Detalhamento do Trabalho Final da Disciplina de Projeto de Arquivos

Filipe da Silva de Oliveira, Ricardo Stoklosa

25 de Novembro de 2019

1 Objetivo

Implementar um sistema gerenciador de dados com 10.000 registros de no mínimo 5 campos, utilizando indices de busca.

2 Dados Escolhidos

A base de dados utilizada no trabalho é composta pelos nomes populares dos bebês nascidos entre 2011 a 2016 na região de Nova York. Cada registro é composto pelos campos:

- Id
- Ano de nascimento
- Gênero
- Etnia
- Nome
- Quantidade de bebês que possuem este nome
- Ranking de popularidade deste nome por ano

A base foi obtida no site de dados públicos da cidade de Nova York, e foi manipulada de tal forma que tivesse 10.000 apenas registros.

Com esses dados é possível consultar qual o nome mais utilizado em determinado ano, qual a posição do ranking daquele ano este nome está, entre outras curiosidades.

3 Etapas do Sistema

A construção do sistema foi desenvolvida em três etapas, a primeira foi a geração do arquivo binário através do *csv*. A segunda foi a manipulação desse arquivo binário, utilizando o esquema de remoção lógica e fisica similar a ATIVIDADE-04 desenvolvida durante as aulas da disciplina, foram adotadas duas *structs*, uma chamada cabeçalho indicando o RRN do registro que foi logicamente removido e outra para indicar qual o proximo RRN que deverá assumir o cabeçalho do arquivo. Sendo assim quando inserido um novo registro ele ocupava a posição de um registro que foi removido pelo método lógico ou ele era adicionado ao final do arquivo, além disso é possível limpar o arquivo dos registros lógicos.

A terceira etapa do sistema é onde foi desenvolvido a Arvore B, infelizmente não conseguimos implementar a Arvore B de tal forma que apenas um nó estivesse em memória principal, logo a Arvore B inteira ocupa a memória principal, porém se pararmos para calcular, cada registro ocupa 224 bytes, 224*10000 obtemos 2MB de informação o que para a capacidade das memórias rams de hoje em dia é um volume muito pequeno de dados. Então tanto as buscas na Arvore B e as buscas sequenciais foram desenvolvidas de tal forma que todos os registros estivessem na memória principal.

4 Indices

Dois indices foram construídos, um para o campo Id e o outro para o campo composto por Ano de Nascimento e Ranking, assim é possível descobrir qual o nome de determinada posição em determinado ano.

5 Resultados Obtidos

Para a comparação das buscas entre a Arvore B utilizando os indices, foi desenvolvido uma busca sequencial sem o uso de índices, e o sistema calculava o tempo em nanosegundos da execução dos dois algoritmos. A média foi calculada levando em consideração 5 execuções, e o algoritmo processado por uma máquina virtual com 1 GB de memória ram e sistema operacional UBUNTU.
Para o índice de id na busca da Arvore B o tempo médio foi de 1683,6 nanosegundos, e para a busca sequencial, o tempo médio foi de 34634,4 nanosegundos.
Para o índice que utilizava o Ano e Ranking o tempo médio na busca da Arvore B foi de 1431 nanosegundos e na busca sequencial utilizando estes mesmos campos o tempo médio foi de 24879,4 nanosegundos.
A busca utilizando Arvore B e Índices mostra-se muito mais eficiente e rápida que apenas a busca sequencial.