DAS 839 NoSQL Systems

Assignment-1

Submitted by :

Group 2:

Abhay bhadouriya – MT2024003

jainish parmar - MT2024065

Jaimin Jadvani – MT2024064

Files Submitted :

1. Report File

Section A

Problem – 1:

```
SELECT Team
FROM Scores
WHERE RunsFor > RunsAgainst AND
      RunsFor <= RunsAgainst + 2
```

This query filters the records from the Scores table based on following two conditions:

1. RunsFor > RunsAgainst → The team must have won the match.
2. RunsFor <= RunsAgainst + 2 → The winning margin must be at most 2 runs.

In Below Screenshot we have illustrated how the query is fetching data from Database.

Out of Dragons, Marines, Swallows only Marines is given in options.

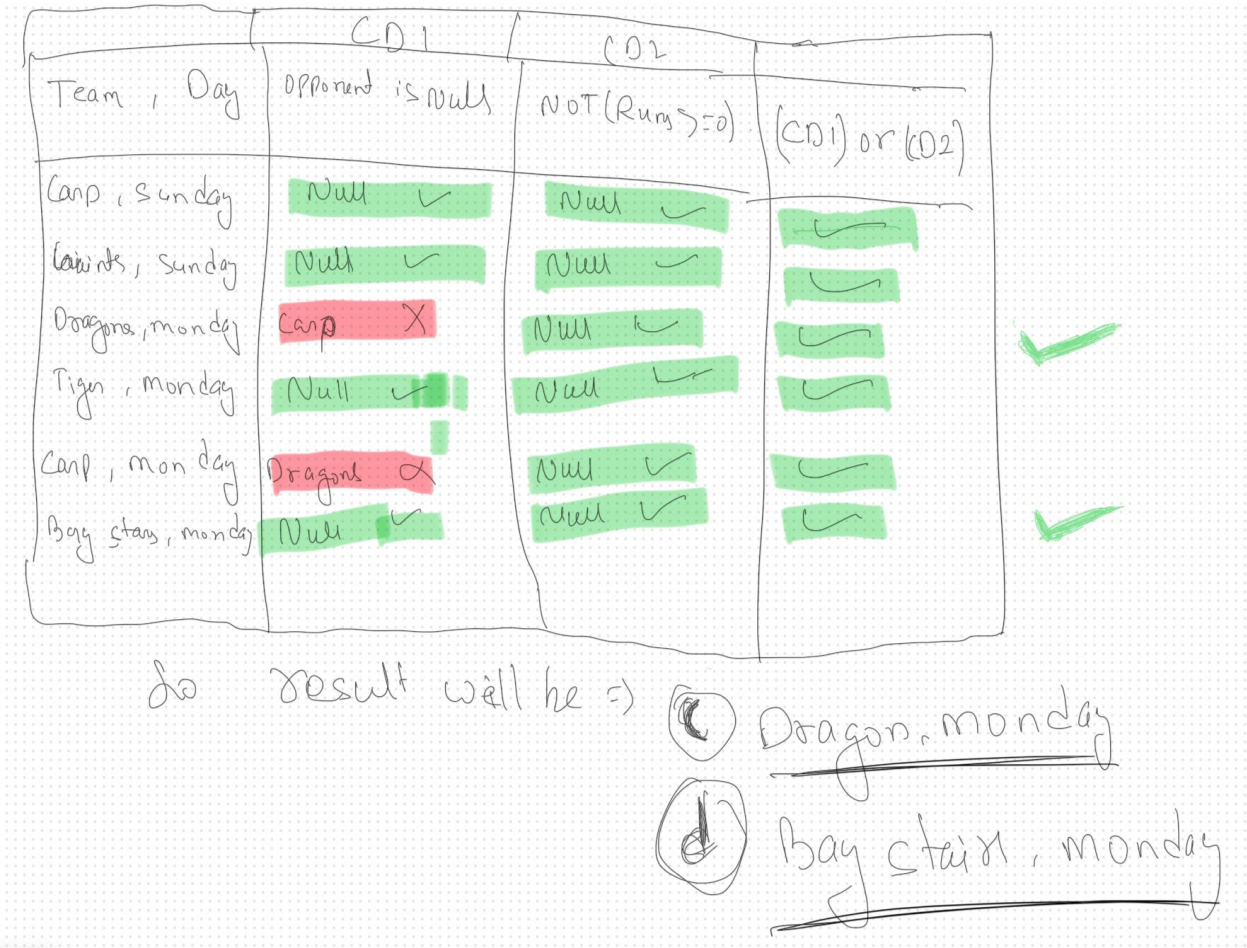
So out ans is **A. Marines**

Problem – 2:

```
SELECT Team, Day
FROM Scores
WHERE Opponent IS NULL OR
      NOT (Runs >= 0)
```

This query filters the records from the Scores table based on following two conditions:

1. Opponent IS NULL: Selects rows where the team did not play on that day.
2. NOT (Runs >= 0): This means Runs is NULL, which happens when the game is not yet concluded or the team did not play.



Applying the Query Conditions:

1. Teams where Opponent IS NULL (did not play):
    a. Sunday: Carp, Giants
    b. Monday: Tigers, Bay Stars
2. Teams where Runs IS NULL (game not concluded or did not play):
    a. Monday: Dragons, Carp

