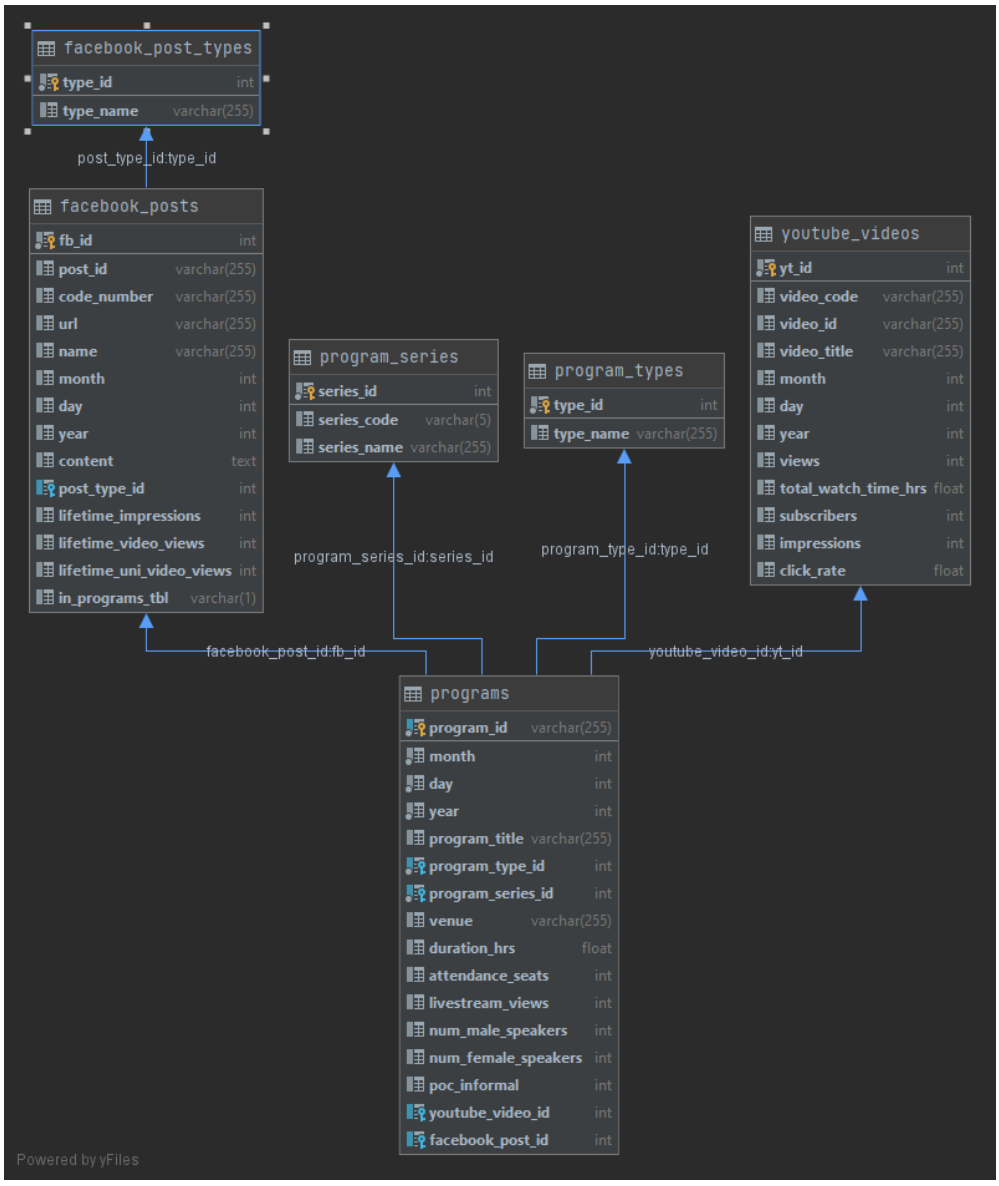
Final Project - Chicago Council on Science and Technology (C2ST) Data

Good Afternoon C2ST and happy end of the quarter!

I've been grinding away at your data, and have developed a database for your programs, facebook, and youtube data to connect each of these sources as you requested.

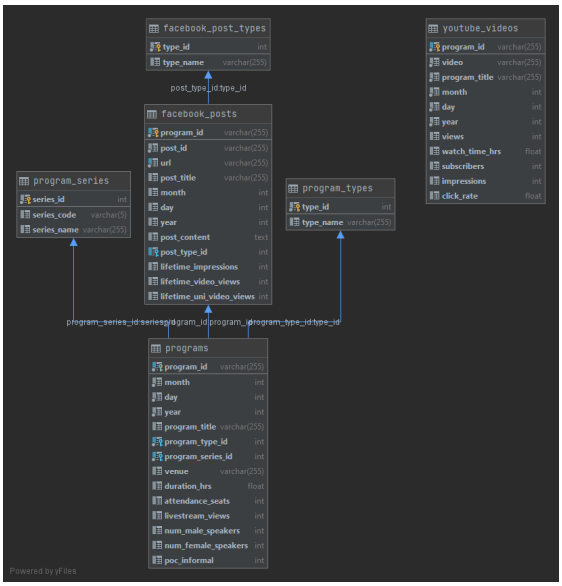
The data required some cleaning and manipulation to normalize it appropriately, so I've included my python code used to modify things prior to importing into the database. A copy of the Google Colab notebook I prepared for data manipulation can be found [on my github repository](). I've also attached the actual .ipynb file.

