DSA5104 Assignment1

Wang Jiangyi, A0236307J
National University of Singapore

1 Task 2: Schema Diagram

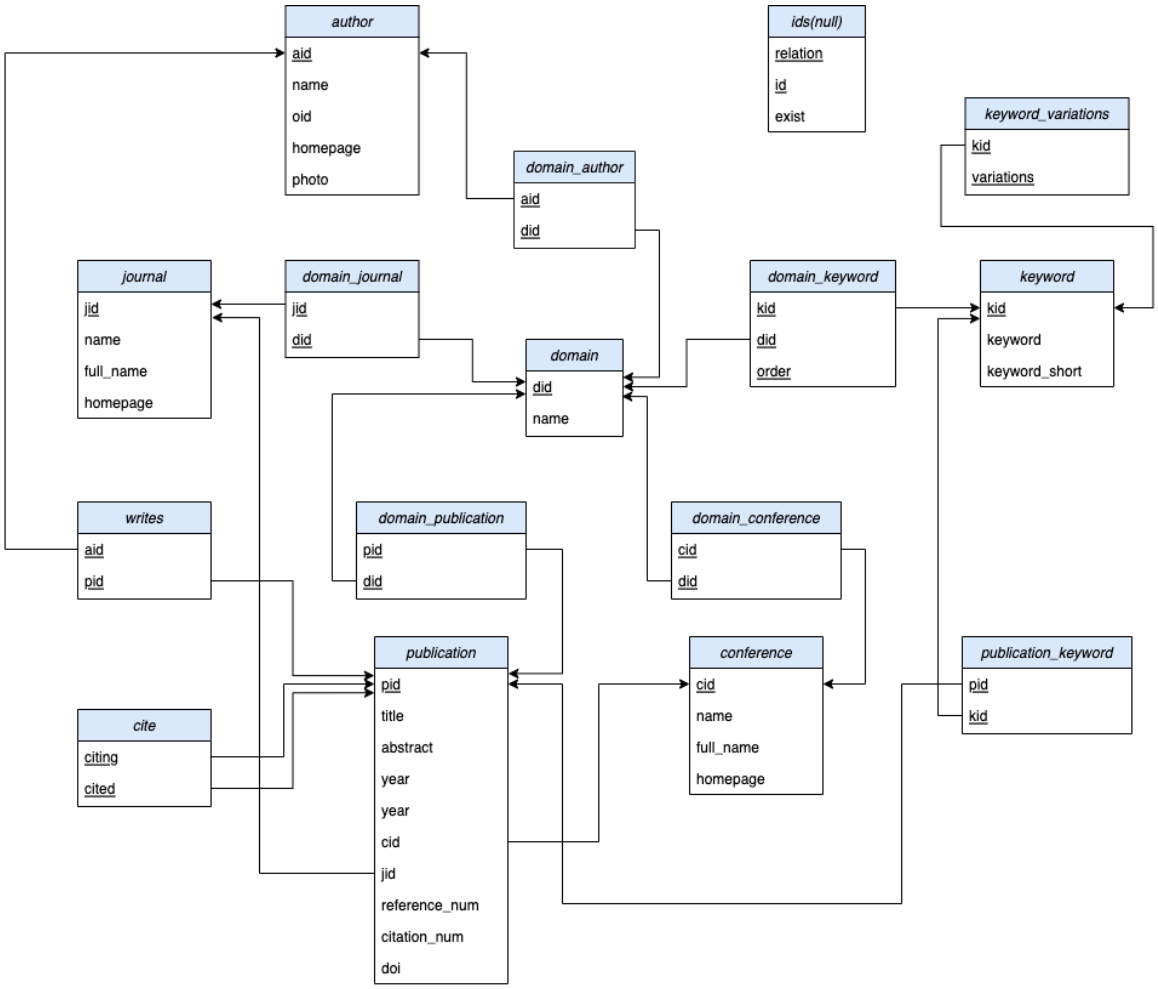

Figure 1: Relational Diagram for MAS

2 Task 3: NLQueries

In the beginning, I pointed out that, **for Query 7 and Query 13**, I have **different understanding of NL Query**. Therefore, I give all the answers corresponding to different understanding.

2.1 Query 1

NL Query: Return me the authors who have papers in PVLDB.