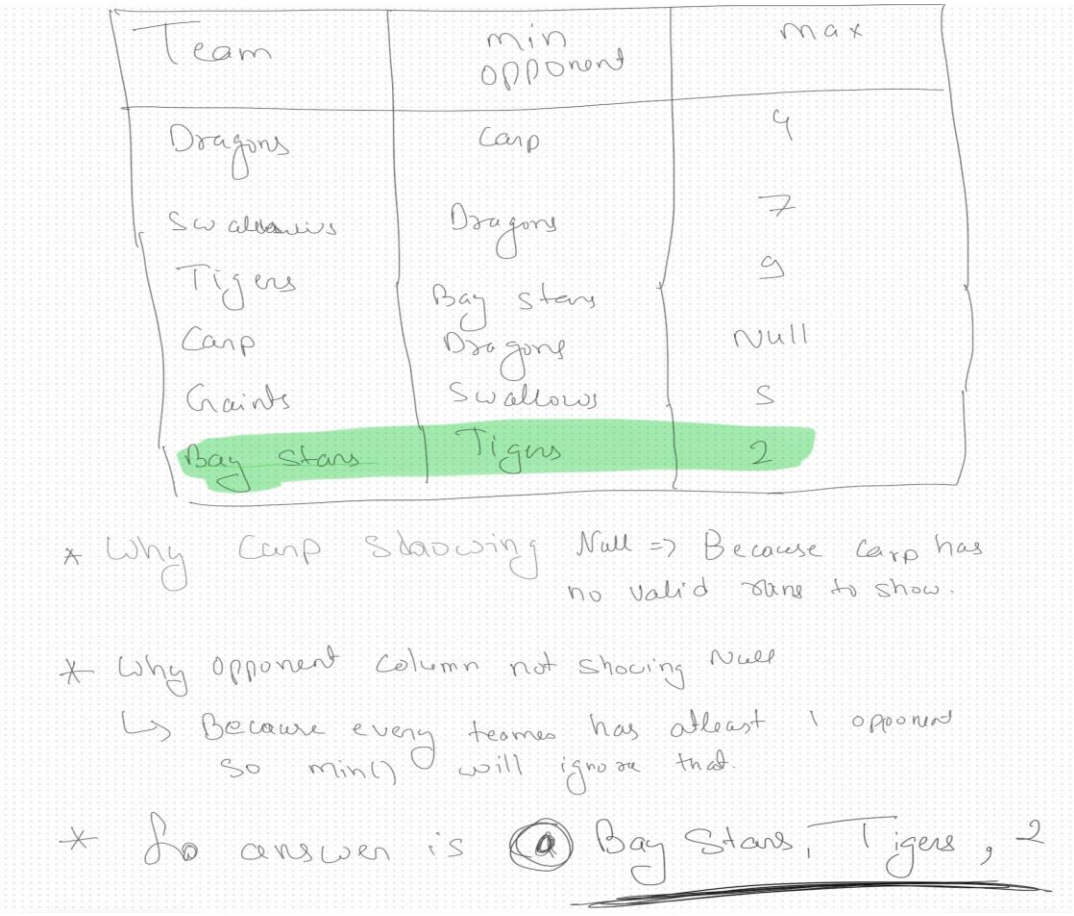
Correct Answers: (c) Dragons, Monday and (d) Bay Stars, Monday

Problem 3:

```
SELECT Team, Min(Opponent), Max(Runs)
FROM Scores
GROUP BY Team
```

This query filters the records from the Scores table based on following conditions:

1. Opponent IS NULL: Selects rows where the team did not play on that day.
2. NOT (Runs >= 0): This means Runs is NULL, which happens when the game is not yet concluded or the team did not play.
3. Min(Opponent): Picks the lexicographically smallest opponent name per team.
4. Max(Runs): Picks the highest runs per team.



Applying the Query Conditions:

1. Teams where Opponent IS NULL (did not play):

a. Sunday: Carp, Giants
b. Monday: Tigers, Bay Stars

2. Teams where Runs IS NULL (game not concluded or did not play):
    a. Monday: Dragons, Carp

Correct Answers: (a) Bay Stars, Tigers, 2

Problem – 4

```
SELECT *
FROM Scores
ORDER BY Runs DESC, Team ASC
```

ORDER BY Runs DESC, Team ASC:

- Sort primarily by Runs in descending order.
- If two rows have the same Runs, sort alphabetically by Team in ascending order.
- NULL values appear at the bottom in ORDER BY DESC.

Team ▲ 2	Day	Opponent	Runs ▼ 1
Tigers	Sunday	Bay Stars	9
Bay Stars	Monday	*NULL*	7
Swallows	Sunday	Dragons	7
Dragons	Monday	Carp	6
Giants	Monday	Swallows	5
Tigers	Monday	*NULL*	5
Dragons	Sunday	Swallows	4
Bay Stars	Sunday	Tigers	2
Swallows	Monday	Giants	0
Carp	Sunday	*NULL*	*NULL*
Carp	Monday	Dragons	*NULL*
Giants	Sunday	*NULL*	*NULL*

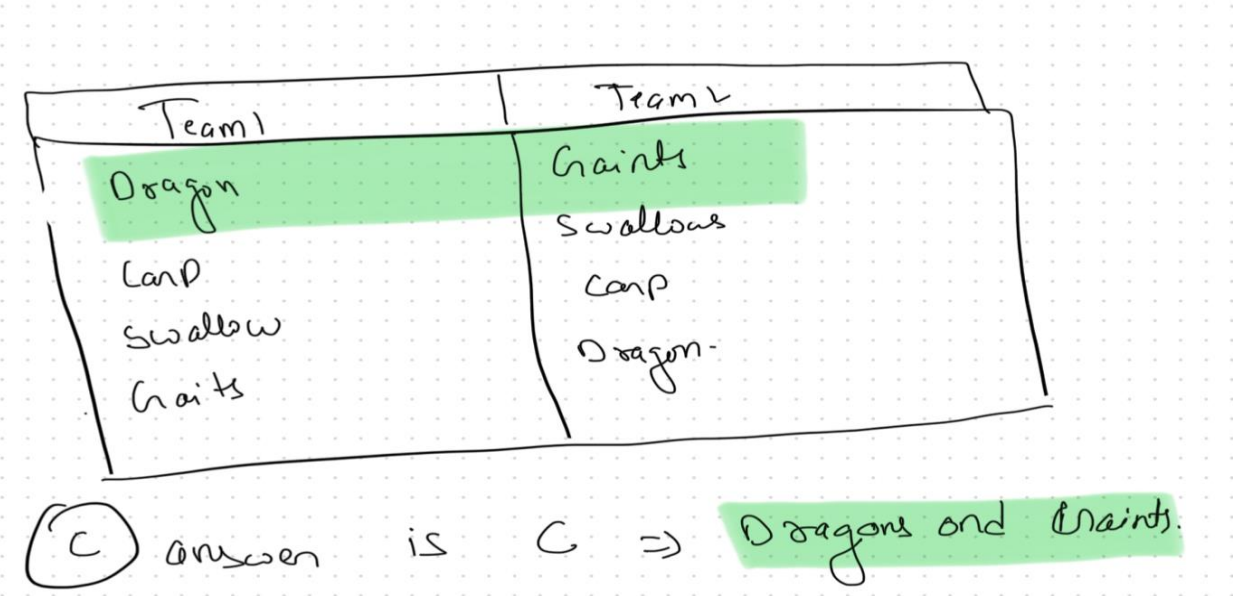
The Option **a) Giants, Sunday, NULL, NULL could appear seventh through twelfth.** Is true because it appreared at 12th postion in sql query. So **A** is the answer

Problem – 5

```
SELECT S1.Team, S2.Team
FROM Scores S1, Scores S2
WHERE S1.Opponent = S2.Opponent
    AND S1.Team <> S2.Team
```

This query:

1. Self-joins the Scores table (treating it as two copies, S1 and S2).
2. Finds rows where both teams played against the same opponent (S1.Opponent = S2.Opponent).
3. Ensures that a team is not paired with itself (S1.Team <> S2.Team).



