Modelagem e Armazenamento da tabela de posts de um Fórum universitário no cassandra

Trabalho 3 de Banco de Dados para Ciência de Dados

Pedro Augusto Benevides Salviano - 790983

Contextualização

O presente trabalho se propõe a apresentar a criação de uma tabela inserção de valores e consultas sobre os mesmos, além da discussão explicando as decisões do modelo.

Dados

Os dados foram criados aleatoriamente com ajuda do ChatGPT.

Script

O script .cql pode ser encontrado em https://github.com/pedro-salviano/uni-forum-cql

```
Unset
//Criar keyspace para desenvolvimento do trabalho (equivalente à database de bancos relacionais)
CREATE KEYSPACE trab3 WITH replication = {'class': 'SimpleStrategy', 'replication_factor': 1};

// Ativar/usar keyspace criado
USE trab3;

// Criar tabela post para registro das mensagens postadas
CREATE TABLE post(
       message_id UUID,
       user_id UUID,
       user_name TEXT,
       user_age INT,
       topic TEXT,
       body TEXT,
       timestamp TIMESTAMP,
       PRIMARY KEY(topic, user_id, timestamp)
);
```

```
// Popular tabela
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 123e4567-e89b-12d3-a456-426614174000, 'alice_smith', 22,
'Matemática', 'Preciso de ajuda com cálculo diferencial.',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 987f1234-e89b-12d3-a456-426614174111, 'bob_jones', 24,
'Matemática', 'Eu posso ajudar! Qual é a sua dúvida?', toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 123e4567-e89b-12d3-a456-426614174000, 'alice_smith', 22,
'Matemática', 'Estou com dificuldade em integrais triplas.',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 654d7890-e89b-12d3-a456-426614174222, 'carol_lee', 21,
'Programação', 'Dicas para aprender Python mais rápido?', '2025-02-23
13:56:50.664000+0000');
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 456a7891-e89b-12d3-a456-426614174333, 'dave_miller', 23,
'História', 'Recomendações de livros sobre Roma Antiga.',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 789b1234-e89b-12d3-a456-426614174444, 'eva_clark', 20,
'Filosofia', 'Discussão sobre ética kantiana.', toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 321c4567-e89b-12d3-a456-426614174555, 'frank_wright', 25,
'Engenharia', 'Dúvidas sobre circuitos elétricos.', toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 963f8527-e89b-12d3-a456-426614174666, 'grace_davis', 22,
'Economia', 'Qual livro introdutório recomendam?', toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 741d2589-e89b-12d3-a456-426614174777, 'harry_wilson', 23,
'Literatura', 'Análise de "1984" de George Orwell.', toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 852e3694-e89b-12d3-a456-426614174888, 'irene_lopez', 21,
'Química', 'Experimentos simples para entender equilíbrio químico.',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
```

```
      (uuid(), 159a7532-e89b-12d3-a456-426614174999, 'jack_martin', 24,
'Administração', 'Gestão de tempo: como melhorar?', toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 8ea295d7-5cad-4bf5-a1d9-2a70a8bf7d02, 'zack_young', 26,
'Psicologia', 'A influência do meio ambiente no comportamento humano.',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 654d7890-e89b-12d3-a456-426614174222, 'carol_lee', 21,
'Programação', 'Quais são os melhores cursos gratuitos de Python?',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 741d2589-e89b-12d3-a456-426614174777, 'harry_wilson', 23,
'Literatura', 'Livros clássicos que todo mundo deveria ler?',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 456a7891-e89b-12d3-a456-426614174333, 'dave_miller', 23,
'História', 'Influência do Império Romano na atualidade.',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 963f8527-e89b-12d3-a456-426614174666, 'grace_davis', 22,
'Economia', 'Como funciona a inflação?', toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 321c4567-e89b-12d3-a456-426614174555, 'frank_wright', 25,
'Engenharia', 'Alternativas ao uso de silício na computação.',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 159a7532-e89b-12d3-a456-426614174999, 'jack_martin', 24,
'Administração', 'Como ser mais produtivo no trabalho remoto?',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 852e3694-e89b-12d3-a456-426614174888, 'irene_lopez', 21,
'Química', 'Diferença entre ligação covalente e iônica.',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 8ea295d7-5cad-4bf5-a1d9-2a70a8bf7d02, 'zack_young', 26,
'Psicologia', 'Como a psicologia cognitiva influencia a IA?',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
```

```
      (uuid(), 654d7890-e89b-12d3-a456-426614174222, 'carol_lee', 21,
'Programação', 'Estruturas de dados mais usadas em desenvolvimento.',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 789b1234-e89b-12d3-a456-426614174444, 'eva_clark', 20,
'Filosofia', 'A relação entre existencialismo e absurdismo.',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 963f8527-e89b-12d3-a456-426614174666, 'grace_davis', 22,
'Economia', 'Impacto da taxa Selic na economia brasileira.',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 741d2589-e89b-12d3-a456-426614174777, 'harry_wilson', 23,
'Literatura', 'Escritores contemporâneos promissores.', toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 456a7891-e89b-12d3-a456-426614174333, 'dave_miller', 23,
'História', 'Como foi a transição do feudalismo para o capitalismo?',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 852e3694-e89b-12d3-a456-426614174888, 'irene_lopez', 21,
'Química', 'O que são catalisadores e como funcionam?', toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 321c4567-e89b-12d3-a456-426614174555, 'frank_wright', 25,
'Engenharia', 'Quais são os desafios da computação quântica?',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 159a7532-e89b-12d3-a456-426614174999, 'jack_martin', 24,
'Administração', 'Como liderar equipes de forma eficaz?',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 8ea295d7-5cad-4bf5-a1d9-2a70a8bf7d02, 'zack_young', 26,
'Psicologia', 'Teorias da motivação e sua aplicação no dia a dia.',
toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
      (uuid(), 987f1234-e89b-12d3-a456-426614174111, 'bob_jones', 24,
'Matemática', 'Como usar transformadas de Fourier?', toTimestamp(now()));
INSERT INTO post (message_id, user_id, user_name, user_age, topic, body,
timestamp) VALUES
```

```
      (uuid(), 123e4567-e89b-12d3-a456-426614174000, 'alice_smith', 22,
'Programação', 'Hey Carol, para aprender python recomendo o livro Entendendo
Algoritmos do Bhargava.', toTimestamp(now()));

// Procurar mensagem específica postada, usando a definição de consulta
esperada do trabalho ("As mensagens devem ser consultadas rapidamente por
usuário e data."), considerando o particionamento por tópico;
SELECT * FROM post WHERE topic = 'Programação' AND user_id =
654d7890-e89b-12d3-a456-426614174222 AND timestamp = '2025-02-23
13:56:50.664000+0000';

// Criar indíce e fazer consulta com agregação por usuário e tópico, para
obter frequência das mensagens
SELECT user_id, user_name, topic, COUNT(*) as qtd_mensagens_no_topico FROM
post GROUP BY topic, user_id;
```

