Busca por Similaridade no CassandraDB

Antonio Mourão, Rafael Pasquini, Rodolfo Villaça, Lasaro Camargos

Motivação

- Busca por similaridade permite pesquisa em dados complexos
- Soluções de armazenamento mais usadas não suportam essa característica

Cassandra

- Altamente escalável
- Particionamento aleatório dos dados entre os nós
- Dados distribuídos uniformemente
- Nós ordenados de acordo com a sequência natural
- Não permite a busca por dados similares

Hamming DHT

- Tabela de espalhamento distribuído
- Permite a busca de conteúdos por similaridade
- Nós com identificadores aleatórios posicionados em um anel virtual
- Ordenados de acordo com a sequência do código de Gray (distância de Hamming igual a 1)

Hamming DHT (Código de Gray)

Decimal	Binario	Decimal	Gray Binario
0	000	0	000
1	001	1	001
2	010	3	011
3	011	2	010
4	100	6	110
5	101	7	111
6	110	5	101
7	111	4	100

Hamming DHT

- Alta probabilidade de dados similares serem colocados em um mesmo nó ou naqueles com quem está em contato direto
- Reduz a média do número de saltos na rede na busca por conteúdos similares
- Avaliada por meio de simulação

CasSIMdra (Cassandra + Hamming DHT)

- Implementar os conceitos da HammingDHT no Cassandra permitindo busca por similaridade
- Dados na forma de vetores
- Uso do RHH para a geração de identificadores
- Nós ordenados no anel seguindo a sequência do código de Gray
- Particionamento dos dados agrega dados similares próximos uns aos outros
- Uso do LSBF para busca dos dados
- A busca é feita com poucos saltos pela rede

CasSIMdra (Cassandra + Hamming DHT)

- Implementar os conceitos da HammingDHT no Cassandra permitindo busca por similaridade
- Dados na forma de vetores
- Uso do RHH para a geração de identificadores
- Nós ordenados no anel seguindo a sequência do código de Gray
- Particionamento dos dados agrega dados similares próximos uns aos outros
- Uso do LSBF para busca dos dados
- A busca é feita com poucos saltos pela rede

CasSIMdra (Cassandra + Hamming DHT)

- Implementar os conceitos da HammingDHT no Cassandra permitindo busca por similaridade
- Dados na forma de vetores
- Uso do RHH para a geração de identificadores
- Nós ordenados no anel seguindo a sequência do código de Gray
- Particionamento dos dados agrega dados similares próximos uns aos outros
- Uso do LSBF para busca dos dados
- A busca é feita com poucos saltos pela rede

LSH / RHH

- Método de redução de espaços com alta dimensionalidade para espaços de menor dimensão
- Mantém pontos próximos no espaço original também próximos no espaço reduzido
- RHH é uma família LSH
- Usadas para gerar identificadores mantendo a similaridade dos objetos

CasSIMdra (Cassandra + Hamming DHT)

- Implementar os conceitos da HammingDHT no Cassandra permitindo busca por similaridade
- Dados na forma de vetores
- Uso do RHH para a geração de identificadores
- Nós ordenados no anel seguindo a sequência do código de Gray
- Particionamento dos dados agrega dados similares próximos uns aos outros
- Uso do LSBF para busca dos dados
- A busca é feita com poucos saltos pela rede

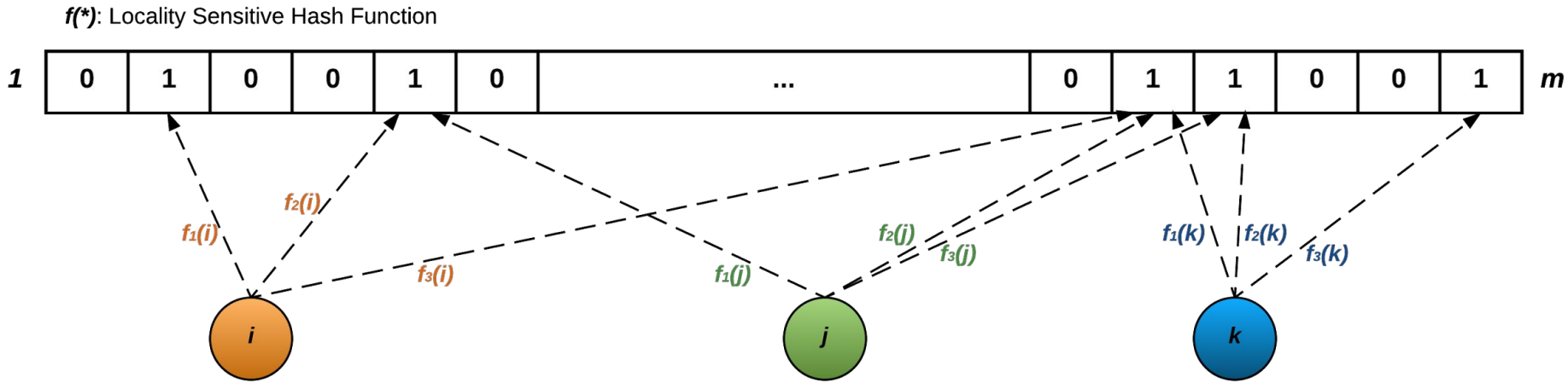
CasSIMdra (Cassandra + Hamming DHT)

