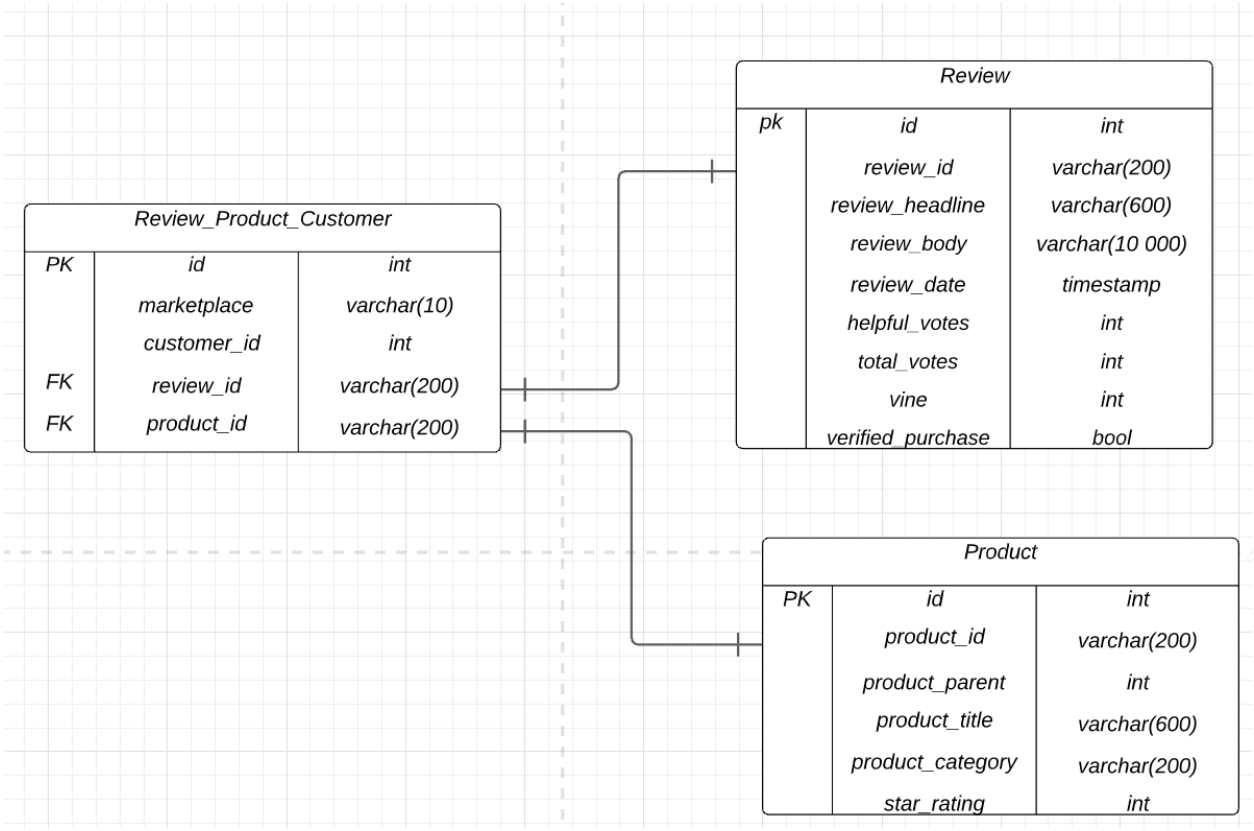