Modelagem

Os requisitos exigiam:

- Cada mensagem deve ter um ID.
- Deve armazenar o ID do usuário que postou a mensagem.
- Deve-se armazenar a idade do usuário que postou a mensagem (por questões de temas sensíveis)
- Deve-se armazenar o tema da mensagem (“política”, “saúde”, “tecnologia”...)
- Deve conter o texto da mensagem.
- Deve-se registrar a data e hora da postagem.
- As mensagens devem ser consultadas rapidamente por usuário e data.

Dessa forma a tabela foi definida como segue.

Unset

```
CREATE TABLE post(
      message_id UUID,
      user_id UUID,
      user_name TEXT,
      user_age INT,
      topic TEXT,
      body TEXT,
      timestamp TIMESTAMP,
      PRIMARY KEY(topic, user_id, timestamp)
);
```

Note que além do user_id e data, foi incluído como chave de partição o tópico, dessa forma o acesso às mensagens pode ser segregado por tópico, permitindo a consulta de

todas mensagens de um tópico, a ordenação por sua vez idealmente seria feito primeiro pelo timestamp e somente depois pelo user_id, no entanto se feito dessa forma não seria possível realizar a consulta da frequência de mensagens por tópico de cada usuário, sem o uso de materialized views (recurso experimental e desabilitado por padrão no cassandra, na versão utilizada para o desenvolvimento).

Da forma implementada há um overhead de ordenação dos resultados como desejado fora do Cassandra, para exibir as mensagens na ordem de publicação.