- Implementar os conceitos da HammingDHT no Cassandra permitindo busca por similaridade
- Dados na forma de vetores
- Uso do RHH para a geração de identificadores
- Nós ordenados no anel seguindo a sequência do código de Gray
- Particionamento dos dados agrega dados similares próximos uns aos outros
- Uso do LSBF para busca dos dados
- A busca é feita com poucos saltos pela rede

CasSIMdra (Cassandra + Hamming DHT)

- Implementar os conceitos da HammingDHT no Cassandra permitindo busca por similaridade
- Dados na forma de vetores
- Uso do RHH para a geração de identificadores
- Nós ordenados no anel seguindo a sequência do código de Gray
- Particionamento dos dados agrega dados similares próximos uns aos outros
- Uso do LSBF para busca dos dados
- A busca é feita com poucos saltos pela rede

Locality Sensitive Bloom Filter (LSBF)

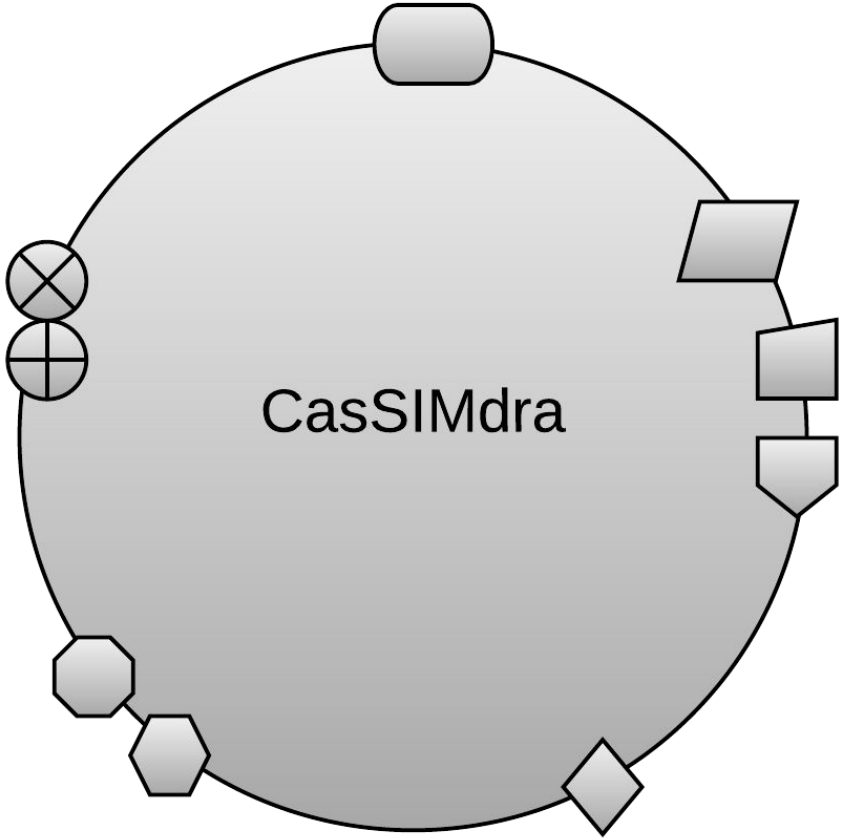


Locality Sensitive Bloom Filter (LSBF)

