Bytewise Fellowship

Task 2: SQL Exercise - RNACentral Schema Queries

Name	Muhammad Ahsan Saleem
Date	16th June, 2024
Submitted to:	Muhammad Bilal

Introduction:

This document presents a series of SQL queries designed to explore and analyze data from the RNACentral schema. The goal is to address specific questions related to RNA structures, pre-computed RNA records, database characteristics, and sequence regions. By executing these queries, we aim to retrieve valuable insights and provide accurate numerical results, which are essential for understanding the data landscape and supporting data-driven decision-making processes. The results and methodologies are documented.

Task 02 - Exercise 🚨

You will be using RNACentral (RNACentral) schema to find and explore the data available in all tables that will help you recognize 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
2. How many pre-computed RNA are present that are still active and got their last release update before 2022
3. How many total pre-computed RNA records for snoRNA and tRNA were recorded in 2011, 2016, 2014, and 2020
4. Can you give me the names of all databases built for RNA with minimum length other than 100, 200, 300, 400, and 15
5. 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.

Way to Submit 🎖

You'll be creating a file with questions mentioned above, then you will paste the potential query you would have built for that, and answer the question numerically from the results your query would have given you. Also, you can attach a partial screenshot of your query output as well. You'll upload that file to your GitHub repository named <Your name>-Data Engineering-BWF.

Query Documentation

Question 1

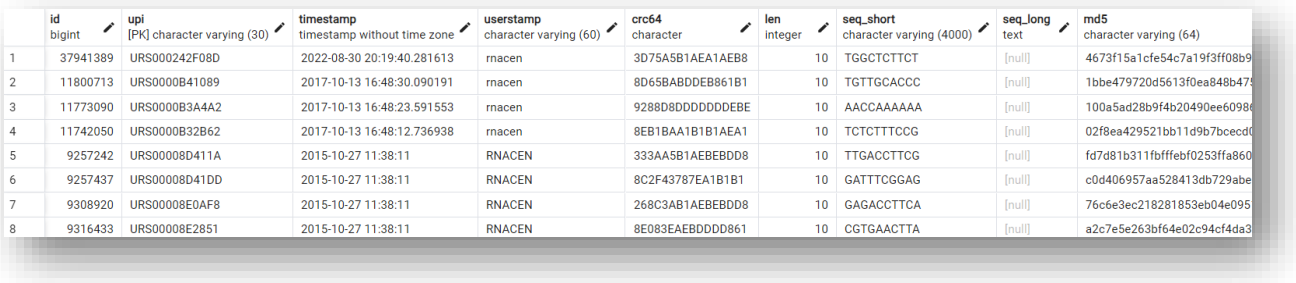
Statement: Write a query to get data having length of Rna structures more than 12 with them being added after 2008.

Query:

SELECT *
FROM rna
WHERE len > 12 AND EXTRACT(YEAR FROM timestamp) > 2008
LIMIT 1000;

```
1  SELECT *
2  FROM rna
3  WHERE len < 12 AND EXTRACT(YEAR FROM timestamp) > 2008
4  LIMIT 1000
```

Screenshot:



	id bigint	upi [PK] character varying (30)	timestamp timestamp without time zone	userstamp character varying (60)	crc64 character	len integer	seq_short character varying (4000)	seq_long text	md5 character varying (64)
1	37941389	URS000242F08D	2022-08-30 20:19:40.281613	rnacen	3D75A5B1AEA1AEB8	10	TGGCTCTTCT	[null]	4673f15a1cfe54c7a19f3ff08b9
2	11800713	URS0000B41089	2017-10-13 16:48:30.090191	rnacen	8D65BABDDEB861B1	10	TGTTGCACCC	[null]	1bbe479720d5613f0ea848b475
3	11773090	URS0000B3A4A2	2017-10-13 16:48:23.591553	rnacen	9288D8DDDDDDDEBE	10	AACCAAAAAA	[null]	100a5ad28b9f4b20490ee60986
4	11742050	URS0000B32B62	2017-10-13 16:48:12.736938	rnacen	8EB1BAA1B1B1AEA1	10	TCTCTTTCCG	[null]	02f8ea429521bb11d9b7bcecd0
5	9257242	URS00008D411A	2015-10-27 11:38:11	RNACEN	333AA5B1AEBEBDD8	10	TTGACCTTCG	[null]	fd7d81b311fbfffebf0253ffa860
6	9257437	URS00008D41DD	2015-10-27 11:38:11	RNACEN	8C2F43787EA1B1B1	10	GATTTCGGAG	[null]	c0d406957aa528413db729abe
7	9308920	URS00008E0AF8	2015-10-27 11:38:11	RNACEN	268C3AB1AEBEBDD8	10	GAGACCTTCA	[null]	76c6e3ec218281853eb04e095
8	9316433	URS00008E2851	2015-10-27 11:38:11	RNACEN	8E083EAEBDDDD861	10	CGTGAACTTA	[null]	a2c7e5e263bf64e02c94cf4da3

Question 2

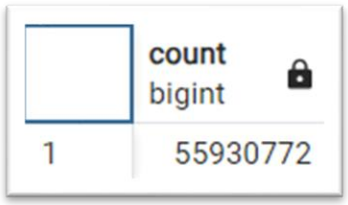
Statement: How many pre-computed RNA are present that are still active and got their last release update before 2022?