A primeira consulta solicitada no trabalho é encontrar uma mensagem específica de um usuário, para ser realizado é necessário saber todos elementos da chave primária, para que não seja feito scan de todos resultados obtidos a partir da consulta dos elementos da chave fornecidos na query, como segue o exemplo:

```
Unset
SELECT
       *
FROM
       post
WHERE
       topic = 'Programação'
       AND
       user_id = 654d7890-e89b-12d3-a456-426614174222
       AND
       timestamp = '2025-02-23 13:56:50.664000+0000'
;
```

A última consulta solicitada foi a que justificou a decisão de ordenar primeiro pelo user_id, “recuperar, para cada usuário, os tipos de mensagens enviadas e a frequência delas”. Ao particionar por topic e ordenar primeiramente pelo user_id podemos utilizar as duas componentes da chave no group by garantindo o resultado correto na busca.

Caso timestamp fosse a primeira chave de ordenação, seria necessário incluí-la no group by, causando com que cada mensagem gerasse uma linha no resultado o que não é o desejado, e caso fosse ocultada do group by não seria possível incluir o usuário no group by, causando com que mensagens de usuários diferentes fossem agregadas na mesma linha do resultado, e o usuário exibido fosse o primeiro encontrado.

```
Unset
SELECT
       user_id,
       user_name,
       topic,
       COUNT(*) as qtd_mensagens_no_topico
FROM
       post
GROUP BY topic, user_id;
```

```
cqlsh:trab3> SELECT user_id, user_name, topic, COUNT(*) as qtd_mensagens_no_topico FROM post GROUP BY topic, user_id;
InvalidRequest: Error from server: code=2200 [Invalid query] message="Group by currently only support groups of columns following their declared order in the PRIMARY KEY"
cqlsh:trab3> SELECT user_id, user_name, topic, COUNT(*) as qtd_mensagens_no_topico FROM post GROUP BY topic, timestamp, user_id;

 user_id                              | user_name    | topic        | qtd_mensagens_no_topico
--------------------------------------+--------------+--------------+-------------------------
 741d2589-e89b-12d3-a456-426614174777 | harry_wilson |   Literatura |                       1
 741d2589-e89b-12d3-a456-426614174777 | harry_wilson |   Literatura |                       1
 741d2589-e89b-12d3-a456-426614174777 | harry_wilson |   Literatura |                       1
 321c4567-e89b-12d3-a456-426614174555 | frank_wright |   Engenharia |                       1
 321c4567-e89b-12d3-a456-426614174555 | frank_wright |   Engenharia |                       1
 321c4567-e89b-12d3-a456-426614174555 | frank_wright |   Engenharia |                       1
 123e4567-e89b-12d3-a456-426614174000 |  alice_smith |   Matemática |                       1
 987f1234-e89b-12d3-a456-426614174111 |    bob_jones |   Matemática |                       1
 123e4567-e89b-12d3-a456-426614174000 |  alice_smith |   Matemática |                       1
 987f1234-e89b-12d3-a456-426614174111 |    bob_jones |   Matemática |                       1
 654d7890-e89b-12d3-a456-426614174222 |    carol_lee |  Programação |                       1
 654d7890-e89b-12d3-a456-426614174222 |    carol_lee |  Programação |                       1
 654d7890-e89b-12d3-a456-426614174222 |    carol_lee |  Programação |                       1
 123e4567-e89b-12d3-a456-426614174000 |  alice_smith |  Programação |                       1
 789b1234-e89b-12d3-a456-426614174444 |    eva_clark |    Filosofia |                       1
 789b1234-e89b-12d3-a456-426614174444 |    eva_clark |    Filosofia |                       1
 159a7532-e89b-12d3-a456-426614174999 |  jack_martin | Administração |                       1
 159a7532-e89b-12d3-a456-426614174999 |  jack_martin | Administração |                       1
 159a7532-e89b-12d3-a456-426614174999 |  jack_martin | Administração |                       1
 456a7891-e89b-12d3-a456-426614174333 |  dave_miller |     História |                       1
 456a7891-e89b-12d3-a456-426614174333 |  dave_miller |     História |                       1
 456a7891-e89b-12d3-a456-426614174333 |  dave_miller |     História |                       1
 963f8527-e89b-12d3-a456-426614174666 |  grace_davis |     Economia |                       1
 963f8527-e89b-12d3-a456-426614174666 |  grace_davis |     Economia |                       1
 963f8527-e89b-12d3-a456-426614174666 |  grace_davis |     Economia |                       1
 8ea295d7-5cad-4bf5-a1d9-2a70a8bf7d02 |   zack_young |   Psicologia |                       1
 8ea295d7-5cad-4bf5-a1d9-2a70a8bf7d02 |   zack_young |   Psicologia |                       1
 8ea295d7-5cad-4bf5-a1d9-2a70a8bf7d02 |   zack_young |   Psicologia |                       1
 852e3694-e89b-12d3-a456-426614174888 |  irene_lopez |      Química |                       1
 852e3694-e89b-12d3-a456-426614174888 |  irene_lopez |      Química |                       1
 852e3694-e89b-12d3-a456-426614174888 |  irene_lopez |      Química |                       1

(31 rows)
```