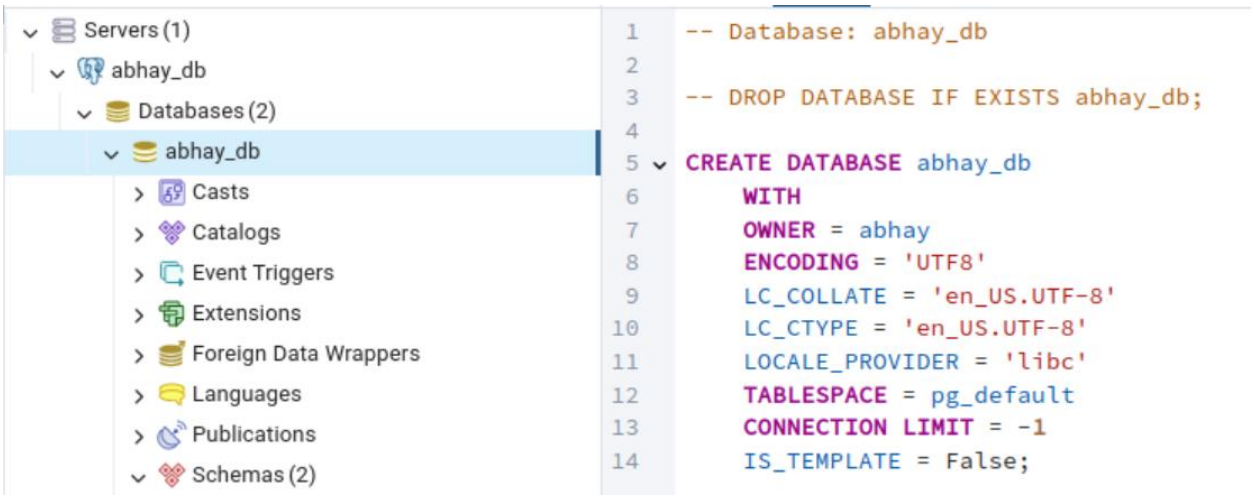
Dragons and Giants Both played against Carp on different days (Sunday & Monday).

Correct options: (c) Dragons, Giants .

Section 2

Problem 1. Bulkloading Data into a PostgreSQL Database

1.1 Database Creation (abhay_db)

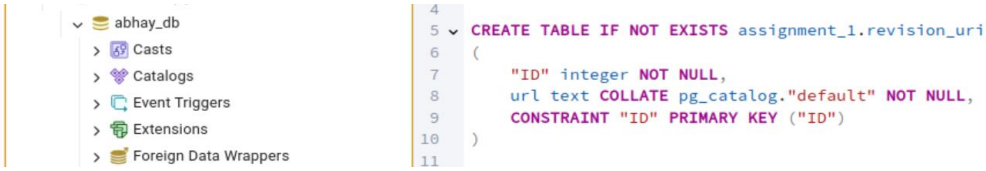


- Created A PostgreSQL database named **abhay_db** and database owner set to **abhay**.
- Encoding is **UTF-8**, with collation and character type as **en_US.UTF-8**.
- The default **tablespace** is used (pg_default).
- No connection limit (-1 means unlimited).
- It is **not a template** database.
- **NOTE: we created table using PGADMIN4 So some of the attribute are by default added by PGADMIN4 it is not an AI generated query**

1.2 Schema Creation (assignment_1)

- Schema named **assignment_1** is created within abhay_db and only authorized for user **abhay**. And create only if it does not already exist.

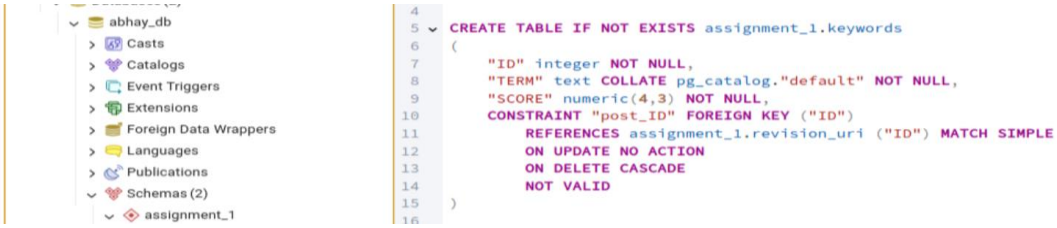
1.3 Table Creation (assignment_1.revision_uri)



- Creating the table revision_uri under the assignment_1 schema.

- And Defining the two columns: ID (integer, primary key) and url (text,not null).

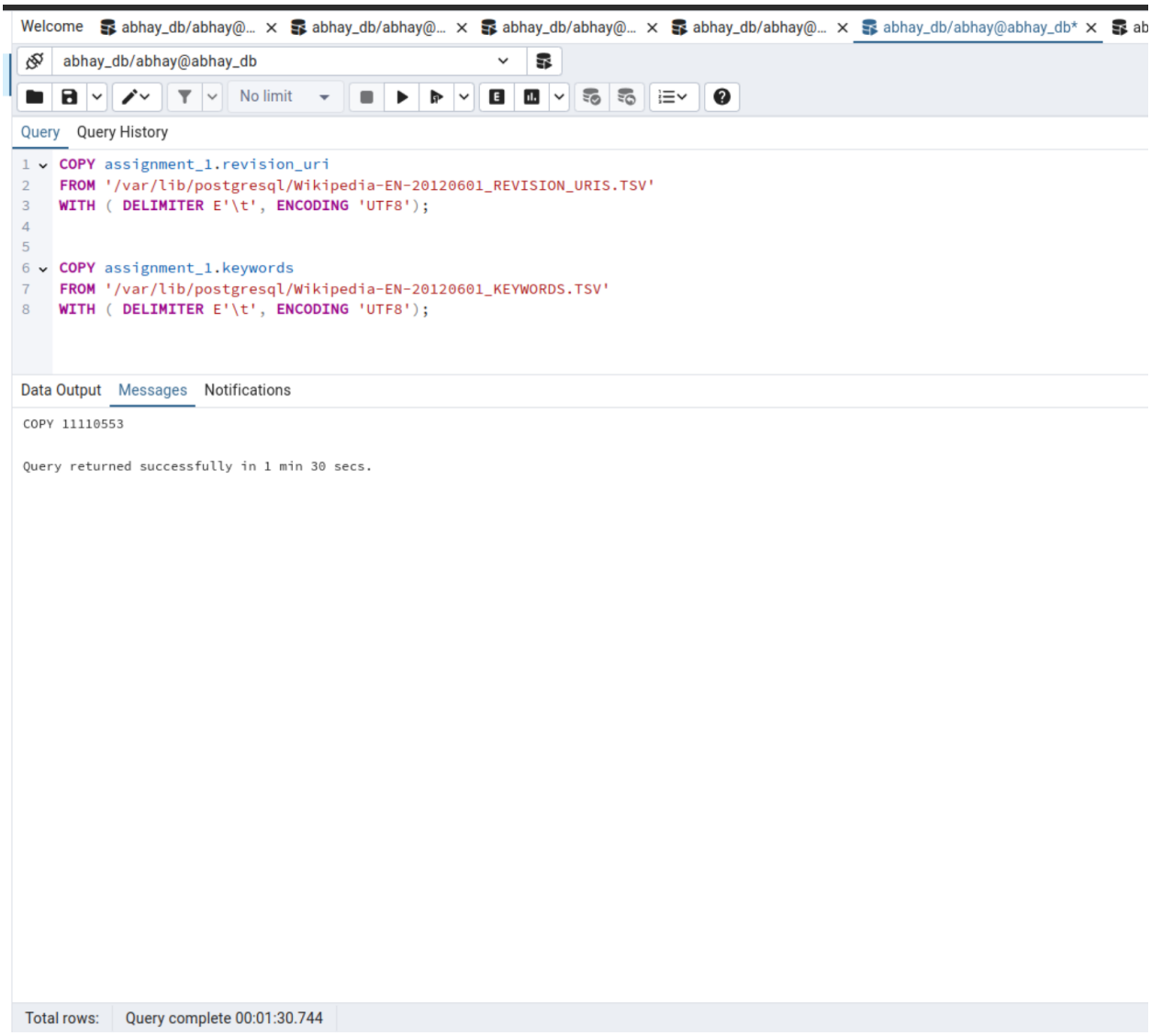
1.4 Table Creation (assignment_1.keywords)



