DBMS 2 - CS3563: Assignment 2 Report

GROUP 31

NAME	ROLL NUMBER
SAI VARDHAN MALTHI	MA19BTECH11010
UMESH KUMAR REDDY MOGATALA	ES19BTECH11005
HEMANTH K	ES19BTECH11003
NAVEEN J	CS19BTECH11009

Deliverables:

- *parser.py*
- *modified_ERD.pdf*
- *loader.py*
- *relational_schema.pdf*
- *bash_file.sh*
- *pg_dump.sql*

Brief description about our implementation:

parser.py

The parser.py file, given the source.txt file, creates a new .csv file. This file contains details of each paper's field(Title, Authors, year, venue…).

Key changes in ERD:

- The main author relation is dropped and instead, we have added another attribute in the research paper entity to identify the main author of a research paper.
- The author entity in modified ERD has author_id as the PK.

---

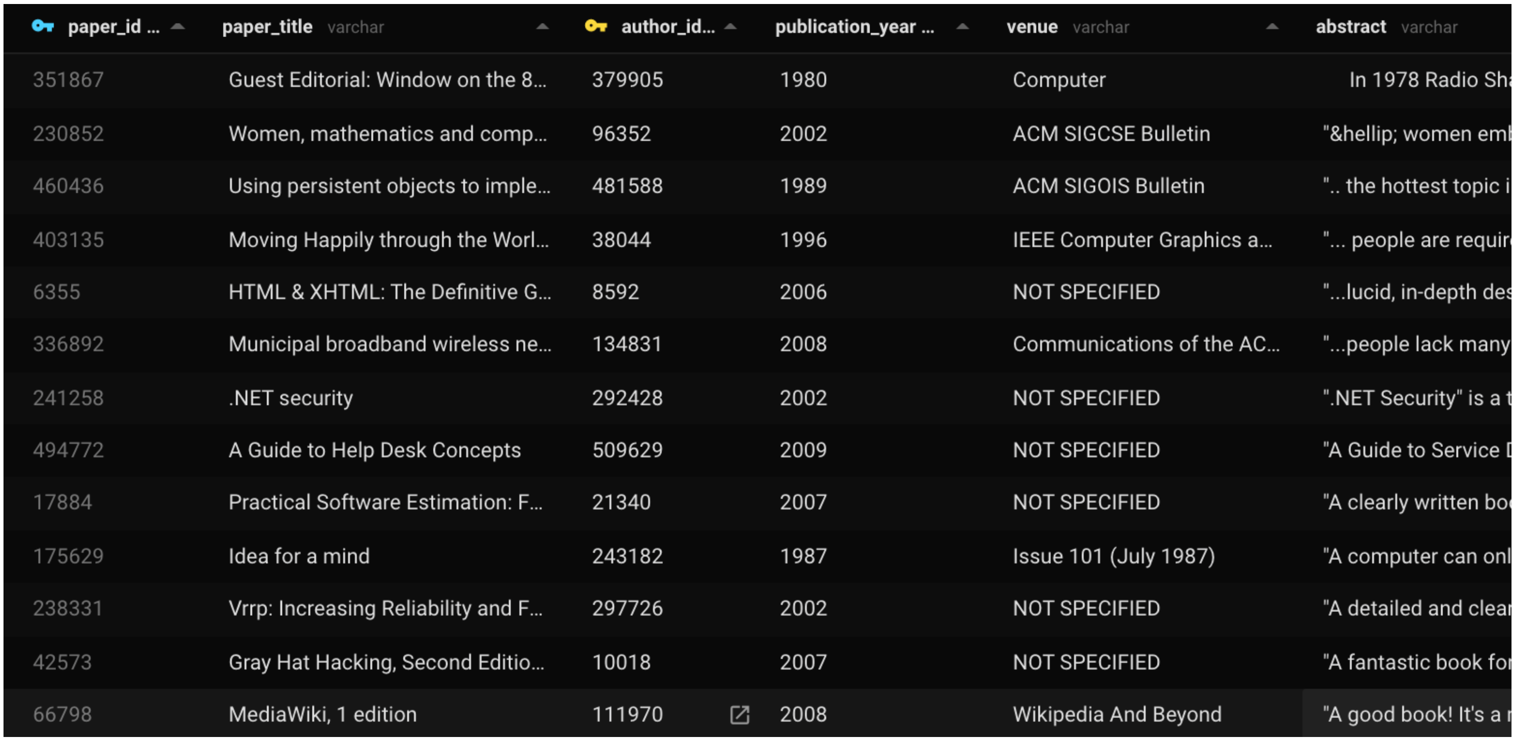
loader.py file

- Before importing the data we created all 4 raw tables corresponding to the modified ER Diagram using the psycopg2 module.
- **Data Cleaning:**
    1. We've taken care of empty fields by filling them with pre-decided text("NOT SPECIFIED").
    2. Repetitions in author names for a single research paper are also handled.
    3. We've also imposed the "NO SELF CITATION" constraint.
- A quick overview of the attributes and constraints of each raw table is provided in the latter part.
- We install required python libraries like ordered-set, psycopg2 by running the .sh script file.

