



Soha Khalid

BWT – Data Engineering

Task 02 – Exercise

You will be using Rnacen (RNACentral) schema to find and explore the data available in all tables that will help you to recognise the potential tables you can query to answer the following questions.

- 1. Write a query to get data having length of Rna structures more than 12 with them being added after 2008.**

```
SELECT * FROM rna
WHERE len > 12
AND timestamp > '2008-12-31'
LIMIT 1000;
```

The screenshot shows a database interface with a query editor and a results table. The query editor contains the following SQL query:

```
1 SELECT * FROM rna
2 WHERE len > 12
3 AND timestamp > '2008-12-31'
4 LIMIT 1000;
```

The results table displays the following data:

	id	bigint	UPI	character varying (30)	timestamp	timestamp without time zone	userstamp	character varying (60)	crc64	character	len	integer	seq_short	character
1	13188704		URS0000C93E60		2018-03-12 14:27:09.711392		rnacen		08B8B144020A070C		375		GTAGTCC	
2	13188705		URS0000C93E61		2018-03-12 14:27:09.711445		rnacen		8F9F2FF5C64271C0		373		TAATACG	
3	13188706		URS0000C93E62		2018-03-12 14:27:09.711471		rnacen		FCDE58284F547928		934		COGGTAA	
4	13188707		URS0000C93E63		2018-03-12 14:27:09.711685		rnacen		67844C38AEFD207C		376		TAAGACG	
5	13188708		URS0000C93E64		2018-03-12 14:27:09.711719		rnacen		C6B543C4F5DC81DA		784		TCCGCGT	
6	13188709		URS0000C93E65		2018-03-12 14:27:09.711907		rnacen		347CE9C3CEDFED6D		2912		ATATAGG	

2. How many pre computed RNA are present that are still active and got their last release update before 2022.

```
SELECT COUNT (*)
FROM rnc_rna_precomputed
WHERE is_active = TRUE
AND last_release < '2022'
LIMIT 1000;
```

The screenshot shows a database interface with a table schema on the left and a SQL query editor on the right. The table schema for `rnc_rna_precomputed` lists 14 columns: `id`, `taxid`, `description`, `upi`, `rna_type`, `update_date`, `has_coordinates`, `databases`, `is_active` (highlighted), `last_release`, `short_description`, `so_rna_type`, `is_locus_representative`, and `assigned_so_rna_type`. The SQL query editor shows the following query:

```
SELECT COUNT (*)
FROM rnc_rna_precomputed
WHERE is_active = TRUE
AND last_release < '2022'
LIMIT 1000;
```

Below the query editor, the 'Data Output' tab is active, displaying a single row of results:

count
70886776

3. How many total pre computed RNA records for snoRNA and tRNA were recorded in 2011, 2016, 2014, and 2020.

```
SELECT *
FROM rnc_rna_precomputed
WHERE rna_type IN ('snoRNA', 'tRNA')
AND TO_CHAR(update_date, 'YYYY') IN ('2011', '2014', '2016', '2020')
LIMIT 1000;
```

```
SELECT COUNT(*)
FROM rnc_rna_precomputed
WHERE (rna_type IN ('snoRNA', 'tRNA')) AND EXTRACT(YEAR FROM update_date) IN
(2011, 2016, 2014, 2020)
LIMIT 1000;
```

Database interface showing a query on the `rnc_rna_precomputed` table. The query filters for `rna_type` in ('snoRNA', 'tRNA') and `update_date` in ('2011', '2014', '2016', '2020').

```

1 SELECT *
2 FROM rnc_rna_precomputed
3 WHERE rna_type IN ('snoRNA', 'tRNA')
4 AND TO_CHAR(update_date, 'YYYY') IN ('2011', '2014', '2016', '2020')
5 LIMIT 2000;

```

	rna_type	update_date	has_coordinates	databases	is_active	last_release	short_description
1	tRNA	2020-08-20	false		false	468	Generic tRNA
2	snoRNA	2020-08-14	true	Rfam	false	460	Small nucleolar RNA SNORD17
3	tRNA	2020-08-20	false	ENA	true	884	tRNA-Val
4	tRNA	2020-04-26	false		false	468	Generic tRNA
5	snoRNA	2020-05-03	true	Ensembl	false	550	Small nucleolar RNA SNORA70
6	tRNA	2020-08-20	false	ENA	true	884	tRNA-Thr-GGT
7	tRNA	2020-08-20	false	ENA	true	884	(soil metagenome) tRNA-Thr