- Creating ID as Primary key integer
- Creating Term as Text and Not null
- Foreign key ensures data integrity, linking keywords to revision_uri.
- Cascade delete ensures that when a related revision_uri entry is deleted, corresponding keywords entries are removed.
- Creating column numeric type for SCORE, allowing up to 4 digits with 3 decimal places.

1.5 Inserting data in assignment_1.revision_uri and assignment_1.keywords with Foreign key constraint

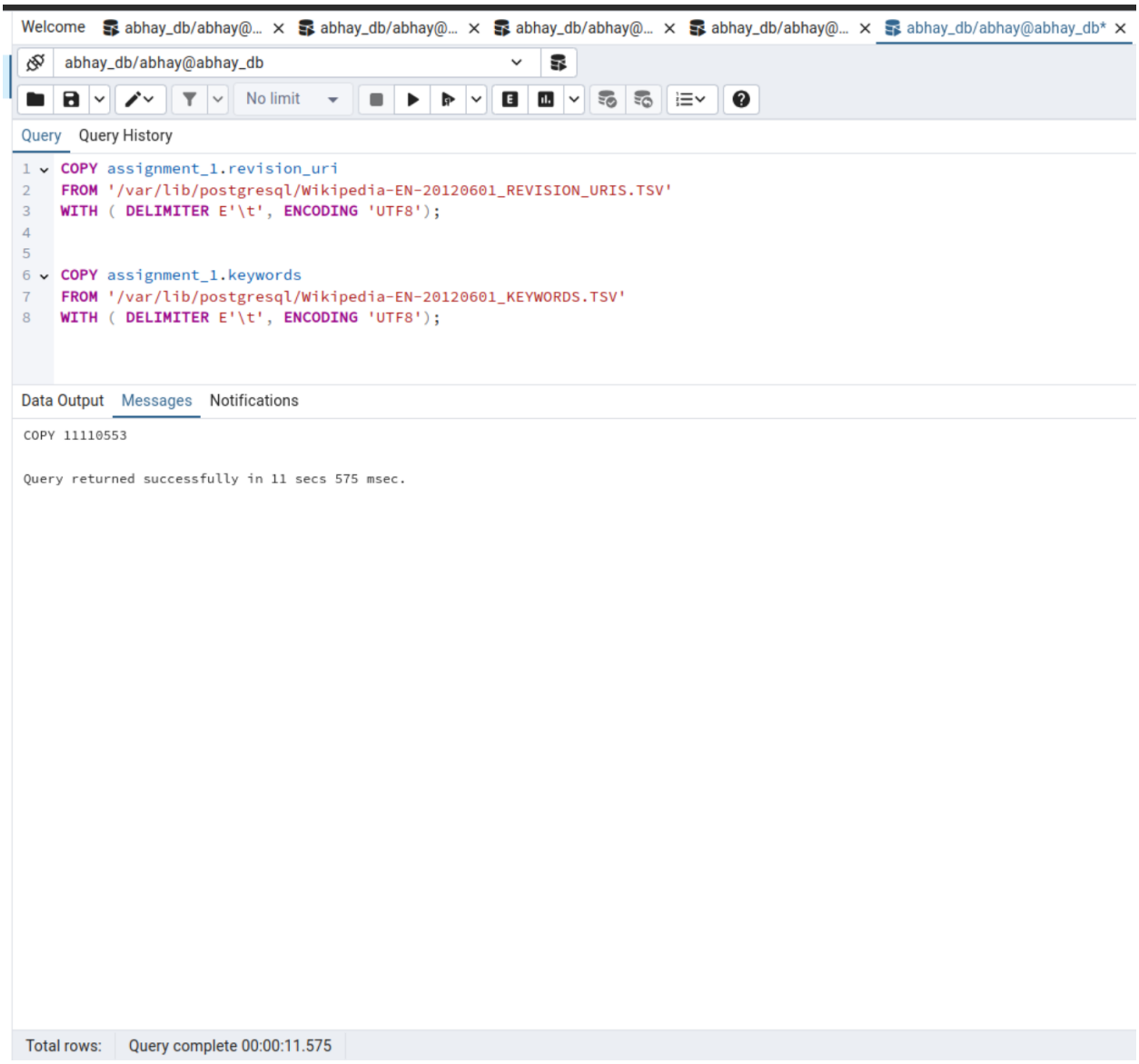
- Importint the data from a TSV files into the revision_uri and keywords table.
- Uses UTF-8 encoding and tab (\t) as a delimiter.
- It took **1 minute 30 seconds** to Successfully copied **11,110,553 rows** in tables, using key constraint.



1.6 Creating keyword table without foreign key constraint

- Again Creating table keywords under assignment_1 schema.
- Defines three columns: ID (integer), TERM (text), and SCORE (numeric with precision 4,3), but without any key constraint.

