Project 1 – databricks

You are getting some source data from yelp company:
yelp_academic_dataset_review.json
yelp_academic_dataset_checkin.json
yelp_academic_dataset_business.json
yelp_academic_dataset_tip.json
yelp_academic_dataset_user.json

Currently, your company would like some queryable datasets instead of looking into these raw yelp data. Please use ONLY databricks to achieve the “delta lake process” within SCD type II.

Requirements:

1. Upload the data into DBFS, and reading (flattern) all json data into a single databricks notebook
2. According to the DA team, you need to join all sub sources into one unified table and drop the useless column + doing data cleansing. After that, They are targeting finding the solution of the problems like:
    a. How many reviews are there for each business?
    b. How many businesses take place in each state, In each city? What kind of business do they have the most in each state, in each city ?
    c. What time do people usually write reviews?
    d. And more… ???
3. Please think of writing and using the code dynamically, which means hardening your logic into functions. One suggestion you can do is to gather raw_to_bronze, bronze_to_silver into functions.
4. Pease do pay attention with these:
    a. Think about non-reasonable negative values, null values, duplicates ?
    b. Maybe a user will write a lot of comments for a single/different business?
    c. And more…
5. Think about this project as an industry based project– which means everything treated as in work (so that please do look after your coding style and comments)
6. In the end, you should have all your bronze and silver tables in Delta format all persisted to your DBFS as well.
7. Commit your final result and source code into github.

yelp_academic_dataset_business.json

"root":{*14 items*
"business_id":string"Pns2l4eNsfO8kk83dixA6A"
"name":string"Abby Rappoport, LAC, CMQ"
"address":string"1616 Chapala St, Ste 2"
"city":string"Santa Barbara"
"state":string"CA"
"postal_code":string"93101"
"latitude":float34.4266787
"longitude":float-119.7111968
"Stars":int 5
"Review_count":int 7
"is_open":int0
"attributes":{*1 item*
"ByAppointmentOnly":string"True"}
"categories":string"Doctors, Traditional Chinese Medicine, Naturopathic/Holistic, Acupuncture, Health & Medical, Nutritionists"
"hours":NULL}

yelp_academic_dataset_checkin.json

"root":{*2 items*
"business_id":string"---kPU91CF4Lq2-WlRu9Lw"
"date":string"2020-03-13 21:10:56, 2020-06-02 22:18:06, 2020-07-24 22:42:27, 2020-10-24 21:36:13, 2020-12-09 21:23:33, 2021-01-20 17:34:57, 2021-04-30 21:02:03, 2021-05-25 21:16:54, 2021-08-06 21:08:08, 2021-10-02 15:15:42, 2021-11-11 16:23:50"}

yelp_academic_dataset_review.json

“root":{*9 items*
"review_id":string"KU_O5udG6zpxOg-VcAEodg"
"user_id":string"mh_-eMZ6K5RLWhZyISBhwA"
"business_id":string"XQfwVwDr-v0ZS3_CbbE5Xw"
"stars":int3
"useful":int0
"funny":int0
"cool":int0
"text":string"If you decide to eat here, just be aware it is going to take about 2 hours from beginning to end. We have tried it multiple times, because I want to like it! I have been to it's other locations in NJ and never had a bad experience. The food is good, but it takes a very long time to come out. The waitstaff is very young, but usually pleasant. We have just had too many experiences where we spent way too long waiting. We usually opt for another diner or restaurant on the weekends, in order to be done quicker."
"date":string"2018-07-07 22:09:11"}

yelp_academic_dataset_tip.json

"root":{*5 items*
"user_id":string"AGNUgVwnZUey3gcPCJ76iw"
"business_id":string"3uLgwr0qeCNMjKenHJwPGQ"
"text":string"Avengers time with the ladies."
"date":string"2012-05-18 02:17:21"
"compliment_count":int0}

yelp_academic_dataset_user.json

"root":{*9 items*
"user_id":string"qVc8ODYU5SZjKXVBgXdI7w"
"name":string"Walker"
"review_count":int585
"yelping_since":string"2007-01-25 16:47:26"
"useful":int7217
"funny":int1259
"cool":int5994
"elite":string"2007"
"friends":NULL}