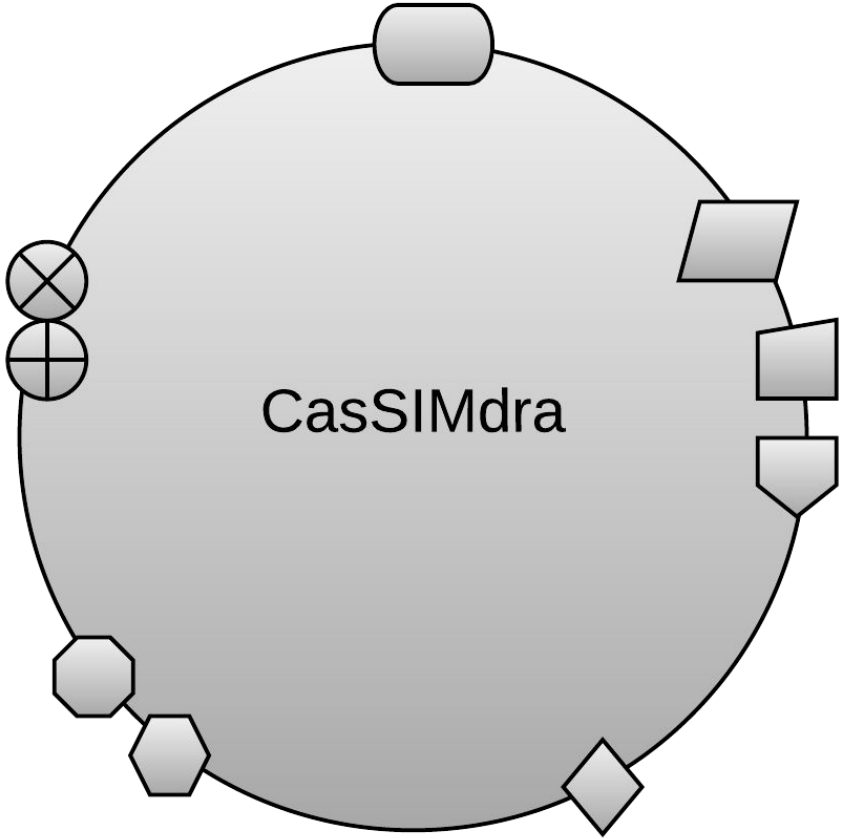
- Usa funções de espalhamento sensíveis à localidade
- Mapeia dados similares para os mesmos índices ou para índices vizinhos
- Verifica-se se algum dos bits na vizinhança do índice gerado, inclusive, é igual a 1
- Quanto maior a vizinhança considerada, menor a similaridade
- Sofrem de falsos negativos

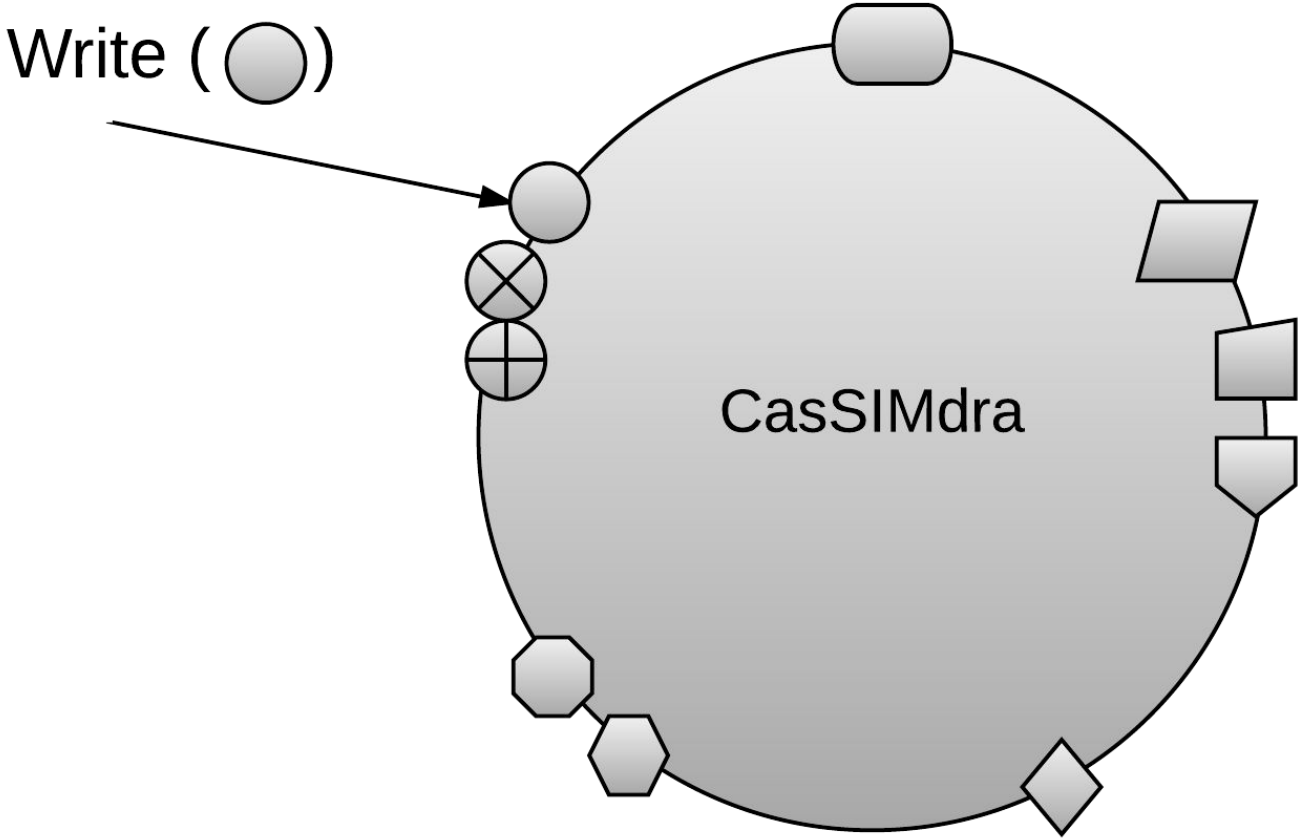
CasSIMdra (Cassandra + Hamming DHT)