1.7 Again Bulk Loading the data



- Importing the data from a TSV files into the revision_uri and keywords table.
- Uses UTF-8 encoding and tab (\t) as a delimiter.
- It took **11 seconds 575 mili seconds** to Successfully copied **11,110,553 rows** in tables, without using foreign key constraint.

Problem 2 - Running Keyword Queries over Wikipedia

2.1 Query 1 - The query returns the **URLs of Wikipedia articles** that contain **all four stemmed keywords** (infantri, reinforc, brigad, and fire). If an article is missing even **one** of the keywords, it will not be included in the results.

Query Query History

```
1  SELECT t1.url
2  FROM assignment_1.revision_uri t1
3  JOIN assignment_1.keywords t2 ON t1."ID" = t2."ID"
4  WHERE t2."TERM" IN ('infantri', 'reinforc', 'brigad', 'fire')
5  GROUP BY t1.url
6  HAVING COUNT(DISTINCT t2."TERM") = 4;
7
```

And the result is

Data Output Messages Notifications

SQL

	url text
1	http://en.wikipedia.org/wiki/2003_invasion_of_Iraq?oldid=495060157
2	http://en.wikipedia.org/wiki/2nd_Armored_Division_(United_States)?oldid=494050221
3	http://en.wikipedia.org/wiki/Albert_Kesselring?oldid=493045790
4	http://en.wikipedia.org/wiki/Anglo-Nepalese_War?oldid=493993887
5	http://en.wikipedia.org/wiki/Battle_of_Boulou?oldid=490141720
6	http://en.wikipedia.org/wiki/Battle_of_Camden?oldid=490509731
7	http://en.wikipedia.org/wiki/Battle_of_Chickamauga?oldid=495189001
8	http://en.wikipedia.org/wiki/Battle_of_Dak_To?oldid=494048995
9	http://en.wikipedia.org/wiki/Battle_of_Fort_Donelson?oldid=495189282
10	http://en.wikipedia.org/wiki/Battle_of_Isandlwana?oldid=495315847
11	http://en.wikipedia.org/wiki/Battle_of_Landen?oldid=476011327
12	http://en.wikipedia.org/wiki/Battle_of_New_Orleans?oldid=494895989
13	http://en.wikipedia.org/wiki/Battle_of_Passchendaele?oldid=495266730
14	http://en.wikipedia.org/wiki/Battle_of_Salamanca?oldid=494418014
15	http://en.wikipedia.org/wiki/Battle_of_Stones_River?oldid=495190457
16	http://en.wikipedia.org/wiki/Battle_of_Vimy_Ridge?oldid=494032814
17	http://en.wikipedia.org/wiki/Battle_of_York?oldid=495560649
18	http://en.wikipedia.org/wiki/Bernard_Montgomery,_1st_Viscount_Montgomery_of_Alamein?oldid=494673033
19	http://en.wikipedia.org/wiki/Brazilian_Armed_Forces?oldid=494507454
20	http://en.wikipedia.org/wiki/British_Expeditionary_Force_(World_War_I)?oldid=483791980
21	http://en.wikipedia.org/wiki/Continuation_War?oldid=494901078
22	http://en.wikipedia.org/wiki/Devil's_Brigade?oldid=493327271
23	http://en.wikipedia.org/wiki/Ethiopian_National_Defense_Force?oldid=494732082
24	http://en.wikipedia.org/wiki/Indonesiau2013Malaysia_confrontation?oldid=495193359
25	http://en.wikipedia.org/wiki/Lion_of_Babylon_(tank)?oldid=491695462
26	http://en.wikipedia.org/wiki/List_of_battles_1801u20131900?oldid=492168405
27	http://en.wikipedia.org/wiki/Luxembourg_Army?oldid=493333664
28	http://en.wikipedia.org/wiki/Mike_Jackson?oldid=493067138
29	http://en.wikipedia.org/wiki/Operation_Pegasus?oldid=483932614

Total rows: 35 Query complete 00:00:01.285

we have assumed that we have to look for the words exactly mentioned in the doc and not look for words whose stemming gives the following outputs. If that is the case then the query would change as follows:

Query Query History

```
1  SELECT t1.url
2  FROM assignment_1.revision_uri t1
3  JOIN assignment_1.keywords t2 ON t1."ID" = t2."ID"
4  WHERE t2."TERM" like '%infantri%' or t2."TERM" like '%reinforc%' or t2."TERM" like '%brigad%'or t2."TERM" like '%fire%'
5  GROUP BY t1.url
6  HAVING COUNT(DISTINCT t2."TERM") = 4;
7
```

Data Output Messages Notifications

SQL Showing rows: 1 to 66 Page No: 1 of 1

	url text
1	http://en.wikipedia.org/wiki/1952?oldid=494937414
2	http://en.wikipedia.org/wiki/1st_Czechoslovak_Armoured_Brigade?oldid=492150931
3	http://en.wikipedia.org/wiki/Akron,_Ohio?oldid=495317194
4	http://en.wikipedia.org/wiki/Army_of_the_Republic_of_Vietnam?oldid=495258787
5	http://en.wikipedia.org/wiki/Auburn,_California?oldid=492217341
6	http://en.wikipedia.org/wiki/Barbecue?oldid=494969208
7	http://en.wikipedia.org/wiki/Battle_of_Boulou?oldid=490141720
8	http://en.wikipedia.org/wiki/Battle_of_Camden?oldid=490509731
9	http://en.wikipedia.org/wiki/Battle_of_Fort_Hindman?oldid=495189359
10	http://en.wikipedia.org/wiki/Battle_of_Landen?oldid=476011327
11	http://en.wikipedia.org/wiki/Battle_of_Mount_Tumbledown?oldid=493723336
12	http://en.wikipedia.org/wiki/Battle_of_New_Orleans?oldid=494895989