SQL Query: (2 methods)

```sql
# Q1: Return me the authors who have papers in PVLDB.

## Method 1:

WITH pid_in_PVLDB AS (
                SELECT pid
                FROM publication
                WHERE jid = (
                        SELECT jid
                        FROM journal
                        WHERE name = 'PVLDB'
                )
        )
SELECT a.name AS author_name
FROM author a
WHERE EXISTS (
        SELECT pid
        FROM pid_in_PVLDB
        WHERE pid IN (
                SELECT b.pid
                FROM writes b
                WHERE b.aid = a.aid
        )
);

## Method 2:

WITH author_flag AS (
                SELECT a.aid, b.name AS author_name
                        , CASE
                                WHEN a.pid IN (
                                                SELECT pid
                                                FROM publication
                                                WHERE jid = (
                                                        SELECT jid
                                                        FROM journal
                                                        WHERE name = 'PVLDB'
                                                )
                                        )
                                THEN 1
                                ELSE 0
                        END AS author_flag
                FROM writes a
                        LEFT JOIN author b ON a.aid = b.aid
        )
SELECT aid, author_name
FROM author_flag
GROUP BY aid, author_name
HAVING sum(author_flag) > 0;
```

Query Result: (1320 rows in total)

aid	author_name
1241	Chun Chen
1329	Philip S. Yu
4766	Yannis Sismanis
5061	Kyuseok Shim
8789	Jens Teubner
9074	Fernando C. Pereira
9728	Stefan Manegold
11196	Juliana Freire
11281	Minghua Chen
11314	Christian S. Jensen
13495	Anthony K. H. Tung
13599	Vivek Narasayya
16868	Phokion G. Kolaitis
20614	Stavros Harizopoulos
24944	Mark D. Hill
28766	Jian-Hua Feng
32465	Sabrina De Capitani d...
35950	Daniel J. Abadi

Figure 2: Result for query 1

2.2 Query 2

NL Query: Return me the organization H. V. Jagadish is in.

SQL Query:

```
# Q2: Return me the organization H. V. Jagadish is in.

SELECT a.name AS author_name, b.name AS organization_name
FROM (
        SELECT oid, name
        FROM author
        WHERE name = 'H. V. Jagadish'
) a
        LEFT JOIN organization b ON a.oid = b.oid;
```

Query Result: (1 row in total)

author_name	organization_name
H. V. Jagadish	University of Michigan

Figure 3: Result for query 2

2.3 Query 3

NL Query: Return me the authors who have papers in VLDB conference before 2002 after 1995.

SQL Query:

```
# Q3: Return me the authors who have papers in VLDB conference before 2002 after 1995.

WITH pid_VLDB_1995_2002 AS (
                SELECT pid
                FROM publication
                WHERE cid = (
                                SELECT cid
                                FROM conference
                                WHERE name = 'VLDB'
                        )
                        AND year > 1995
                        AND year < 2002
        )
```

```sql
SELECT b.aid AS author_id, b.name AS author_name
FROM (
        SELECT DISTINCT aid
        FROM writes
        WHERE pid IN (
                SELECT pid
                FROM pid_VLDB_1995_2002
        )
) a
        LEFT JOIN author b ON a.aid = b.aid;
```

Query Result: (984 rows in total)

author_id	author_name
636544	Praveen Seshadri
1037685	Miron Livny
1480663	Raghu Ramakrishnan
61208	Ashish Kumar Gupta
850509	Sunita Sarawagi
939673	Jeffrey Naughton
2016347	Prasad M. Deshpande
2209729	Rakesh Agrawal
56151118	Sameet Agarwal

Figure 4: Result for query 3

2.4 Query 4

NL Query: Return me the authors who have cooperated both with "H. V. Jagadish" and "Divesh Srivastava".