[SNAPSHOTS] **Raw tables after populating data:**

1. **ResearchPaper -** *contains information about all the research papers.*

Primary key: paper_id
Foreign key: author_id

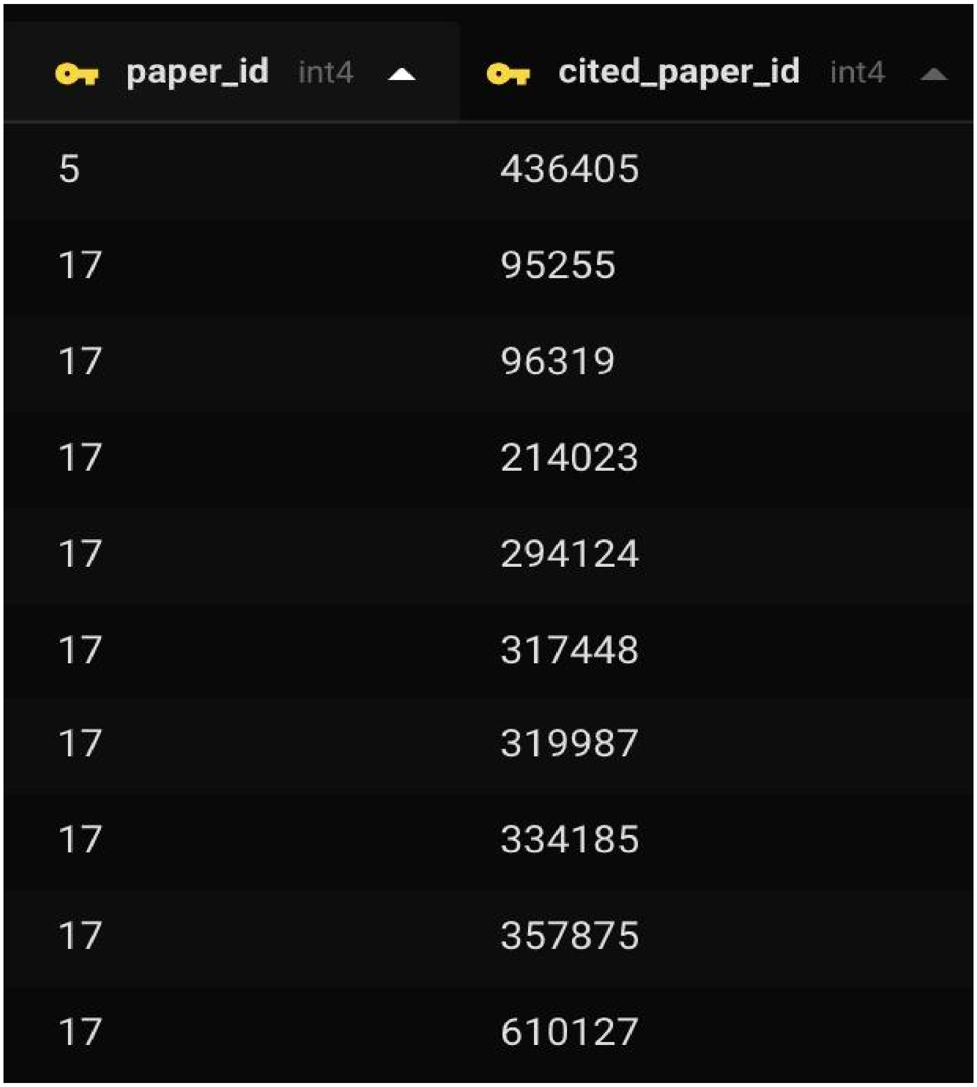


paper_id ...	paper_title varchar	author_id...	publication_year ...	venue varchar	abstract varchar
351867	Guest Editorial: Window on the 8...	379905	1980	Computer	In 1978 Radio Sh
230852	Women, mathematics and comp...	96352	2002	ACM SIGCSE Bulletin	"&hellip; women em
460436	Using persistent objects to imple...	481588	1989	ACM SIGOIS Bulletin	".. the hottest topic i
403135	Moving Happily through the Worl...	38044	1996	IEEE Computer Graphics a...	"... people are requir
6355	HTML & XHTML: The Definitive G...	8592	2006	NOT SPECIFIED	"...lucid, in-depth de
336892	Municipal broadband wireless ne...	134831	2008	Communications of the AC...	"...people lack many
241258	.NET security	292428	2002	NOT SPECIFIED	".NET Security" is a t
494772	A Guide to Help Desk Concepts	509629	2009	NOT SPECIFIED	"A Guide to Service
17884	Practical Software Estimation: F...	21340	2007	NOT SPECIFIED	"A clearly written bo
175629	Idea for a mind	243182	1987	Issue 101 (July 1987)	"A computer can onl
238331	Vrrp: Increasing Reliability and F...	297726	2002	NOT SPECIFIED	"A detailed and clea
42573	Gray Hat Hacking, Second Editio...	10018	2007	NOT SPECIFIED	"A fantastic book fo
66798	MediaWiki, 1 edition	111970	2008	Wikipedia And Beyond	"A good book! It's a

2. Citation - *contains id's of all papers that are cited*

Primary key: (paper_id,cited_paper_id)