I then modified the comma separated value exports from my notebook, removed the indexing rows auto generated from the notebook export, and saved them as plain text files to prepare for importing into the database. Once all the data files were prepared appropriately, I got to work defining the database schema within DataGrip. I've included a visualization of the database below.



Some Important Notes:

- I experienced a lot of errors when importing my data into DataGrip, namely regarding duplicate data.
    - This seems to be a shortcoming of my understanding of the data and my data normalization, and for that, I apologize.
    - I ended up needing to discard these duplicate values, and due to time constraints, did not have time to refactor the data and restructure the database to accommodate these duplicate values.
        - Discarding duplicates eliminates valuable information about values themselves.
        - Given more time, I would like to create a new table specifically for unique values and having a second table that keeps track of these numbers on specific days, which I believe would resolve this issue.
    - I encountered a similar issue with the facebook post data.
        - I think this is moreso related to a shortcoming of facebook's analytic reporting.
        - Given more time, I would like to employ some NLP techniques to the post content to extract the most meaningful words from the post content to provide a more specific label for those with posts that do not match specific program titles.
- I also encountered some difficulty with DataGrip's autogenerated visualizations and my foriegn keys.
    - I thought I correctly identified my foreign keys, and saw DataGrip did not recognize the relationships between the programs table and the youtube_videos table.
    - It turns out that my foreign keys were, in fact, not specified correctly, which resulted in this diagram.



    - I added some new columns and modified my foreign keys accordingly.

Here are the answers to the questions you asked in your last email:

1. What are the earliest C2ST programs?

The earliest C2ST programs are “Anniversary of Sputnik” on 10-5-2007 and “Nasa Administrative Griffin” on 10-29-2007.

2. Which program had the most in-person attendance?

“Ted X Windy City Contrast” on 2-23-2013 had the most in person attendance at 750 seats.

   a. Which program had the most livestream attendance (views)?

“The Earliest Child” on 10-16-2019 had the most livestream views at 346, while the Neutrino 2020 Physics Slam on 7-2-2020 had the second highest livestream views at 341.

3. What programs had the largest combined views and impressions (from facebook and youtube)? Note: Only some programs exist in all 3 data sets

On Facebook: “The Earliest Child” had the most lifetime post impressions (2928). This post also has the most video views at 788 total lifetime views.
On YouTube: “Biomechanics of Running: The Science of Movement - Steven McCaw” has the most YouTube impressions (87317) and video views (8893).
“The Array of Things” is the only program that has both a YouTube video and a Facebook post. Lifetime Facebook impressions and views are 855 and 150, respectively. Youtube impressions and views are 1047 and 69 respectively.

4. How does the top 3 live-streamed programs compare to the top 3 overall online views?

Top 3 Livestream Views:

1. “The Earliest Child” 346 views
2. “Neutrino 2020 Physics Slam” 341 views
3. “Art, Science, and the Elegant Universe” 301 views

Top 3 YouTube Views

1. “Biomechanics of Running: The Science of Movement - Steven McCaw” 8893 views
2. “Women in Science Symposium 2012 Closing Keynote with Dr. Hafidi 05/12/12” 4583 views
3. “The Science of Addiction” 3845 views

Top 3 Facebook Views

1. The Earliest Child 788 views
2. “3D Printing PPE” 587 views
3. “Does Recycling Still Matter?” 481 views

5. Does the top 3 viewed videos also have the highest watch time?

Yes and no. The video with the most views (biomechanics of running) does also have the highest watch time (1859 hours). But the video with the second most views (women in science symposium 2012) has the 5th highest total watch time (539 hours). The Science of addiction (3rd highest views) has the 4th highest total watch time (560 hours).

   a. If not, which video has the highest watch time?

The top 2 and 3 videos with the highest watch time are “Fermilab and the New Frontiers of Physics” (1583 hours) and “The Origins of Genus Homo” (915 hours)

b. What are the click through rates for each of these videos?

All of these videos have a click through rate of less than 10%, with the lowest click rate (3.9%) for Fermilab and the New Frontiers of Physics. The biomechanics of running has the third highest click through rate at 4.49%.

6. Does the Facebook post with the most impressions also have the most views? What about the other posts with the most impressions?

The Earliest Child has the most impressions and the most views. Surprisingly, post with the second highest impressions (2540), CERN has the fourth lowest views (384) on the top 5 impressions list

7. Which program series is the most popular?

The most popular program series is the Physical Science series. It has an average in person attendance rate of about 117 seats and an average livestream count of about 41 views. I’ve attached the table for average attendance and livestream views for each program series for your reference.

program_series_id	series_name	avg_attend	avg_liveview
1	Physical Science	116.7188	41.0625
2	Science and Society	107.0581	18.0000
3	Climate, Energy, and Environment	106.6279	22.8605
4	Life Science	83.7255	24.4314
5	Health and Wellness	51.6857	28.7143
6	Technology and Engineering	52.6333	39.4000