# Assignment 2

Due date: January 26, 2014, 11:55pm IST

**General Instructions**

1. You will submit 2 files: Fname1Lname1-Fname2Lname2.sql and Fname1Lname1-Fname2Lname2.pdf, corresponding to the dataset and a timings report, respectively.
2. The .sql files are run automatically using the psql command using the \i option, so please ensure that there are no syntax errors in the file. *If we are unable to run your file, you get an automatic reduction to 0 marks.*

   To understand how to run many queries at once from text file, a dummy query file example.sql is available in `www.cse.iitd.ac.in/~cs5090257/example.sql`. To run example.sql in postgres, type the following command in the terminal:

   sudo -u postgres psql dbname

   and then type

   \i /address/to/example.sql

   This command will run all the queries listed in example.sql at once.
3. The format of the file should be as follows. You can have a preamble where you may create views if you like (please note that no procedures are allowed (this has to be pure sql), and correspondingly have a cleanup section where these views are removed. One line should identify the query number (note the two hyphens before and after the query number), followed by the actual, syntactically correct SQL query. Leave a blank line after each query.

   - -PREAMBLE- -

   OPTIONAL DEFINITIONS

   - -1- -

   SQL QUERY

   - -2- -

   SQL QUERY

   - -3- -

   SQL QUERY

   - -CLEANUP- -

   CLEANUP EVERYTHING YOU CREATED HERE
4. Many of the queries below require an 'ORDER BY' clause. If you made an error in this clause, your answer will be evaluated as incorrect and zero marks will be awarded.
5. Errors with respect to equality and inequality conditions will also be evaluated as incorrect and zero marks will be awarded.
6. Assume set semantics, unless stated otherwise.
7. No changes are allowed in the i) data, ii) loading process, iii) attribute names, iv) table names
8. The .pdf file should contain a bar graph corresponding to the dataset. The graphs should report the timings for each query of the dataset (X-axis legend is the query number, Y-axis legend is the time taken).
9. **TODO** Moodle instructions

# 1 Dataset

## 1.1 Instructions

1. In this assignment you will analyze publication data i.e. DBLP, the reference citation website created and maintained by Michael Ley. The analysis has to be done in postgres. We are providing you cleaned up data and you can download it from `www.cse.iitd.ac.in/~cs5090257/unified.txt`. You can get more information about the DBLP from `http://dblp.uni-trier.de/xml/docu/dblpxml.pdf`. The cleaned up data only contains the information about the publications that are of type articles.

2. Your database should include following five tables and you should use only these tables while writing solution of the queries. You can create temporary tables while handling any SQL query but you should include SQL queries for creating and deleting these temporary tables.

    1. Paper
    **Attributes :** a) PaperId: Int, b) Title: text, c) year: Int, d) VenueId: Int
    2. Author
    **Attributes :** a) AuthorId: Int, b) name: String
    3. PaperByAuthors
    **Attributes :** a) AuthorId: Int, b) PaperId: Int
    4. Citation
    **Attributes :** a) Paper1Id: Int, b) Paper2Id: Int
    *Note :* Paper2Id is paperId of the cited paper and it is cited by Paper1Id, i.e. Paper1Id cites Paper2Id. There are few entries in the dataset which cites paper that are not available into the dataset. So you have to ignore those entries.
    5. Venue
    **Attributes :** a) VenueId: Int, b) name: varchar(50), c) type: varchar(50)
    *Note:* Here 'name' is acronym of Venue and 'type' is the type of Venue like conference/journal etc. You will be able to extract these information from key associated with each paper in the dataset.

## 1.2 Queries

1. For each type of venue, count the total number of publication presented at that type of venue. Your query should return a set of (publication-type, count) pairs. For example: (conference, 20000), (journal, 30000), ...

2. What is the average number of authors per paper? Output column: the aggregate value.

3. List the titles of all entities with more than 20 authors order by alphabetically.

4. List the names of all authors who never published as a single author order by alphabetically.

5. Find the top 20 authors with the largest number of publications order by first rank and on tie, order by alphabetically.

6. List the names of all authors who published as a single author more than 100 times order by alphabetically.

7. List the names of authors who have never published a paper in a journal order by alphabetically.

8. List the names of authors who have **only** journal publications order by alphabetically.

9. List the names of the authors who have published at least two papers in 2012 **and** at least three papers in 2013 order by alphabetically.

10. Find the top 20 authors with the largest number of publications in 'corr' journal order by rank.

11. Name the authors that have four papers or more in 'amc' journal order by alphabetically.

12. Two major journals are 'ieicet' and 'tcs' . Find all authors who published at least 10 'ieicet' papers but never published a 'tcs' paper order by alphabetically.

13. Compute the total number of publications per year in DBLP during 2004-2013. Output Column: (Year, No. of Publication).

14. How many (distinct) authors have papers with query optimization in the title (case insensitive) order by alphabetically? Output column: count.

15. Write an SQL query that lists the titles of the most cited papers order by decreasing number of citation. Tie breaking?

16. Write an SQL query that list the titles of all publications that have been cited more than 10 times order by alphabetically.

17. Write an SQL query that list the titles of all publications that are cited more than they cite by a margin of at least 10 order by alphabetically.

18. Write an SQL query that list the titles of papers that have never been cited order by alphabetically.

19. Write an SQL query that list the names of all authors who have published papers citing other papers by themselves order by alphabetically.

20. Who are the authors who published in any 'corr' journal between 2009 and 2013 but not in 'ieicet' 2009 journal? Output column: author.name order by alphabetically.

21. What is (are) the journal(s) that has (have) a strictly increasing number of papers every year from 2009 to 2013? Output column: journal.name order by alphabetically. You cannot list the years from 2009 to 2013 individually in the query. You can only use 2009 and 2013 as constants, but not anything in between. (Hint: You may create a temporary table, but make sure to drop it after the query. Points will be deducted if the temporary table is not dropped.)

22. What is(are) the journal/year(s) (e.g., corr 2001) which had the largest number of papers? Output columns: journal.name, journal.year.