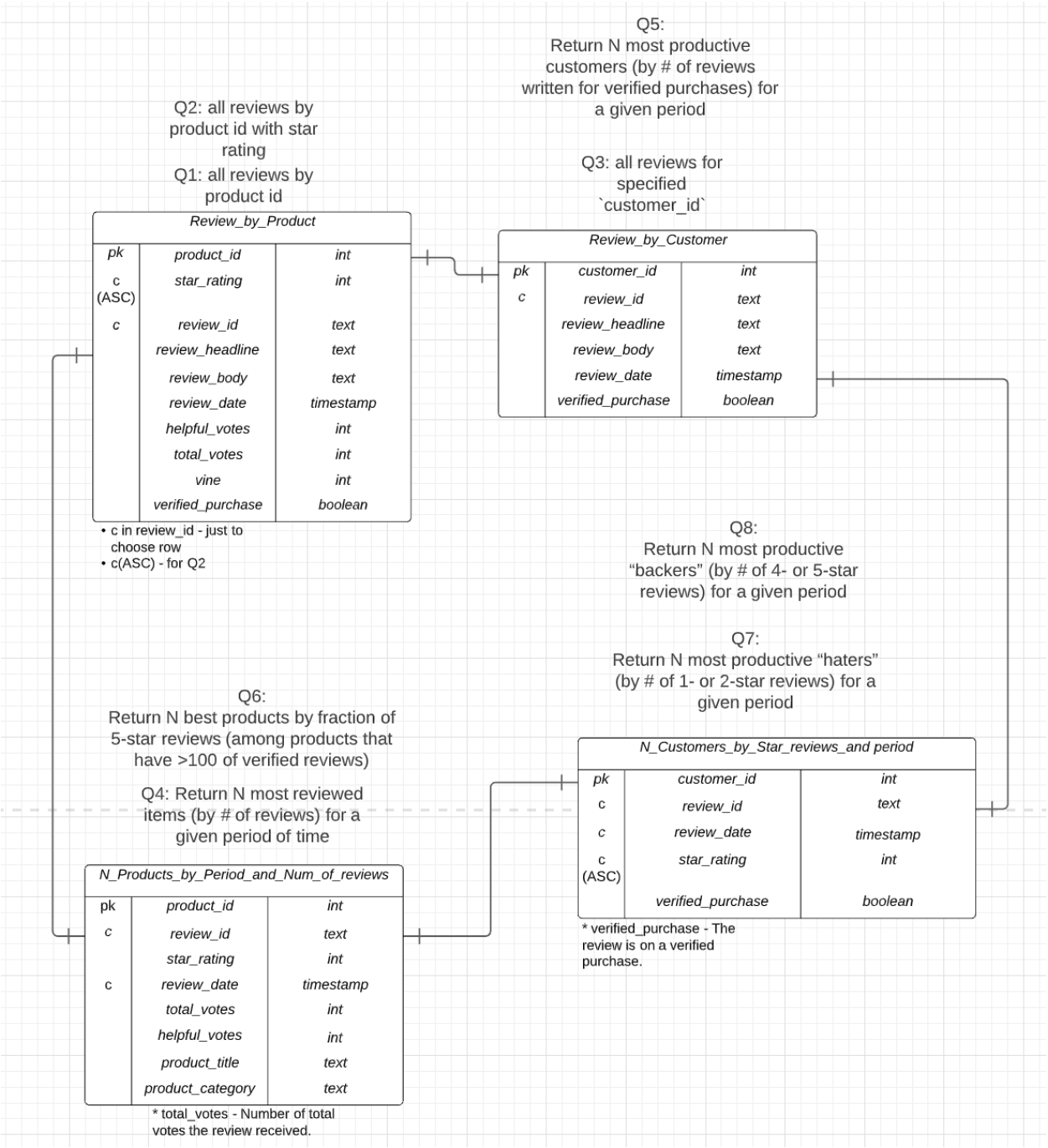
Petryshyn Sofiia
Homework 1
Git: LINK