Total rows: 2000 of 2000 Query complete 00:00:07.405 Ln 7, Col 1

Database interface showing a query on the `rnc_rna_precomputed` table. The query counts the number of records for `rna_type` in ('snoRNA', 'tRNA') and `update_date` in ('2011', '2016', '2014', '2020').

```

1
2
3 SELECT COUNT(*)
4 FROM rnc_rna_precomputed
5 WHERE (rna_type IN ('snoRNA', 'tRNA')) AND EXTRACT(YEAR FROM update_date) IN (2011, 2016, 2014, 2020)
6 LIMIT 1000;

```

	count
1	70872

4. Can you give me the names of all databases built for RNA with minimum length other than 100, 200, 300, 400, and 15.

SELECT display_name

FROM rnc_database

WHERE min_length NOT IN (100, 200, 300, 400, 15);

The screenshot shows a database management interface. On the left, a tree view displays the 'rnc_database' structure, including a list of 19 columns: id, timestamp, userstamp, descr, current_release, full_descr, alive, for_release, display_name, project_id, avg_length, min_length, max_length, num_sequences, num_organisms, description, url, example, and reference. The 'Columns (19)' folder is expanded, and the 'display_name' column is highlighted.

The main query editor shows the following SQL query:

```
SELECT display_name
FROM rnc_database
WHERE min_length NOT IN (100, 200, 300, 400, 15);
```

The 'Data Output' tab displays the results of the query. The first column is 'display_name', which is a character varying (60) type. The results are as follows:

display_name	
1	ENA
2	GENCODE
3	MGnify
4	GeneCards
5	RDP
6	snoRNA Database
7	Rfam

The status bar at the bottom indicates 'Total rows: 48 of 48' and 'Query complete 00:00:01.481'.

- Can you get complete 500 records of sequences for active regions and name your column as myregions in which you are getting the region name column value. Then tell me what different chromosomes with exon_count we have for regions including center, east and north using the name you set for your column.

```
SELECT region_name AS myregions
```

```
FROM rnc_sequence_regions
```

```
LIMIT 500;
```

For last part no result can be retrieved from data because no such value named “center, north and east” existed in the given database.

> rnc_sequence_features

> rnc_sequence_regions

> rnc_taxonomy

> temp_bad_is_active

> validate_layout_counts

> validate_layout_hits

> xref

> xref_not_unique

> xref_p1_deleted

> xref_p1_deleted_old

> xref_p1_not_deleted

> xref_p1_not_deleted_old

> xref_p2_deleted

> xref_p2_deleted_old

> xref_p2_not_deleted

> xref_p2_not_deleted_old

> xref_p3_deleted

> xref_p3_deleted_old

> xref_p3_not_deleted

> xref_p3_not_deleted_old

> xref_p4_deleted

> xref_p4_deleted_old

> xref_p4_not_deleted

> xref_p4_not_deleted_old

> xref_p5_deleted

> xref_p5_not_deleted

> xref_p6_deleted

pfmegrnargs/reader@soha

No limit

Query Query History

1 SELECT region_name AS myregions

2 FROM rnc_sequence_regions

3 LIMIT 500;

4

5

6

Data Output Messages Notifications

myregions

text

1 URS00006F9F83_10020@KN672353.1/2907709-2907773:-

2 URS00006F9F83_10020@KN672353.1/2908196-2908260:+

3 URS00006F9F83_10020@KN672353.1/3567704-3567768:-

4 URS00006F9F83_10020@KN672353.1/4727654-4727718:+

5 URS00006F9F83_10020@KN672353.1/5119365-5119429:+

6 URS00006F9F83_10020@KN672353.1/515849-515913:+

7 URS00006F9F83_10020@KN672353.1/6064424-6064488:-

Total rows: 500 of 500 Query complete 00:00:01.708