Proposta de TCC

Estudo do efeito de variações de Bloom filters no desempenho de consultas no banco de dados Apache Cassandra

Aluno: Matheus Barbosa Silva
Supervisores: Guilherme Oliveira Mota e Yoshiharu Kohayakawa

1 Introdução

A manipulação de dados em larga escala (*big data*) demanda o uso de ferramentas de armazenamento e consulta de dados mais ágeis que as convencionais e de alta disponibilidade. Nesse contexto, o sistema de banco de dados distribuído Apache Cassandra apresenta-se como uma das soluções mais amplamente utilizadas no mercado.

De modo a garantir maior eficiência e durabilidade no armazenamento de dados, arquivos do tipo **SSTable** são utilizados no armazenamento de dados de forma persistente e imutável. Assim como outras ferramentas de bancos de dados com foco em Big Data, o Apache Cassandra utiliza Bloom Filters (certas estruturas de dados probabilísticas) para armazenar e mapear um conjunto de SSTables com maior agilidade, mas com o custo de resultados falsos positivos.

2 Objetivos

A proposta de trabalho de conclusão de curso apresentada objetiva medir e analisar o impacto do uso de estruturas de dados alternativas fundamentadas em filtros de Bloom (como os *cuckoo filters*) no desempenho de consultas realizadas no software de banco de dados NoSQL Apache Cassandra. Para isso, deve-se descrever didaticamente o funcionamento de filtros de Bloom e *cuckoo filters*, assim como analisar os mecanismos e parâmetros que permitem mitigar as ocorrências de falsos positivos.

A exposição teórica dos filtros alternativos deve permitir a projeção de resultados de ganho de desempenho real nas consultas no Apache Cassandra. Assim, deve-se comparar o ganho de desempenho projetado na exposição teórica ao ganho real obtido nos experimentos com consultas em um banco de dados Apache Cassandra modificado para utilizar um filtro alternativo.

3 Cronograma

O desenvolvimento do projeto deve seguir o seguinte cronograma:

Atividade	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Pesquisa: Bloom filters	×							
Pesquisa: cuckoo hashing	×	×						
Pesquisa: cuckoo filters		×	×					
Pesquisa: Apache Cassandra			×	×				
Implementação: cuckoo filter no Apache Cassandra				×				
Análise de desempenho de consultas no Apache Cassandra				×	×	×		
Desenvolvimento da monografia		×	×	×	×	×	×	
Desenvolvimento de materiais para apresentação							×	×

4 Referências

[1] **Andrei e Michael(2004)** Andrei Broder e Michael Mitzenmacher. Network Applications of Bloom Filters: A Survey. Internet Mathematics, 1.

[2] **Bin Fan** *et. al*(**2014**) Bin Fan, David G. Andersen, Michael Kaminsky, Michael D. Mitzenmacher. Cuckoo Filter: Practically Better Than Bloom. CoNEXT'14.

[3] **Mitzenmacher e Eli(2005)** Michael Mitzenmacher, Eli Upfal. *Probability and Computing*. primeira edição.

[4] **Sharafat e Muhammad(2018)** Sharafat Ibn Mollah Mosharraf, Muhammad Abdullah Adnan. *Improving Query Execution Performance in Big Data using Cuckoo Filter*. 2018 IEEE International Conference on Big Data.