# Finding Data – Data Source

0. The data we are collecting is from three different Data Source:

1.

2.

3.

## Data Cleanup & Analysis

\*The type of transformation needed for this data (cleaning, joining, filtering, aggregation):

0. We write code in Python with Jupiter in three different ipynb for three resources:

a. RedditData.ipynb

b.TwittData.ipynb

c. Youtube.ipynb

In each of above ipynb files, we exact data and transform dataset into Pandas DataFrame which is the table we are going to use. The **final Data Frames** we produced from APIs of above three resources were all named as **new\_topics\_data\_df.**

1. Added ‘Source’ column to each of the three Data Frames named as new\_topics\_data\_df, the content for this column is respectively ‘Reddit’, ‘Twitter’, ‘Youtube’.

2. Transformed Datatype of column with Date and Time into pandas **Datetime64ns** data type, then strip of Time and keep Data in column, then convert Datetime64ns column into **String type** in order to load table into database

3. Converted **Integer** (Int32 or Int64) data type in Data Frames into **String** in order to load table into database

4. Replaced NaN content if existing with value None for all the three dataframes.

5. After all the above transform and cleaning job, the final tables are ready to be loaded to database, there names are still **new\_topics\_data\_df.**

6. We load three tables named as **new\_topics\_data\_df** into relational database MYSQL. The schema name is crypto\_db.

The three tables were loaded into crpto\_db as three tables named as below:

1. bitcoin\_reddit : save **new\_topics\_data\_df** into csv file; import csv files into database crypto\_db named as table Bitcoin\_reddit.
2. bitcoin\_twitter: use SQLAlchemy library in Python to load **new\_topics\_data\_df** into database crypto\_db named as table Bitcoin\_twitter.
3. bitcoin\_youtube: use SQLAlchemy library in Python to load **new\_topics\_data\_df** into database crypto\_db named as table Bitcoin\_youtube.

The above can be reproduced by running the three ipynb files.

=========================================================

\* \*\*E\*\*xtract: your original data sources and how the data was formatted (CSV, JSON, MySQL, etc).

\* \*\*T\*\*ransform: what data cleaning or transformation was required.

\* \*\*L\*\*oad: the final database, tables/collections, and why this was chosen.