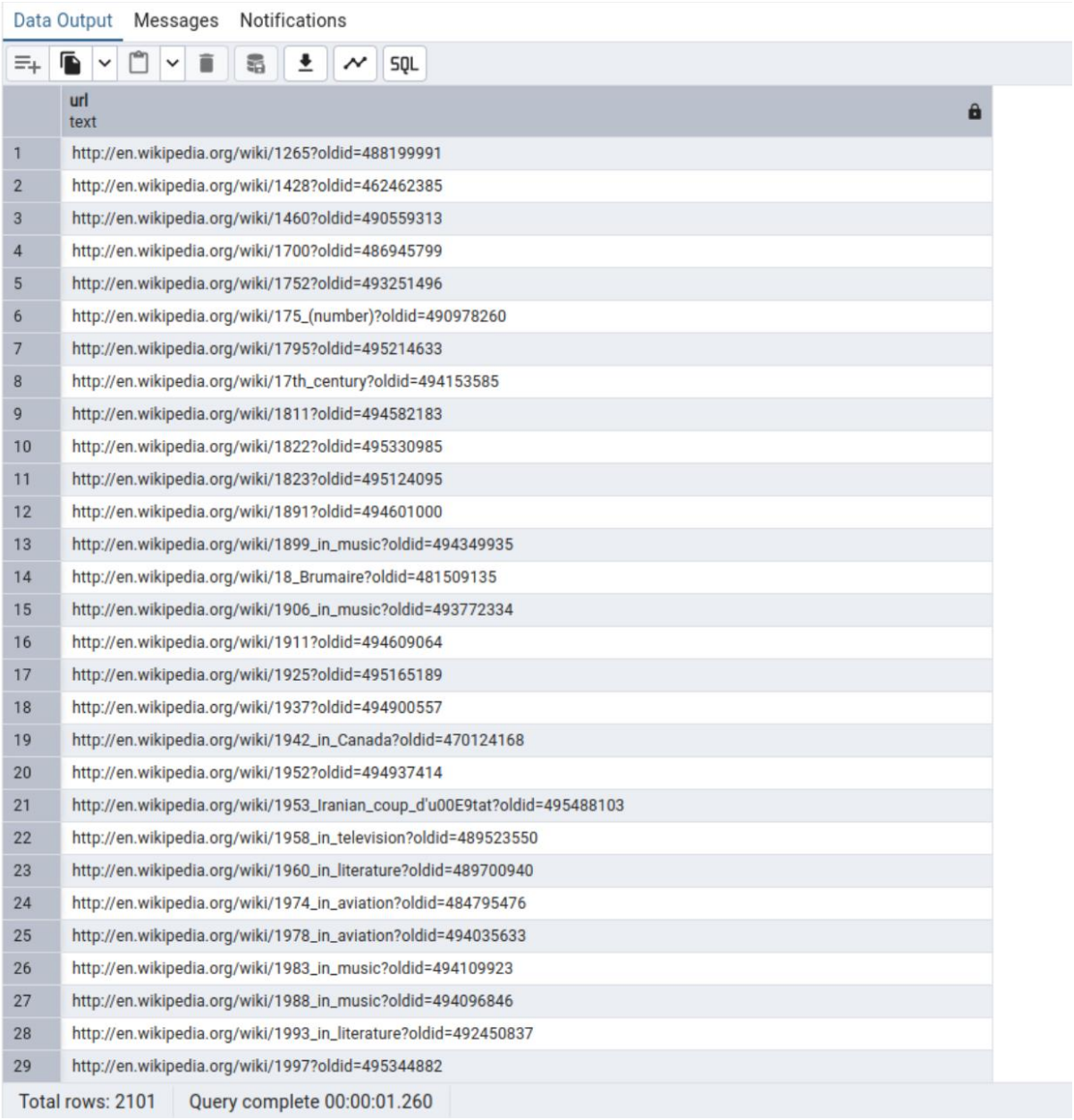
Total rows: 66 Query complete 00:00:02.476 LF

2.2 Query 2 - Using a similar line of approach(of filtering rows with any of the 4 given words) with frequency as above. Note that for 'any of' we just need to have the count of words for a particular article id to be greater than 0. This means that there are any of the following words present. The query for this is as follows:

Query Query History

```
1  SELECT t1.url
2  FROM assignment_1.revision_uri t1
3  JOIN assignment_1.keywords t2 ON t1."ID" = t2."ID"
4  WHERE t2."TERM" IN ('infantri', 'reinforc', 'brigad', 'fire')
5  GROUP BY t1.url
6  HAVING COUNT(DISTINCT t2."TERM") = 1;
7
```

And out put is this



Data Output | Messages | Notifications

SQL

	url text
1	http://en.wikipedia.org/wiki/1265?oldid=488199991
2	http://en.wikipedia.org/wiki/1428?oldid=462462385
3	http://en.wikipedia.org/wiki/1460?oldid=490559313
4	http://en.wikipedia.org/wiki/1700?oldid=486945799
5	http://en.wikipedia.org/wiki/1752?oldid=493251496
6	http://en.wikipedia.org/wiki/175_(number)?oldid=490978260
7	http://en.wikipedia.org/wiki/1795?oldid=495214633
8	http://en.wikipedia.org/wiki/17th_century?oldid=494153585
9	http://en.wikipedia.org/wiki/1811?oldid=494582183
10	http://en.wikipedia.org/wiki/1822?oldid=495330985
11	http://en.wikipedia.org/wiki/1823?oldid=495124095
12	http://en.wikipedia.org/wiki/1891?oldid=494601000
13	http://en.wikipedia.org/wiki/1899_in_music?oldid=494349935
14	http://en.wikipedia.org/wiki/18_Brumaire?oldid=481509135
15	http://en.wikipedia.org/wiki/1906_in_music?oldid=493772334
16	http://en.wikipedia.org/wiki/1911?oldid=494609064
17	http://en.wikipedia.org/wiki/1925?oldid=495165189
18	http://en.wikipedia.org/wiki/1937?oldid=494900557
19	http://en.wikipedia.org/wiki/1942_in_Canada?oldid=470124168
20	http://en.wikipedia.org/wiki/1952?oldid=494937414
21	http://en.wikipedia.org/wiki/1953_Iranian_coup_d'u00E9tat?oldid=495488103
22	http://en.wikipedia.org/wiki/1958_in_television?oldid=489523550
23	http://en.wikipedia.org/wiki/1960_in_literature?oldid=489700940
24	http://en.wikipedia.org/wiki/1974_in_aviation?oldid=484795476
25	http://en.wikipedia.org/wiki/1978_in_aviation?oldid=494035633
26	http://en.wikipedia.org/wiki/1983_in_music?oldid=494109923
27	http://en.wikipedia.org/wiki/1988_in_music?oldid=494096846
28	http://en.wikipedia.org/wiki/1993_in_literature?oldid=492450837
29	http://en.wikipedia.org/wiki/1997?oldid=495344882

Total rows: 2101 | Query complete 00:00:01.260

2.3 Query 3

Initially, we retrieve all records from the keywords table that include 'reinforc'. Simultaneously, we fetch rows from keywords containing any of the remaining three terms. Next, we eliminate those entries that share the same IDs as 'reinforc'. The outcome should yield records with article IDs that contain 'reinforc' but exclude the other three.