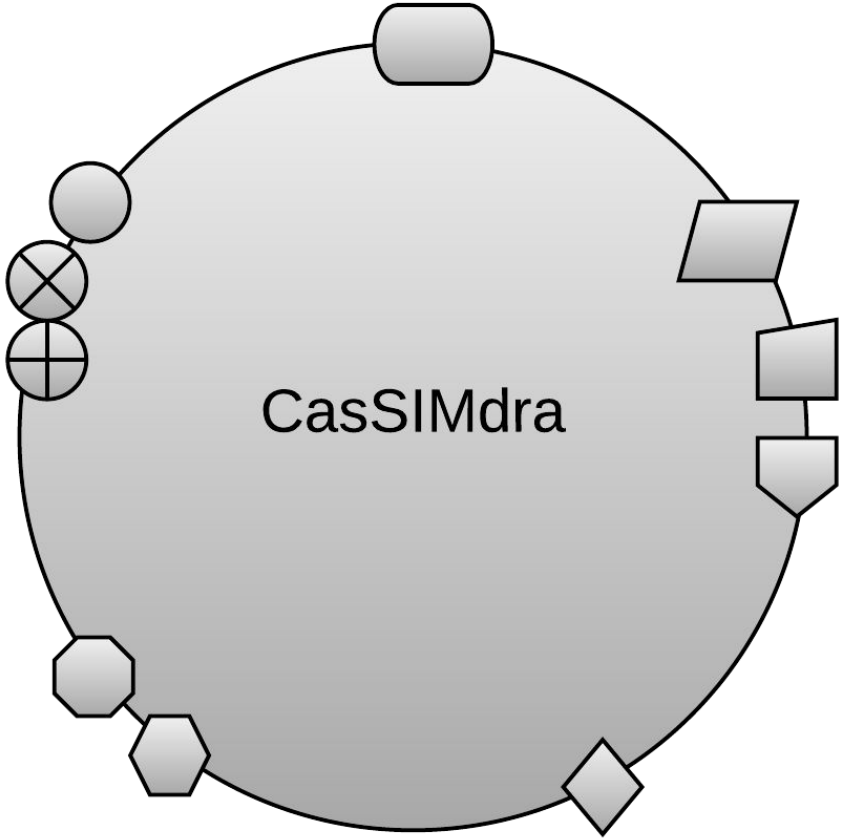
- Implementar os conceitos da HammingDHT no Cassandra permitindo busca por similaridade
- Dados na forma de vetores
- Uso do RHH para a geração de identificadores
- Nós ordenados no anel seguindo a sequência do código de Gray
- Particionamento dos dados agrega dados similares próximos uns aos outros
- Uso do LSBF para busca dos dados
- A busca é feita com poucos saltos pela rede