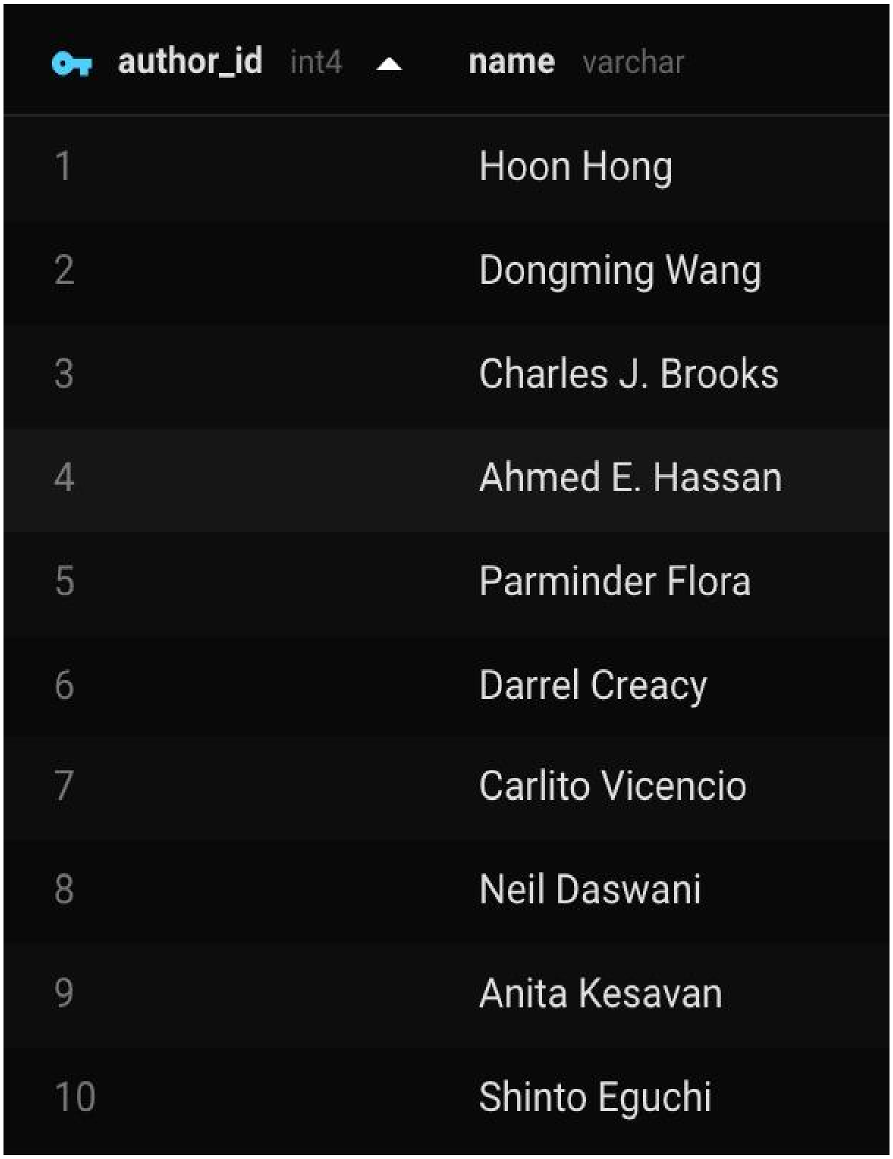
Foreign keys: paper_id,cited_paper_id



🔑 paper_id int4 ▲	🔑 cited_paper_id int4 ▲
5	436405
17	95255
17	96319
17	214023
17	294124
17	317448
17	319987
17	334185
17	357875
17	610127

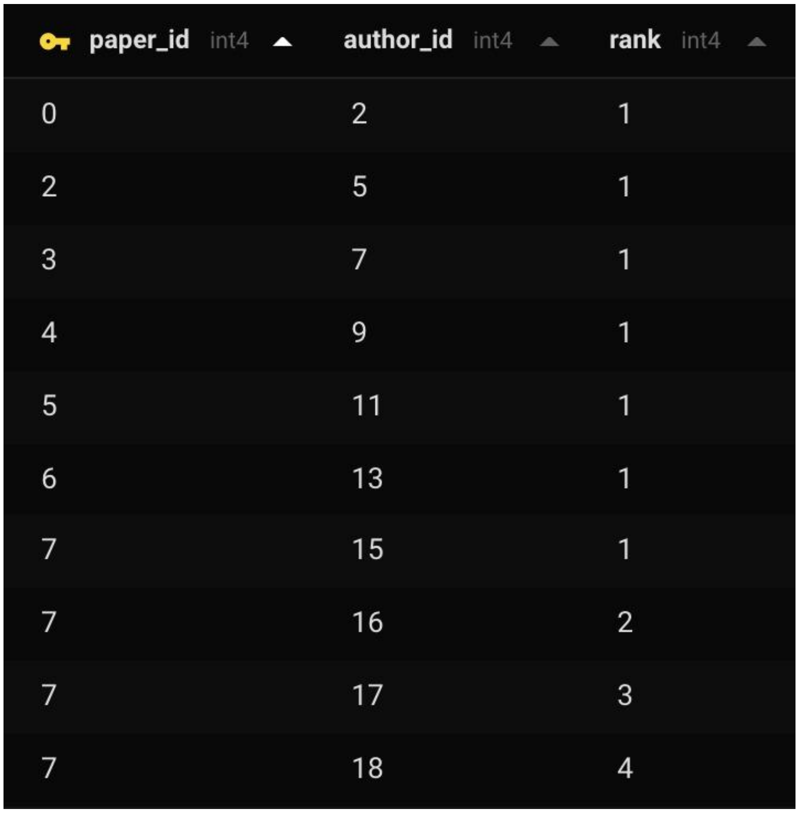
3. **AuthorInfo -** *every author is identified with a unique number*

Primary key: author_id
Foreign key: N/A



author_id int4 ▲	name varchar
1	Hoon Hong
2	Dongming Wang
3	Charles J. Brooks
4	Ahmed E. Hassan
5	Parminder Flora
6	Darrel Creacy
7	Carlito Vicencio
8	Neil Daswani
9	Anita Kesavan
10	Shinto Eguchi

4. **CoAuthors -** *has information about all the co-authors associated with the research papers along with their contribution(rank is given according to their contribution)*
Primary key:(paper_id, rank)
Foreign keys: paper_id,author_id



paper_id int4	author_id int4	rank int4
0	2	1
2	5	1
3	7	1
4	9	1
5	11	1
6	13	1
7	15	1
7	16	2
7	17	3
7	18	4

NOTE: (author_id,paper_id) can also be used as a primary key for the above table.