Query Query History

```
1  SELECT url as url
2  FROM assignment_1.revision_uri t0
3  WHERE t0."ID" IN (
4      SELECT t1."ID" as article_id
5      FROM assignment_1.keywords t1
6      WHERE t1."TERM" = 'reinforc'
7      AND t1."ID" NOT IN (
8          SELECT t2."ID" as article_id
9          FROM assignment_1.keywords  t2
10         WHERE t2."TERM" IN ('infantri', 'brigad', 'fire')
11     )
12 );
13
```

Data Output Messages Notifications

Showing rov

	url text
1	http://en.wikipedia.org/wiki/Millipede?oldid=494910496
2	http://en.wikipedia.org/wiki/Child_labour?oldid=495084550
3	http://en.wikipedia.org/wiki/Revolutionary_Girl_Utena?oldid=494097487
4	http://en.wikipedia.org/wiki/Castine,_Maine?oldid=495034474
5	http://en.wikipedia.org/wiki/Gambling?oldid=492205691
6	http://en.wikipedia.org/wiki/Acoustics?oldid=494462936
7	http://en.wikipedia.org/wiki/Politics_of_Greece?oldid=494977862
8	http://en.wikipedia.org/wiki/Anglicanism?oldid=495439272
9	http://en.wikipedia.org/wiki/Henry_VII_of_England?oldid=493305259

Total rows: 199 Query complete 00:00:02.832

2.4 Query 4 The process of fetching rows stays the same as in the first question. The query retrieves URLs from assignment_1.revision_uri by joining them with assignment_1.keywords on ID. It filters records containing all four specified keywords ('infantri', 'reinforc', 'brigad', 'fire'), sums their scores for each URL, and ranks them in descending order of total_score.

Welcome abhay_db/abhay@abhay_db*

abhay_db/abhay@abhay_db

No limit

Query Query History

```
1  SELECT t1.url, SUM(t2."SCORE") AS total_score
2  FROM assignment_1.revision_uri t1
3  JOIN assignment_1.keywords t2 ON t1."ID" = t2."ID"
4  WHERE t2."TERM" IN ('infantri', 'reinforc', 'brigad', 'fire')
5  GROUP BY t1.url
6  HAVING COUNT(DISTINCT t2."TERM") = 4
7  ORDER BY total_score DESC;
8
```