SQL Query:

```sql
# Q4:  Return me the authors who have cooperated both with "H. V. Jagadish" and "Divesh
    Srivastava".

WITH coop_HVJ AS (
                SELECT DISTINCT aid AS coop_HVJ_aid
                FROM writes
                WHERE pid IN (
                                SELECT pid AS pid_HVJ
                                FROM writes
                                WHERE aid = (
                                        SELECT aid
                                        FROM author
                                        WHERE name = 'H. V. Jagadish'
                                )
                        )
                        AND aid <> (
                                SELECT aid
                                FROM author
                                WHERE name = 'H. V. Jagadish'
                        )
        ),
        coop_DS AS (
                SELECT DISTINCT aid AS coop_DS_aid
                FROM writes
                WHERE pid IN (
                                SELECT pid AS pid_DS
                                FROM writes
                                WHERE aid = (
                                        SELECT aid
                                        FROM author
                                        WHERE name = 'Divesh Srivastava'
                                )
                        )
```

```
                        AND aid <> (
                                SELECT aid
                                FROM author
                                WHERE name = 'Divesh Srivastava'
                        )
        )
SELECT b.name AS coop_author_name
FROM (
        # coop_HVJ_aid interesects coop_DS_aid

        SELECT coop_HVJ_aid AS coop_aid
        FROM coop_HVJ
        WHERE coop_HVJ_aid IN (
                SELECT coop_DS_aid
                FROM coop_DS
        )
) a
        LEFT JOIN author b ON a.coop_aid = b.aid;
```

Query Result: (48 rows in total)

coop_author_name
S Muthukrishnan
Panagiotis G. Ipeirotis
Luis Gravano
Nick Koudas
Lauri Pietarinen
Shurug Al-khalifa
Cong Yu
Andrew Nierman
Jignesh M. Patel
Laks Lakshmanan

Figure 5: Result for query 4

2.5 Query 5

NL Query: Return me the authors who have more papers on VLDB than ICDE.

SQL Query:

```
# Q5: Return me the authors who have more papers on VLDB than ICDE.

WITH VLDB AS (
                SELECT a.aid AS author_id, count(publication_VLDB.pid) AS count_VLDB
                FROM writes a
                        LEFT JOIN (
                                SELECT pid
                                FROM publication
                                WHERE cid = (
                                        SELECT cid
                                        FROM conference
                                        WHERE name = 'VLDB'
                                )
                        ) publication_VLDB
                        ON a.pid = publication_VLDB.pid
                GROUP BY a.aid
        ),
        ICDE AS (
                SELECT a.aid AS author_id, count(publication_ICDE.pid) AS count_ICDE
                FROM writes a
                        LEFT JOIN (
                                SELECT pid
                                FROM publication
```

```
                                    WHERE cid = (
                                            SELECT cid
                                            FROM conference
                                            WHERE name = 'ICDE'
                                    )
                            ) publication_ICDE
                            ON a.pid = publication_ICDE.pid
                    GROUP BY a.aid
            )
SELECT b.name AS author_name
FROM (
            SELECT a.author_id AS author_id, a.count_VLDB AS count_VLDB, b.count_ICDE AS
    count_ICDE
            FROM VLDB a
                    LEFT JOIN ICDE b ON a.author_id = b.author_id
            WHERE a.count_VLDB > b.count_ICDE
) a
            LEFT JOIN author b ON a.author_id = b.aid;
```

Query Result: (2627 rows in total)

author_name
Debra E. Vandermeer
Yannis Sismanis
Michael J. Lopez
Kyuseok Shim
Hans-dieter Ehrich
David Botzer
Jurgen Wasch
Ricardo A. Baeza-yates
Shlomo Moran
Peter M. Schwarz

Figure 6: Result for query 5

2.6 Query 6

NL Query: Return me the authors who have cited papers of H. V. Jagadish.

SQL Query: (2 methods)

```
# Q6: Return me the authors who have cited papers of H. V. Jagadish

WITH citing_HVJ AS (
                    SELECT citing
                    FROM cite
                    WHERE cited IN (
                            SELECT pid
                            FROM writes
                            WHERE aid = (
                                    SELECT aid
                                    FROM author
                                    WHERE name = 'H. V. Jagadish'
                            )
                    )
            )
SELECT DISTINCT b.aid AS author_id, b.name AS author_name
FROM citing_HVJ a
            LEFT JOIN (
                    SELECT a1.pid, a2.aid, a2.name
                    FROM writes a1
                            LEFT JOIN author a2 ON a1.aid = a2.aid
            ) b
            ON a.citing = b.pid
WHERE b.aid IS NOT NULL;
```

```
25
26 # method 2
27
28 WITH citing_HVJ AS (
29                 SELECT citing
30                 FROM cite
31                 WHERE cited IN (
32                         SELECT pid
33                         FROM writes
34                         WHERE aid = (
35                                 SELECT aid
36                                 FROM author
37                                 WHERE name = 'H. V. Jagadish'
38                         )
39                 )
40         )
41 SELECT aid AS author_id, name AS author_name
42 FROM author
43 WHERE aid IN (
44         SELECT aid
45         FROM writes
46         WHERE pid IN (
47                 SELECT citing
48                 FROM citing_HVJ
49         )
50 );
```