1. Design a data model corresponding to the API requirements

RDBMS:



Cassandra Data Model:



2. Create a database RDBMS
- Separate data frame

```
# Work with each file seperately

for file in dir_files_list[0:1]:
# for file in dir_files_list[1:2]:
# for file in dir_files_list[2:3]:
# for file in dir_files_list[3:4]:
    print('Processing: ', file)
    df = pd.read_csv(os.path.join(DIR, file), sep='\t', error_bad_lines=False, quoting=csv.QUOTE_NONE)


df.head()
```

```
Processing:  amazon_reviews_us_Books_v1_00.tsv
```

	marketplace	customer_id	review_id	product_id	product_parent	product_title	product_category	star_rating	helpful_votes	total_votes	vine	verif
0	US	25933450	RJOVP071AVAJO	0439873800	84656342	There Was an Old Lady Who Swallowed a Shell!	Books	5	0	0	N	
1	US	1801372	R1ORGBETCDW3AI	1623953553	729938122	I Saw a Friend	Books	5	0	0	N	
2	US	5782091	R7TNRFQAOUTX5	142151981X	678139048	Black Lagoon, Vol. 6	Books	5	0	0	N	
3	US	32715830	R2GANXKDIFZ6OI	014241543X	712432151	If I Stay	Books	5	0	0	N	
4	US	14005703	R2NYB6C3R8LVN6	1604600527	800572372	Stars 'N Strips Forever	Books	5	2	2	N	

```
df_new = df[['marketplace', 'customer_id', 'review_id', 'product_id']]
df_new.reset_index(level=0, inplace=True)
df_new.columns = ['id', 'marketplace', 'customer_id', 'review_id', 'product_id']
df_new.to_csv('./saved/Review_Product_Customer.csv', index=False)
df_new.head()
```

	id	marketplace	customer_id	review_id	product_id
0	0	US	25933450	RJOVP071AVAJO	0439873800
1	1	US	1801372	R1ORGBETCDW3AI	1623953553
2	2	US	5782091	R7TNRFQAOUTX5	142151981X
3	3	US	32715830	R2GANXKDIFZ6OI	014241543X
4	4	US	14005703	R2NYB6C3R8LVN6	1604600527

```
df_new = df[['review_id', 'review_headline', 'review_body', 'review_date', \
             'helpful_votes', 'total_votes', 'vine', 'verified_purchase']]
df_new.reset_index(level=0, inplace=True)
df_new.columns = ['id', 'review_id', 'review_headline', 'review_body', 'review_date', \
             'helpful_votes', 'total_votes', 'vine', 'verified_purchase']
df_new.to_csv('./saved/Review.csv', index=False)
df_new.head()
```

	id	review_id	review_headline	review_body	review_date	helpful_votes	total_votes	vine	verified_purchase
0	0	RJOVP071AVAJO	Five Stars	I love it and so does my students!	2015-08-31	0	0	N	Y
1	1	R1ORGBETCDW3AI	Please buy "I Saw a Friend"! Your children wil...	My wife and I ordered 2 books and gave them as...	2015-08-31	0	0	N	Y
2	2	R7TNRFQAOUTX5	Shipped fast.	Great book just like all the others in the ser...	2015-08-31	0	0	N	Y
3	3	R2GANXKDIFZ6OI	Five Stars	So beautiful	2015-08-31	0	0	N	N
4	4	R2NYB6C3R8LVN6	Five Stars	Enjoyed the author's story and his quilts are ...	2015-08-31	2	2	N	Y

```
df_new = df[['product_id', 'product_parent','product_title','product_category','star_rating']]
df_new.reset_index(level=0, inplace=True)
df_new.columns = ['id', 'product_id', 'product_parent','product_title','product_category','star_rating']
df_new.to_csv('./saved/Product.csv', index=False)
df_new.head()
```

	id	product_id	product_parent	product_title	product_category	star_rating
0	0	0439873800	84656342	There Was an Old Lady Who Swallowed a Shell!	Books	5
1	1	1623953553	729938122	I Saw a Friend	Books	5
2	2	142151981X	678139048	Black Lagoon, Vol. 6	Books	5
3	3	014241543X	712432151	If I Stay	Books	5
4	4	1604600527	800572372	Stars 'N Strips Forever	Books	5

- Create an instance for mySQL
- Open console in GCP
    - Connect to the instance:
    - `gcloud sql connect mysql-instance-28-03-21 --user=root`
    - Use created database:
    - `use big_data_import_tables_from_csv;`
    - Create 3 tables

```
CREATE TABLE Review_Product_Customer (
id INT NOT NULL AUTO_INCREMENT,
marketplace VARCHAR(10),
customer_id INT,
review_id VARCHAR(200),
product_id VARCHAR(200),
PRIMARY KEY (id)
);

CREATE TABLE Review (
id INT NOT NULL AUTO_INCREMENT,
review_id VARCHAR(200) NOT NULL,
review_headline VARCHAR(600),
review_body VARCHAR(10000),
review_date DATE,
helpful_votes INT,
total_votes INT,
verified_purchase BOOL,
PRIMARY KEY (id)
);

CREATE TABLE Product (
id INT NOT NULL AUTO_INCREMENT,
product_id VARCHAR(200) NOT NULL,
product_parent INT,
product_title VARCHAR(600),
product_category VARCHAR(200),
star_rating INT,
PRIMARY KEY (id)
);
```

```
mysql> SHOW TABLES;
+---------------------------+
| Tables_in_rdbms_database_1 |
+---------------------------+
| Product                   |
| Review                    |
| Review_Product_Customer   |
+---------------------------+
3 rows in set (0.03 sec)
```