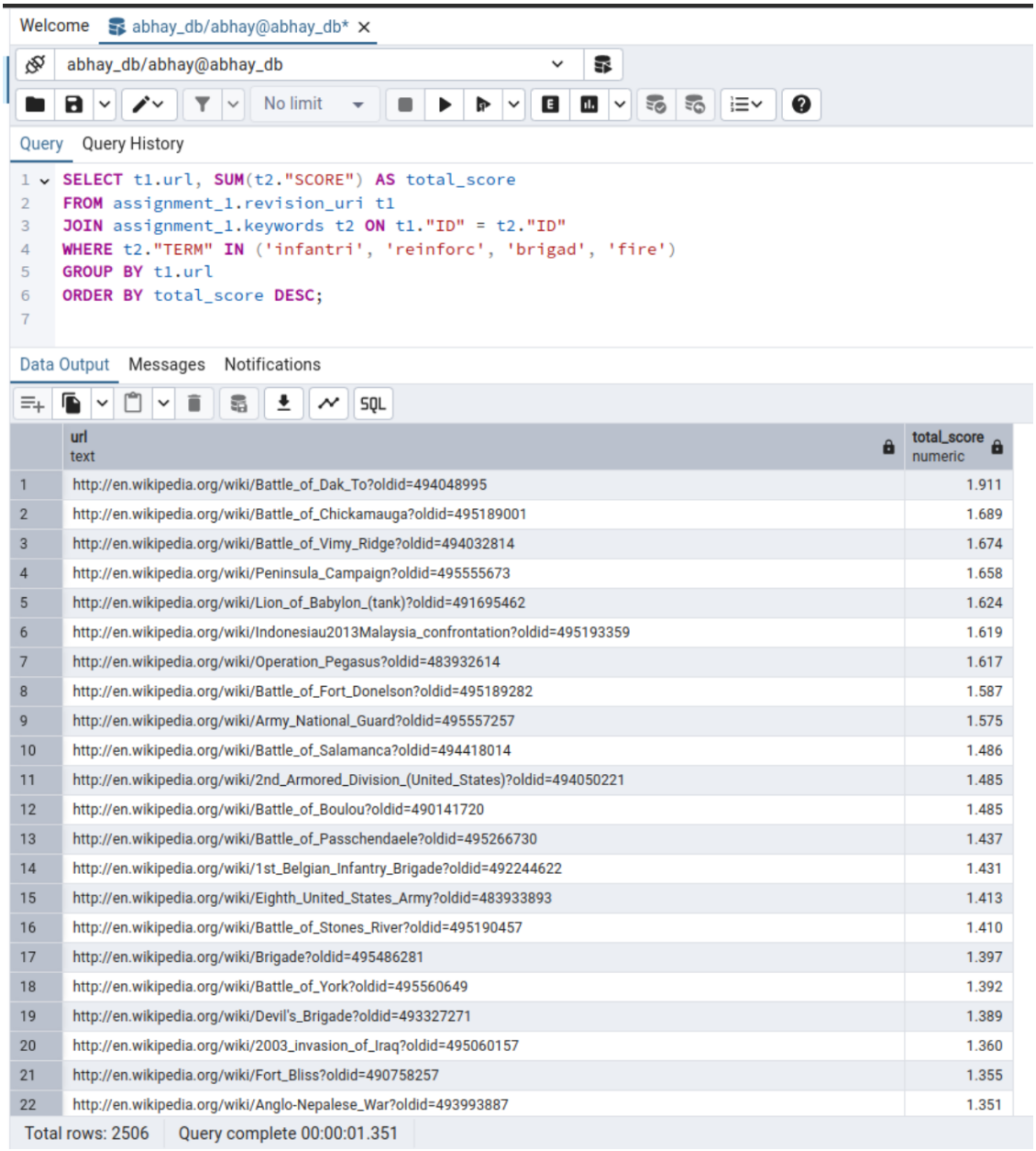
Data Output Messages Notifications

SQL

	url text	total_score numeric
1	http://en.wikipedia.org/wiki/Battle_of_Dak_To?oldid=494048995	1.911
2	http://en.wikipedia.org/wiki/Battle_of_Chickamauga?oldid=495189001	1.689
3	http://en.wikipedia.org/wiki/Battle_of_Vimy_Ridge?oldid=494032814	1.674
4	http://en.wikipedia.org/wiki/Peninsula_Campaign?oldid=495555673	1.658
5	http://en.wikipedia.org/wiki/Lion_of_Babylon_(tank)?oldid=491695462	1.624
6	http://en.wikipedia.org/wiki/Indonesiau2013Malaysia_confrontation?oldid=495193359	1.619
7	http://en.wikipedia.org/wiki/Operation_Pegasus?oldid=483932614	1.617
8	http://en.wikipedia.org/wiki/Battle_of_Fort_Donelson?oldid=495189282	1.587
9	http://en.wikipedia.org/wiki/Battle_of_Salamanca?oldid=494418014	1.486
10	http://en.wikipedia.org/wiki/Battle_of_Boulou?oldid=490141720	1.485
11	http://en.wikipedia.org/wiki/2nd_Armored_Division_(United_States)?oldid=494050221	1.485
12	http://en.wikipedia.org/wiki/Battle_of_Passchendaele?oldid=495266730	1.437
13	http://en.wikipedia.org/wiki/Battle_of_Stones_River?oldid=495190457	1.410
14	http://en.wikipedia.org/wiki/Battle_of_York?oldid=495560649	1.392
15	http://en.wikipedia.org/wiki/Devil's_Brigade?oldid=493327271	1.389
16	http://en.wikipedia.org/wiki/2003_invasion_of_Iraq?oldid=495060157	1.360
17	http://en.wikipedia.org/wiki/Anglo-Nepalese_War?oldid=493993887	1.351
18	http://en.wikipedia.org/wiki/British_Expeditionary_Force_(World_War_I)?oldid=483791980	1.350
19	http://en.wikipedia.org/wiki/United_States_invasion_of_Panama?oldid=494343919	1.338
20	http://en.wikipedia.org/wiki/Battle_of_Camden?oldid=490509731	1.332
21	http://en.wikipedia.org/wiki/Battle_of_Isandlwana?oldid=495315847	1.316
22	http://en.wikipedia.org/wiki/Luxembourg_Army?oldid=493333664	1.283

Total rows: 35 Query complete 00:00:01.286

2.5 Query 5 The logic for retrieving rows remains the same as in the second question. However, this query retrieves URLs from assignment_1.revision_uri by joining them with assignment_1.keywords on ID. It filters records containing any of the specified keywords ('infantri', 'reinforc', 'brigad', 'fire'), calculates the total score for each URL by summing keyword scores, and ranks them in descending order of total_score.



	url text	total_score numeric
1	http://en.wikipedia.org/wiki/Battle_of_Dak_To?oldid=494048995	1.911
2	http://en.wikipedia.org/wiki/Battle_of_Chickamauga?oldid=495189001	1.689
3	http://en.wikipedia.org/wiki/Battle_of_Vimy_Ridge?oldid=494032814	1.674
4	http://en.wikipedia.org/wiki/Peninsula_Campaign?oldid=495555673	1.658
5	http://en.wikipedia.org/wiki/Lion_of_Babylon_(tank)?oldid=491695462	1.624
6	http://en.wikipedia.org/wiki/Indonesiau2013Malaysia_confrontation?oldid=495193359	1.619
7	http://en.wikipedia.org/wiki/Operation_Pegasus?oldid=483932614	1.617
8	http://en.wikipedia.org/wiki/Battle_of_Fort_Donelson?oldid=495189282	1.587
9	http://en.wikipedia.org/wiki/Army_National_Guard?oldid=495557257	1.575
10	http://en.wikipedia.org/wiki/Battle_of_Salamanca?oldid=494418014	1.486
11	http://en.wikipedia.org/wiki/2nd_Armored_Division_(United_States)?oldid=494050221	1.485
12	http://en.wikipedia.org/wiki/Battle_of_Boulou?oldid=490141720	1.485
13	http://en.wikipedia.org/wiki/Battle_of_Passchendaele?oldid=495266730	1.437
14	http://en.wikipedia.org/wiki/1st_Belgian_Infantry_Brigade?oldid=492244622	1.431
15	http://en.wikipedia.org/wiki/Eighth_United_States_Army?oldid=483933893	1.413
16	http://en.wikipedia.org/wiki/Battle_of_Stones_River?oldid=495190457	1.410
17	http://en.wikipedia.org/wiki/Brigade?oldid=495486281	1.397
18	http://en.wikipedia.org/wiki/Battle_of_York?oldid=495560649	1.392
19	http://en.wikipedia.org/wiki/Devil's_Brigade?oldid=493327271	1.389
20	http://en.wikipedia.org/wiki/2003_invasion_of_Iraq?oldid=495060157	1.360
21	http://en.wikipedia.org/wiki/Fort_Bliss?oldid=490758257	1.355
22	http://en.wikipedia.org/wiki/Anglo-Nepalese_War?oldid=493993887	1.351

Total rows: 2506 Query complete 00:00:01.351

2.6 Query 6

First, we find articles that have the keyword 'reinforc' but do not contain 'fire', 'brigad', or 'infantri'. Once we have these articles, we rank them based on their scores. If the total score of 'fire', 'brigad', and 'infantri' is less than 'reinforc', then articles with a higher 'reinforc' score are ranked higher. To achieve this, we join the 'reinforc' scores with the total scores of the other three keywords. Articles are sorted first by total score (descending) and then by 'reinforc' score (descending). Finally, we join with the URL table to get the final list of ranked articles.

Query | Query History

```sql
SELECT t0."url" as url
FROM assignment_1.revision_uri t0
INNER JOIN (
    SELECT *, (temp1.sumone + temp2.sumfour) as f_score
    FROM (
        SELECT tk1."ID" as id1, SUM(tk1."SCORE") as sumone
        FROM assignment_1.keywords tk1
        WHERE tk1."TERM" = 'reinforc'
        GROUP BY id1
    ) as temp1
    LEFT JOIN (
        SELECT tk2."ID" as id2, COUNT(tk2."ID") as count, SUM(tk2."SCORE") as sumfour
        FROM assignment_1.keywords tk2
        WHERE tk2."TERM" IN ('brigad', 'fire', 'infantri')
        GROUP BY id2
    ) as temp2
    ON temp1.id1 = temp2.id2
    WHERE temp2.count IS NULL OR temp2.count != 3
) as comparision
ON t0."ID" = comparision.id1
ORDER BY comparision.f_score DESC, comparision.sumone DESC;

```

Data Output | Messages | Notifications

SQL | Showing rows: 1 to 37

	url text
1	http://en.wikipedia.org/wiki/Fallingwater?oldid=494470323
2	http://en.wikipedia.org/wiki/Rialto_Towers?oldid=493614024
3	http://en.wikipedia.org/wiki/Boat_building?oldid=495412452
4	http://en.wikipedia.org/wiki/Audio_multicore_cable?oldid=492290178
5	http://en.wikipedia.org/wiki/Eureka_Tower?oldid=492656657
6	http://en.wikipedia.org/wiki/Trial_and_error?oldid=492260541

Total rows: 376 | Query complete 00:00:01.467