Query Result: (7245 rows in total)

author_id	author_name
1241	Chun Chen
1242	Chun-Yi Shi
1329	Philip S. Yu
1486	Lukas Relly
2087	Rossana Maria de Ca...
2945	Clara Pizzuti
3253	Philip W. Trinder
3503	Young-whun Lee
3662	Debra E. Vandermeer
3715	Christian Capelle

Figure 7: Result for query 6

2.7 Query 7

NL Query: Return me all the papers, which contain the keyword "Natural Language".

***Note**: There are two different understanding of this NL Queriy. One is, the keyword contains "Natural Language" and the other is, the keyword is excatly "Natural Language". The 2 different SQL Queries can ben shown as follows:*

SQL Query 1, exactly 'Natural Language':

```
1  # Q7:  Return me all the papers, which contain the keyword "Natural Language".
2
3  SELECT b.pid AS publication_id, b.title AS paper_name
4  FROM (
5          SELECT DISTINCT pid
6          FROM publication_keyword
7          WHERE kid IN (
8                  SELECT kid
9                  FROM keyword
10                 WHERE keyword = 'Natural Language'
11         )
12 ) a
13         LEFT JOIN publication b ON a.pid = b.pid;
```

Query Result: (11232 rows in total)

publication_id	paper_name
38	Using Natural Language for Database Design
323	Exploiting lexical regularities in designing natural language systems
474	Reasoning about Information Change
511	Typed Logics With States
902	User-Needs Analysis and Design Methodology for an Automated Documentation Generator
1426	Making Systems Sensitive to the User's Time and Working Memory Constraints
1460	MultiDimensional User Models for Multimedia I/O in the Maintenance Consultant
1474	Natural language processing using a propositional semantic network with structured variables
1922	SUPPORTING FLEXIBILITY AND TRANSMUTABILITY: MULTI-AGENT PROCESSING AND ROLE-SWITCHING I...
2152	A generic algorithm for generating spoken monologues

Figure 8: Result for query 7, version 1

SQL Query 2, contains 'Natural Language':

```
# Q7:  Return me all the papers, which contain the keyword "Natural Language".

SELECT b.pid AS publication_id, b.title AS paper_name
FROM (
        SELECT DISTINCT pid
        FROM publication_keyword
        WHERE kid IN (
                SELECT kid
                FROM keyword
                WHERE keyword LIKE '%Natural Language%'
        )
) a
        LEFT JOIN publication b ON a.pid = b.pid;
```

Query Result: (19208 rows in total)

publication_id	paper_name
38	Using Natural Language for Database Design
323	Exploiting lexical regularities in designing natural language systems
474	Reasoning about Information Change
511	Typed Logics With States
692	Higher-Order Logic Programming as Constraint Logic Programming
807	Concept-Based Retrieval using Controlled Natural Language
834	Empirical learning of natural language processing tasks
902	User-Needs Analysis and Design Methodology for an Automated Documentation Generator
1416	Integrated Natural Language Generation Systems
1426	Making Systems Sensitive to the User's Time and Working Memory Constraints

Figure 9: Result for query 7, version 2

2.8 Query 8

NL Query: Return me all the researchers in database area in University of Michigan.

SQL Query:

```
# Q8: Return me all the researchers in database area in University of Michigan.

SELECT aid AS author_id, name AS author_name
FROM author
WHERE oid = (
                SELECT oid
                FROM organization
                WHERE name = 'University of Michigan'
        )
        AND aid IN (
                SELECT aid
                FROM domain_author
                WHERE did = (
                        SELECT did
                        FROM domain
```

```
                        WHERE name = 'Databases'
                    )
        );
```

Query Result: (146 rows in total)

author_id	author_name
73839	Andrew Nierman
75475	Shuming Bao
105297	Wee Teck Ng
141842	Mark E. Deppe
144550	T. Ceccarelli
160200	Michael J. Cafarella
244284	Alan G. Merten
326823	Toby J. Teorey
360926	Brahim Medjahed
365133	Mark S. Ackerman

Figure 10: Result for query 8

2.9 Query 9

NL Query: Return me the number of papers written by H. V. Jagadish, Yunyao Li, and Cong Yu.