Query:

SELECT COUNT(*)
FROM rnc_rna_precomputed
WHERE is_active = TRUE AND EXTRACT(YEAR FROM update_date) < 2022;

```sql
SELECT COUNT(*)
FROM rnc_rna_precomputed
WHERE is_active = TRUE AND EXTRACT(YEAR FROM update_date) < 2022
```

Screenshot:



	count bigint
1	55930772

Question 3

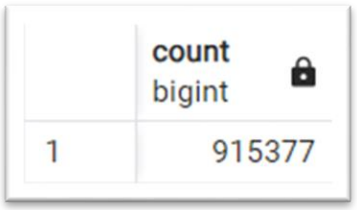
Statement: How many total pre-computed RNA records for snoRNA and tRNA were recorded in 2011, 2016, 2014, and 2020?

Query:

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

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

Screenshot:



	count bigint
1	915377

Question 4

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

Query:

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

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

Screenshot:

	display_name character varying (60)	min_length bigint
1	ENA	10
2	GENCODE	32
3	MGnify	27
4	GeneCards	16
5	RDP	1337
6	snoRNA Database	45
7	Rfam	24

Question 5

Statement: 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.

Query:

SELECT DISTINCT rnc_sequence_regions.id, rnc_rna_precomputed.id, rnc_sequence_regions.region_name AS myregion, rnc_sequence_regions.exon_count
FROM rnc_rna_precomputed
INNER JOIN
rnc_sequence_regions ON rnc_rna_precomputed.id = rnc_sequence_regions.urs_taxid
WHERE is_active = TRUE
LIMIT 500;

```
44  SELECT DISTINCT rnc_sequence_regions.id, rnc_rna_precomputed.id, rnc_sequence_regions.region_name AS myregion, rnc_sequence_regions.exon_count
45  FROM rnc_rna_precomputed
46  INNER JOIN
47  rnc_sequence_regions ON rnc_rna_precomputed.id = rnc_sequence_regions.urs_taxid
48  WHERE is_active = TRUE
49  LIMIT 500;
```

Screenshot:

`

	id integer	id character varying (44)	myregion text	exon_count integer
1	180828	URS00000003C7_4896	URS00000003C7_4896@I/2362324-2362771:+	1
2	181644	URS000000086E_80884	URS000000086E_80884@CACQ02003602/5181-5237:+	1
3	181645	URS000000086E_80884	URS000000086E_80884@CACQ02008496/758-814:+	1
4	181646	URS000000086E_80884	URS000000086E_80884@CACQ02008999/276-332:-	1
5	182352	URS0000001578_403677	URS0000001578_403677@supercont1.68/74803-74863:+	1
6	182936	URS0000001C12_332648	URS0000001C12_332648@10/2349948-2350037:+	2
7	182939	URS0000001C12_332648	URS0000001C12_332648@5/2159942-2160020:-	2
8	182940	URS0000001C12_332648	URS0000001C12_332648@8/1272226-1272302:-	2

Explanation:

- Due to issues with region names not matching those mentioned in the question and performance problems when using DISTINCT on exon_count, an alternative logical query was used.

Additional Note:

- For the fifth query, the exact region names given in the problem statement did not match the available data. Additionally, using DISTINCT for exon_count caused significant performance issues, leading to a long execution time without results. The query provided above is an logical alternative. DISTINCT has been used with all the queries.

- **The query with DISTINCT on exon_count is as follows:**

```
WITH distinct_exons AS (
SELECT DISTINCT exon_count
FROM rnc_sequence_regions AS rsr
JOIN rnc_rna_precomputed AS rrp ON rsr.urs_taxid = rrp.id
WHERE rrp.is_active = TRUE
LIMIT 500
)
SELECT
rsr.id,
rrp.id,
rsr.region_name AS myregion,
rsr.exon_count
FROM
rnc_sequence_regions AS rsr
JOIN
rnc_rna_precomputed AS rrp ON rsr.urs_taxid = rrp.id
JOIN
distinct_exons AS de ON rsr.exon_count = de.exon_count;
```