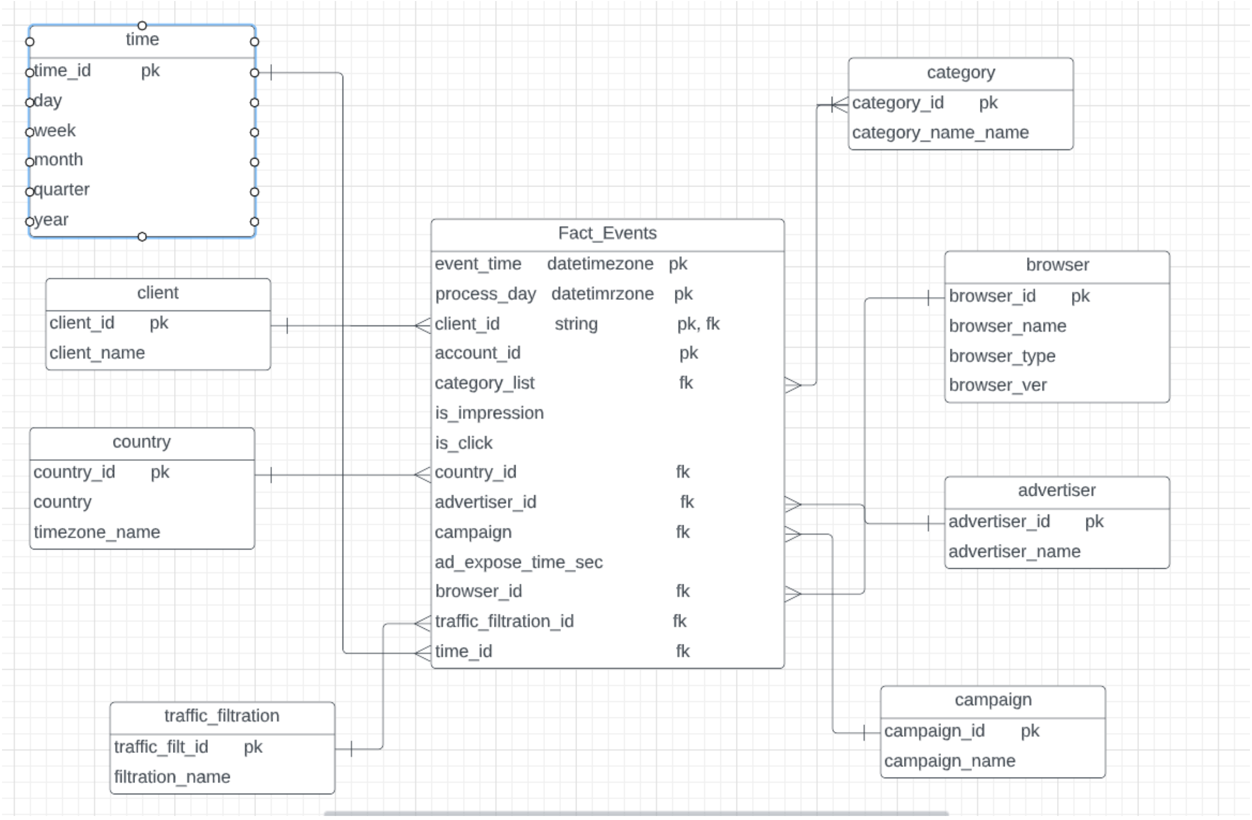
Answer to questions:

1. ERD:



Notes:

- Since advertiser and campaign can be empty it should be treated in the model by creating a value of -1 in the dimension tables for advertiser and campaign with name:NA, and add this foreign key -1 whenever needed to the fact table.
- The category_list foreign key should refer to the category dimension in order to retrieve the top 100 categories.
- The fact table is the basis for all aggregation tables needed:
    - The aggregation table can either be created using SQL and by creation of tables/views
    - Or by using a BI visualization tool (e.g., Tableau, Looker, Excel) to aggregate the data and filter it by desired fields.

2. Create MRR table, STG table and the final DWH table.
    a. Ingest all data as is into the MRR table by daily partitions
    b. For the STG part
        i. Apply the time zone conversions on the partitioned data and take care of nulls and empty values before inserting into the fact table
        ii. Check for validity of data of the dimension table (check if the foreign key for the dimension exists in the dimension table, and if not, insert it)
        iii. Check that the data types correspond to what is defined in the staging table

    c. While checking for data validity, include check regarding data correctness such as:
        i. Non missing values for event_time.
        ii. Boolean fields should be as boolean, unless decided on multiple categories (is_impression, is_click).
    d. On the accessibility level:
        i. Is the data on the DWH level accessible by the BI visualization tool?
        ii. Are all relevant data formats accessible (e.g., PDF).
        iii. make sure all privacy conditions were applied as needed (masking client and account names when necessary).

3. Dimensions are:
    a. Time
    b. Client
    c. Country
    d. Traffic_filtration
    e. Category
    f. Browser
    g. Advertiser
    h. campaign
4. When the data is being received in a delay from the source it is being indicated in the process_day field.
   So, for example, if today is July 10, 2022, and the dashboard should have updated data up to yesterday, that means that we should see in the dashboard the whole complete data for : July 9th, 2022.

   So, if today is the 10th, 2022, I should wait until all my event_times come up to 23:59, 9th July and at least one event should come in on July 10th, 2022, in order to make sure that the system is completely done with sending all events for the 9th.
   Then I can be sure the whole data for the 9th day is received and I can start the process of data creation for the 9th of July.