SQL Query:

```
# Q9:  Return me the number of papers written by H. V. Jagadish, Yunyao Li, and Cong Yu.

WITH satisfied_papers AS (
                SELECT a.pid
                FROM (
                        SELECT pid AS pid
                            , CASE
                                    WHEN aid IN (
                                                    SELECT aid
                                                    FROM author
                                                    WHERE name IN ('H. V. Jagadish', '
    Yunyao Li', 'Cong Yu')
                                                )
                                    THEN 1
                                    ELSE 0
                                END AS flag
                        FROM writes
                ) a
                GROUP BY a.pid
                HAVING sum(a.flag) = 3
        )
SELECT count(pid) AS count_papers
FROM satisfied_papers;
```

Query Result: (1 row in total)

count_papers
3

Figure 11: Result for query 9

2.10 Query 10

NL Query: Return me the number of papers written by H. V. Jagadish in each year.

SQL Query:

```
# Q10: Return me the number of papers written by H. V. Jagadish in each year.

SELECT b.year AS year, count(b.pid) AS count_paper_each_year
```

```
FROM (
        SELECT pid
        FROM writes
        WHERE aid = (
                SELECT aid
                FROM author
                WHERE name = 'H. V. Jagadish'
        )
) a
        LEFT JOIN publication b ON a.pid = b.pid
GROUP BY b.year
ORDER BY b.year ASC;
```

Query Result: (29 rows in total)

year	count_paper_each_y...
0	13
1984	1
1986	1
1987	7
1988	8
1989	14
1990	8
1991	8
1992	18
1993	8

Figure 12: Result for query 10

2.11 Query 11

NL Query: Return me the number of citations of "Making database systems usable" in each year.

SQL Query:

```
# Q11: Return me the number of citations of "Making database systems usable" in each year.

WITH paper_citing_MDSU AS (
                SELECT citing AS citing_pid
                FROM cite
                WHERE cited = (
                        SELECT pid
                        FROM test.publication
                        WHERE title = 'Making database systems usable'
                )
        )
SELECT b.year AS year, count(b.pid) AS count_citation_each_year
FROM paper_citing_MDSU a
        LEFT JOIN publication b ON a.citing_pid = b.pid
GROUP BY b.year
ORDER BY b.year ASC;
```

Query Result: (5 rows in total)

year	count_citation_each_year
0	1
2008	5
2009	11
2010	8
2011	4

Figure 13: Result for query 11

2.12 Query 12

NL Query: Return me the author who has the most number of papers in the VLDB conference.

SQL Query:

```sql
# Q12: Return me the author who has the most number of papers in the VLDB conference.

WITH count_paper_author_VLDB AS (
                SELECT b.aid AS author_id, count(a.pid) AS count_paper_VLDB
                FROM (
                        SELECT pid
                        FROM publication
                        WHERE cid = (
                                SELECT cid
                                FROM conference
                                WHERE name = 'VLDB'
                        )
                ) a
                        LEFT JOIN writes b ON a.pid = b.pid
                GROUP BY b.aid
        )
SELECT b.name AS author_name, a.count_paper_VLDB
FROM (
        SELECT author_id, count_paper_VLDB
        FROM count_paper_author_VLDB
        WHERE count_paper_VLDB = (
                SELECT max(count_paper_VLDB)
                FROM count_paper_author_VLDB
        )
) a
        LEFT JOIN author b ON a.author_id = b.aid;
```

Query Result: (1 row in total)

author_name	count_paper_VLDB
H. V. Jagadish	37

Figure 14: Result for query 12

2.13 Query 13

NL Query: Return me the conferences, which have more than 60 papers containing keyword "Relational Database".