Exemplo timestamp é a primeira chave de ordenação e foi incluída no group by, com modelagem descartada.

```
cqlsh:trab3> SELECT user_id, user_name, topic, COUNT(*) as qtd_mensagens_no_topico FROM post GROUP BY topic;

 user_id                              | user_name    | topic        | qtd_mensagens_no_topico
--------------------------------------+--------------+--------------+-------------------------
 741d2589-e89b-12d3-a456-426614174777 | harry_wilson |   Literatura |                       3
 321c4567-e89b-12d3-a456-426614174555 | frank_wright |   Engenharia |                       3
 123e4567-e89b-12d3-a456-426614174000 |  alice_smith |   Matemática |                       4
 654d7890-e89b-12d3-a456-426614174222 |    carol_lee |  Programação |                       4
 789b1234-e89b-12d3-a456-426614174444 |    eva_clark |    Filosofia |                       2
 159a7532-e89b-12d3-a456-426614174999 |  jack_martin | Administração |                       3
 456a7891-e89b-12d3-a456-426614174333 |  dave_miller |     História |                       3
 963f8527-e89b-12d3-a456-426614174666 |  grace_davis |     Economia |                       3
 8ea295d7-5cad-4bf5-a1d9-2a70a8bf7d02 |   zack_young |   Psicologia |                       3
 852e3694-e89b-12d3-a456-426614174888 |  irene_lopez |      Química |                       3

(10 rows)
```

Exemplo timestamp omitido no group by, com modelagem descartada

```
cqlsh:trab3> SELECT user_id, user_name, topic, COUNT(*) as qtd_mensagens_no_topico FROM post GROUP BY topic, user_id;

 user_id                              | user_name    | topic        | qtd_mensagens_no_topico
--------------------------------------+--------------+--------------+-------------------------
 741d2589-e89b-12d3-a456-426614174777 | harry_wilson |   Literatura |                       3
 321c4567-e89b-12d3-a456-426614174555 | frank_wright |   Engenharia |                       3
 123e4567-e89b-12d3-a456-426614174000 |  alice_smith |   Matemática |                       2
 987f1234-e89b-12d3-a456-426614174111 |    bob_jones |   Matemática |                       2
 123e4567-e89b-12d3-a456-426614174000 |  alice_smith |  Programação |                       1
 654d7890-e89b-12d3-a456-426614174222 |    carol_lee |  Programação |                       3
 789b1234-e89b-12d3-a456-426614174444 |    eva_clark |    Filosofia |                       2
 159a7532-e89b-12d3-a456-426614174999 |  jack_martin | Administração |                       3
 456a7891-e89b-12d3-a456-426614174333 |  dave_miller |     História |                       3
 963f8527-e89b-12d3-a456-426614174666 |  grace_davis |     Economia |                       3
 8ea295d7-5cad-4bf5-a1d9-2a70a8bf7d02 |   zack_young |   Psicologia |                       3
 852e3694-e89b-12d3-a456-426614174888 |  irene_lopez |      Química |                       3

(12 rows)
```

Exemplo resultado correto esperado, com a modelagem implementada.

Na questão da última consulta também foi solicitada a criação de um índice, que eu optei por não criar, uma vez que todas consultas relevantes incluem somente os campos que já são parte da chave. Para efeito de demonstração irei criar um índice na idade, para que nos permite fazer consultas de igualdade da idade, sem necessidade de allow filtering e consequente risco, como segue:

```
Unset
CREATE INDEX idx_user_age ON post(user_age);
SELECT * FROM post WHERE user_age = 20;
```

Resultado: