countries are same and we found some discrepancy and fixed it. Like – Russian v/s Russian Federation and Republic of Kore v/s South Korea.

8. Finally, once all the dimensions are cleaned and ready, we create the fact table for YouTube just like how we created one for Netflix in point 4 with all the necessary dimensions as foreign keys and quantifiable features as metrics.

Running Sample queries

A) Using Python:

```
In [101]: sql = """SELECT * FROM `cis9440-project-group12.phase_3.content_nflx_dim` LIMIT 100;"""
```

```
In [106]: df = client.query(sql).to_dataframe()
          df.head()
```

Out[106]:

	show_id	video_id	country
0	s487	hWLjYJ4Bzvl	United Kingdom
1	s487	bNcj9iR956M	United Kingdom
2	s487	-KeFvjm_hcA	United Kingdom
3	s487	tQR5G3kvfNQ	United Kingdom
4	s487	VaGcPRMY5UM	United Kingdom

```
In [103]: sql = """SELECT n.show_id, y.video_id, y.country FROM `cis9440-project-group12.phase_3.nflx_fact` n
                   JOIN `cis9440-project-group12.phase_3.yt_fact` y ON
                   n.date_id = y.date_id
                   WHERE y.country IN ('United Kingdom', 'United States') LIMIT 10;"""
```

```
In [105]: df_1 = client.query(sql).to_dataframe()
          df_1.head()
```

Out[105]:

	show_id	video_id	country
0	s487	hWLjYJ4Bzvl	United Kingdom
1	s487	bNcj9iR956M	United Kingdom
2	s487	-KeFvjm_hcA	United Kingdom
3	s487	tQR5G3kvfNQ	United Kingdom
4	s487	VaGcPRMY5UM	United Kingdom

B) Using Google Big Query:

```
1  Select COUNT(*) AS Total_records from `cis9440-project-group12.phase_3.nflx_fact`;
```

Query results SAVE RESULTS EXPLORE DATA

Query complete (0.3 sec elapsed, 0 B processed)

Job information | Results | JSON | Execution details

Row	Total_records
1	7777

RUN SAVE SCHEDULE MORE

```
SELECT y.country, COUNT(*) FROM `cis9440-project-group12.phase_3.nflx_fact` n
    JOIN `cis9440-project-group12.phase_3.yt_fact` y ON
    n.date_id = y.date_id
    WHERE y.country IN (
        SELECT c.country FROM `cis9440-project-group12.phase_3.country_dim` c
    ) GROUP BY 1
```

Query results

SAVE RESULTS EXPLORE DATA

Query complete (0.0 sec elapsed, cached)

Job information | Results | JSON

Row	country	f0_
1	Canada	155623
2	Germany	155793
3	France	155070
4	United Kingdom	149122
5	India	140296
6	Japan	77621
7	South Korea	127585
8	Mexico	154873
9	Russia	155496
10	United States	155819