***Note**: There are two different understanding of this NL Queriy. One is, the keyword contains "Relational Database" and the other is, the keyword is excatly "Relational Database". The 2 different SQL Queries can ben shown as follows:*

SQL Query 1, exactly "Relational Database": (2 methods)

```sql
# Q13: Return me the conferences, which have more than 60 papers containing keyword "
    Relational Database".

WITH paper_contain_RD AS (
                SELECT DISTINCT pid
                FROM publication_keyword
                WHERE kid IN (
                        SELECT kid
                        FROM keyword
                        WHERE keyword = 'Relational Database'
                )
        )
SELECT a.name AS conference_name, count(b.pid) AS count_paper_in_conference
FROM conference a
        LEFT JOIN (
                SELECT pid, cid
                FROM publication
                WHERE pid IN (
                        SELECT *
                        FROM paper_contain_RD
                )
        ) b
```

```
        ON a.cid = b.cid
GROUP BY a.name
HAVING count_paper_in_conference > 60
ORDER BY count_paper_in_conference DESC;

# method 2

WITH conference_id_count AS (
                SELECT cid, count(pid) AS count_cid
                FROM publication
                WHERE pid IN (
                        SELECT pid
                        FROM publication_keyword
                        WHERE kid = (
                                SELECT kid
                                FROM keyword
                                WHERE keyword = 'Relational Database'
                        )
                )
                GROUP BY cid
        )
SELECT name
FROM conference
WHERE cid IN (
        SELECT cid
        FROM conference_id_count
        WHERE count_cid > 60
);
```

Query Result: (5 rows in total)

conference_name	count_paper_in_conference
ICDE	157
DEXA	122
SIGMOD	97
VLDB	93
ER(OOER)	65

Figure 15: Result for query 13, version 1

SQL Query 2, contains "Relational Database": (2 methods)

```
# Q13: Return me the conferences, which have more than 60 papers containing keyword "
    Relational Database".

WITH paper_contain_RD AS (
                SELECT DISTINCT pid
                FROM publication_keyword
                WHERE kid IN (
                        SELECT kid
                        FROM keyword
                        WHERE keyword LIKE '%Relational Database%'
                )
        )
SELECT a.name AS conference_name, count(b.pid) AS count_paper_in_conference
FROM conference a
        LEFT JOIN (
                SELECT pid, cid
                FROM publication
                WHERE pid IN (
                        SELECT *
                        FROM paper_contain_RD
                )
        ) b
        ON a.cid = b.cid
GROUP BY a.name
HAVING count_paper_in_conference > 60
ORDER BY count_paper_in_conference DESC;

# method 2
```

```
28
29 WITH conference_id_count AS (
30                 SELECT cid, count(pid) AS count_cid
31                 FROM publication
32                 WHERE pid IN (
33                         SELECT pid
34                         FROM publication_keyword
35                         WHERE kid IN (
36                                 SELECT kid
37                                 FROM keyword
38                                 WHERE keyword LIKE '%Relational Database%'
39                         )
40                 )
41                 GROUP BY cid
42         )
43 SELECT name
44 FROM conference
45 WHERE cid IN (
46         SELECT cid
47         FROM conference_id_count
48         WHERE count_cid > 60
49 );
```

Query Result: (7 rows in total)

conference_name	count_paper_in_conference
ICDE	212
DEXA	156
VLDB	143
SIGMOD	133
ER(OOER)	69
PODS	67
SAC	62

Figure 16: Result for query 13, version 2

2.14 Query 14

NL Query: Return me the number of papers published on PVLDB each year after 2000.

SQL Query:

```
1  # Q14:  Return me the number of papers published on PVLDB each year after 2000.
2
3  SELECT year AS year, count(pid) AS count_paper_each_year
4  FROM publication
5  WHERE jid = (
6                  SELECT jid
7                  FROM journal
8                  WHERE name = 'PVLDB'
9          )
10         AND year > 2000
11 GROUP BY year
12 ORDER BY year ASC;
```

Query Result: (5 rows in total)

year	count_paper_each_y...
2008	162
2009	160
2010	189
2011	29
2014	1

Figure 17: Result for query 14

2.15 Query 15

NL Query: Return me the paper after 2000 in VLDB conference with the most citations.

SQL Query:

```sql
# Q15:  Return me the paper after 2000 in VLDB conference with the most citations

WITH VLDB_paper_after_2000_citation AS (
                SELECT pid AS publication_id, title AS publication_title, citation_num AS
    count_citation
                FROM publication
                WHERE cid = (
                                SELECT cid
                                FROM conference
                                WHERE name = 'VLDB'
                        )
                        AND year > 2000
        )
SELECT *
FROM VLDB_paper_after_2000_citation
WHERE count_citation = (
        SELECT max(count_citation)
        FROM VLDB_paper_after_2000_citation
);
```

Query Result: (1 row in total)

publication_id	publication_title	count_citation
133771	Generic Schema Matching with Cupid	798

Figure 18: Result for query 15