CasSIMdra

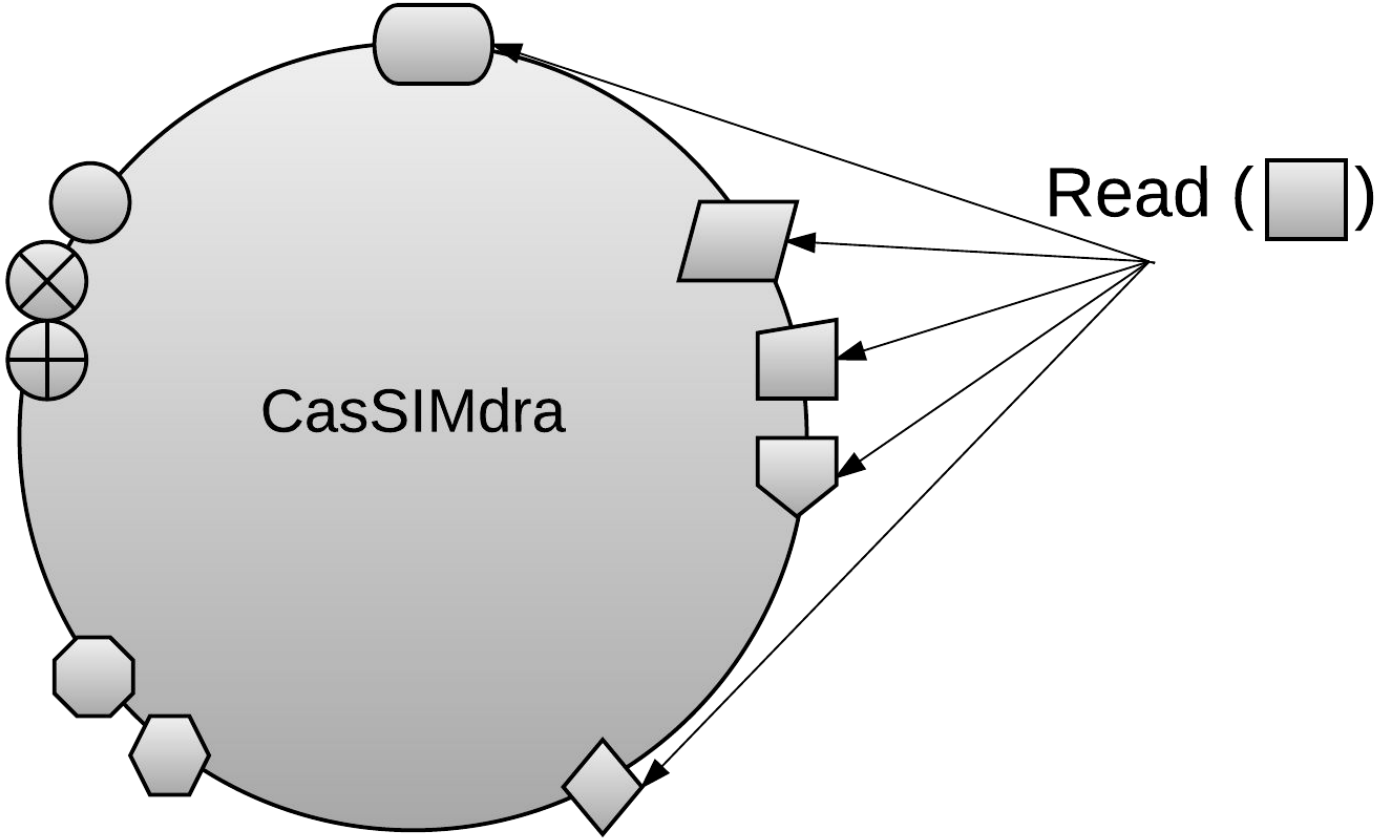
Write ( )




Write ( )
CasSIMdra



Read ( )


CasSIMdra
Read ( )

Dúvidas?

Leiam o artigo!!!

Seção de Poster

GitHub: https://github.com/pluxos/cassandra-sim

Contatos:

- antoniomourao@comp.ufu.br
- rafael.pasquini@ufu.br
- rodolfo.villaca@ufes.br
- lasaro@ufu.br