- Cloud Storage >
  Choose bucket >
  Upload files (upload .csv files formed for the tables)
- Import sql file to GCP to a concrete database:
  File
  Overview >
  Import >
  Choose data from the bucket >
  .CSV >
  Choose database >
  Select name for a table
  After that we have a table full of data

```
mysql> SELECT * FROM Review_Product_Customer LIMIT 1, 2;
+----+-------------+-------------+-----------------+------------+
| id | marketplace | customer_id | review_id       | product_id |
+----+-------------+-------------+-----------------+------------+
|  2 | US          |    25933450 | RJOVP071AVAJO   | 0439873800 |
|  3 | US          |    32715830 | R2GANXKDIFZ6OI  | 014241543X |
+----+-------------+-------------+-----------------+------------+
2 rows in set (0.02 sec)
```

```
mysql> SELECT * FROM Review LIMIT 1, 2;
+----+---------------+----------------+-----------------------------------+-------------+--------------+------------+-------------------+
| id | review_id     | review_headline | review_body                       | review_date | helpful_votes | total_votes | verified_purchase |
+----+---------------+----------------+-----------------------------------+-------------+--------------+------------+-------------------+
|  2 | RJOVP071AVAJO | Five Stars     | I love it and so does my students! | 2015-08-31  |             0 |          0 |                 0 |
|  3 | R2GANXKDIFZ6OI | Five Stars     | So beautiful                      | 2015-08-31  |             0 |          0 |                 0 |
+----+---------------+----------------+-----------------------------------+-------------+--------------+------------+-------------------+
2 rows in set (0.02 sec)
```

```
mysql> SELECT * FROM Product LIMIT 1, 5;
+----+------------+----------------+-----------------------------------------------------------------------+------------------+-------------+
| id | product_id | product_parent | product_title                                                         | product_category | star_rating |
+----+------------+----------------+-----------------------------------------------------------------------+------------------+-------------+
|  2 | 0439873800 |       84656342 | There Was an Old Lady Who Swallowed a Shell!                          | Books            |           5 |
|  3 | 014241543X |      712432151 | If I Stay                                                             | Books            |           5 |
|  4 | 1604600527 |      800572372 | Stars 'N Strips Forever                                               | Books            |           5 |
|  5 | 0399170863 |      559876774 | The Liar                                                              | Books            |           2 |
|  6 | 1517007240 |      299984591 | Devil in the Details (Book 2: The Monastery Murders) (Volume 2)       | Books            |           5 |
+----+------------+----------------+-----------------------------------------------------------------------+------------------+-------------+
5 rows in set (0.03 sec)
```

3. Create a Cassandra database

The same way creation of the files

```
CREATE TABLE Review_by_Product (
  product_id int,
  star_rating int,
  review_id text,
  review_headline text,
  review_body text,
  review_date timestamp,
  helpful_votes int,
  total_votes int,
  vine int,
  verified_purchase boolean,
```

```
  PRIMARY KEY (product_id)
) WITH CLUSTERING ORDER BY (star_rating ASC, review_id);


CREATE TABLE Review_by_Customer (
  customer_id int,
  review_id text,
  review_headline text,
  review_body text,
  review_date timestamp,
  verified_purchase boolean,
  PRIMARY KEY (customer_id)
) WITH CLUSTERING ORDER BY (review_id);


CREATE TABLE N_Customers_by_Star_reviews_and period (
  customer_id int,
  review_id text,
  review_date timestamp,
  star_rating int,
  verified_purchase boolean,
  PRIMARY KEY (customer_id)
) WITH CLUSTERING ORDER BY (review_id);


CREATE TABLE N_Products_by_Period_and_Num_of_reviews (
  product_id int,
  review_id text,
  star_rating int,
  review_date timestamp,
  helpful_votes int,
  total_votes int,
  product_title text,
  product_category text,
  PRIMARY KEY (product_id)
) WITH CLUSTERING ORDER BY (review_date, review_id);
```