DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

CS-322 2022 Project Part 2

Per-Group Deliverable 3 Queries

You can name the output columns as you wish. The order of the result set columns (output schema), and query specification matter. We will explicitly specify when we mean INDICIA PUBLISHER instead of PUBLISHER. The queries you provide must execute on your database (no execution errors, exactly as specified) – otherwise, the corresponding query is graded with 0. You can use any functionality that Oracle SQL allows, and you should consult the SQL documentation for the DBMS you use – we provide hints in some queries about what you need to look for. There are 14 queries in total in this deliverable.

1. For each of the 10 publishers that *began* publishing first (the first 10 based on the year – limit the top 10 in the query even if there are ties), display in how many languages each of them published series in.
   Output schema: {Publisher Name, Series Language Count}

2. Display the publisher names that have published more than 500 series, along with the number of series. Show the result in descending order of the count of published series.
   Output schema: {Publisher Name, Series Count}

3. We are interested in the brand group names with more than 100 indicia publishers under the brand group. Display the brand group name and the corresponding count.
   Output schema: {Brand Group Name, Count}

4. Show the brand group names with the largest number of Belgian Indicia Publishers (Indicia Publisher's country at least partially matches "Belgium"). Show the Brand Group Names and the resulting largest number.
   Output schema: {Brand Group Name, Largest Number of Belgian Indicia Publishers}

5. Find Indicia Publishers that have published at least 400 single-issue series. Display their names and the count (of single-issue series), and order the result by the count descending. Remember that issues may have multiple instances with the same indicia publisher and series, and single-issue series are the ones that have exactly one such pair-value (indicia_publisher, series) occurrence in the issues.
   Output schema: {Indicia Publisher Name, Count}

6. What is the most reprinted story for an issue? Display the issue ID, the story ID with the most reprints, and the corresponding reprint count. Sort the output based on the reprint count, and display the top 5 results. Count only the immediate reprints (the stories that match the origin ID).
   Output schema: {Issue ID, Story ID, Count}

7. Which heroes are featured in stories (feature attribute) that have **all** three genres: humor, crime, and romance.
   Output schema: {Feature}

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

8. Considering the PUBLICATION_DATE column in GCD_ISSUE table – it was kept as a string due to having a variety of date formats. Write a SQL query to extract the years from the PUBLICATION_DATE column that has a 'DD/MM/YYYY' date format. Display the distinct years only once, in ascending order.
   *Hint: You will find useful the Oracle documentation on **TO_DATE** function, **extract** function to get the particular part/value from the date, as well as using the **default null on conversion error** in **to_date** function to handle formatting issues that happen since other date formats exist in the data and for some the matching will fail – therefore instead of failing to return a null value.*
   Output schema: {Extracted Year Values}

9. Related to question 8: we are interested in the counts of years of the PUBLICATION_DATE field of GCD_ISSUE table in different formats. To limit the scope of this question, write a SQL query that returns the count of years in given formats, in this order: 'DD/MM/YYYY', 'MM/DD/YYYY', 'MONTH YYYY'. Return the result in a single line, with 3 columns (one for each format), and a single row (only one value, the count, per column).
   Hint: You will find useful the Oracle documentation on TO_DATE function, as well as using the default null on conversion error in to_date function to handle the formatting errors. Note that COUNT(*) returns all the rows, including null values, while COUNT(with_specified_column) does not count null values – so you can omit the null check in the where clause in that case, if sufficient as a check. Recall that you can use subqueries to simulate multiple input tables to achieve the required output.
   Output schema: {COUNT 'DD/MM/YYYY', COUNT 'MM/DD/YYYY', COUNT 'MONTH YYYY'}

10. Finally, related to the date format and extraction, display the total number of issues per year between 1965 and 1975, including both years in the result. The year should be taken from PUBLICATION_DATE, with 'MONTH YYYY' format. Display the results by years ascending (1965 to 1975).
    Output schema: {Count Per Year, Year}

11. What are the titles of the stories that have been reprinted at least 30 times? Print the reprint count along with the title. Order the output by the reprint count descending (higher to lower).
    Output schema: {Story Title, Reprint Count}

12. We are interested in the top 10 countries based on the publisher(s) with the longest time publishing, meaning the longest duration between the year they began and ended publishing. Take into consideration the year range between **and including** 1600 for the year began and 2020 for the year ended (**both included**). Display for those countries the longest duration of publisher publishing, as well as the average duration of publishing. Order the top 10 countries by the maximum duration descending (highest first).
    Output schema: {Country Name, Max Years Publishing Per Country, Average Years Per Country}

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

13. List all Marvel heroes that appear in Marvel-DC story crossovers. Marvel and DC are considered to be (parts of) Indicia Publisher names, and the heroes are described in the story *feature* attribute. Therefore, we are interested in heroes that appear in purely Marvel stories (without DC) AND in the ones that have both Marvel and DC in the corresponding Indicia publisher name. When comparing strings, you must transform all the strings to lowercase (there is a corresponding function to use – check the documentation) and use partial string matches – so for Marvel and DC you need to find Indicia Publishers that partially match 'marvel' and 'dc' anywhere in the string. Display distinct, lowercase hero/feature strings – but do not manually deduplicate them further. Make sure that in the final result once you combine purely Marvel heroes with crossover heroes, you match their names partially as well!
Output schema: {Crossover Feature Heroes}

14. For every country that has at least 200 publishers (based on Publisher Country location, they don't need to have published any series, and the series they published can be in other countries than the Publisher Country), print the top 2 publishers by the number of series published (in any country, do not enforce that the series country is the same as publisher country).
Hint: it is highly likely that you will need to use over/partition by, as well as filtering on row_number(). Break down the query into intermediate sub-tables, and check how to use over/partition by for top-k queries.
Output schema: {Country Name, Publisher Name}