UNIVERSIDADE BEIRA INTERIOR

Mestrado em Engenharia Informática 2019/20

Tecnologias de Base de Dados

2º Trabalho Prático

Versão 3.6 (22-Nov-2019)

Tema: Desempenho de base de dados

1. Introdução

O desempenho de bases de dados pode ser influenciado por vários fatores. Na figura seguinte (Fig. 1) apresentamos os principais fatores que podem ajudar a melhorar o desempenho de base de dados.

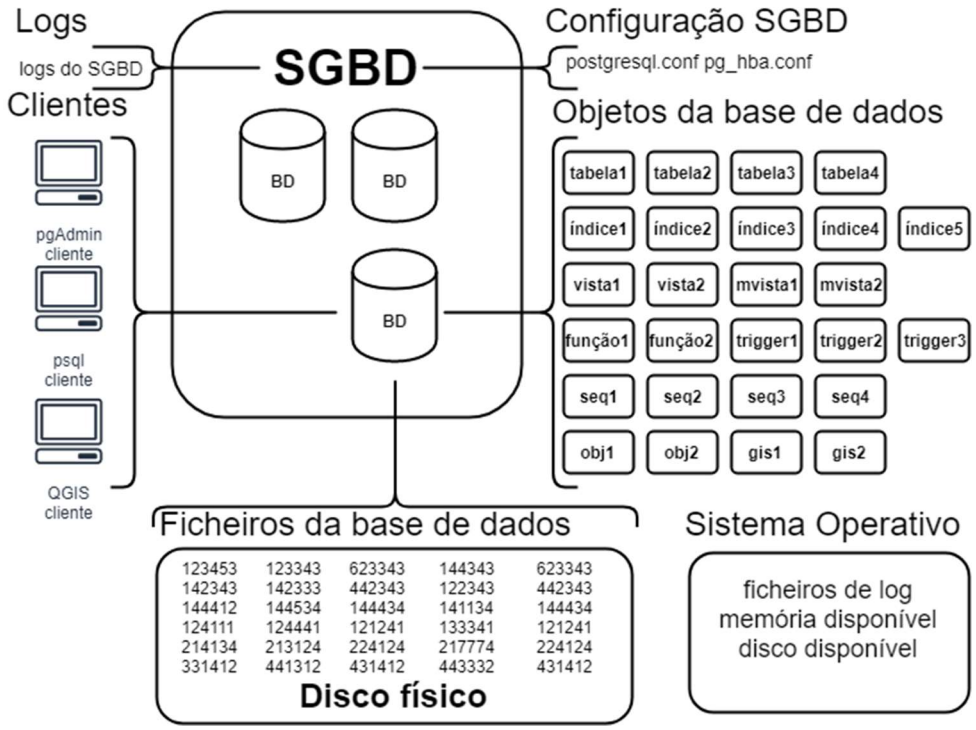

Figura 1 Componentes relevantes para o desempenho de uma base de dados

Trabalho

O trabalho proposto consiste em implementar uma base de dados de grande dimensão. Testar o seu desempenho inicial e desenvolver um conjunto de estratégias para a otimizar. No seguimento do trabalho anterior vamos usar dados reais, neste caso, os dados fornecidos pela Autoridade Nacional de Emergência e Proteção Civil, que disponibiliza informação em tempo real de todos as ocorrências que ocorrem no país através do seu site (http://www.prociv.pt/). Os datasets podem ser obtidos no site (http://centraldedados.pt/protecao_civil/) que possui 3 anos de ocorrências, de 2016 estão registadas 121187 ocorrências, no ano seguinte 217989 e em 2018 mais 234806. No total existem 573 982 registos diferenciados (https://github.com/centraldedados/protecao_civil/). Como o nosso objetivo é ter uma base de dados de grande dimensão vamos juntar a estes os dados de 2019 e depois aumentar o seu número para 1 000 000 000 de registos usando o DBSchema para gerar os dados aleatórios em falta. Na figura seguinte (Fig. 2) descreve-se os procedimentos a seguir passo a passo para implementar a otimização ao nível da base de dados.

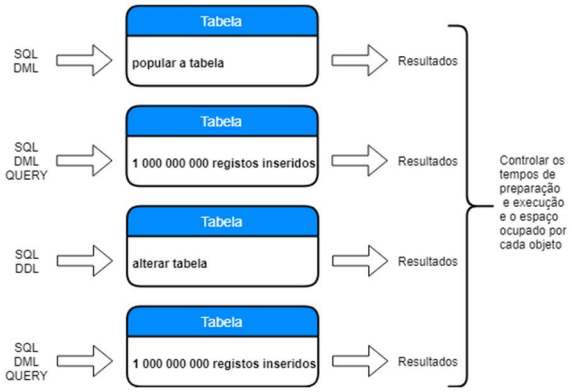

Figura 2 Procedimento passo a passo para otimizar tabelas com muitos registos.

Pretende-se que a implemente a base de dados de modo a que possa dar respostas em tempo útil a questões do tipo: quantas ocorrências acontecem em média por dia, qual o distrito que envolve mais meios, qual o tempo médio das ocorrências, quantas ocorrências com barcos ocorrem por ano, e com avião, qual a ocorrência que mobilizou mais meios durante mais tempo em cada ano. Mostre a lista ordenada por ordem decrescente com os vários tipos de ocorrências. Quantos operacionais terrestres e aéreos são usados por ano em cada distrito, qual a ocorrência com mais meios no verão, qual o distrito com mais incêndios, qual o mês em que o total de meios envolvidos é maior. Mostre também informação geográfica, por exemplo qual é a distância das ocorrências a hospitais, esquadras da polícia e bombeiros, quantas ocorrências por freguesia.

2. Objetivos do trabalho

1. Aprender a efetuar a gestão operacional do SGBD/BD:
    a. Gerir o hardware disponível: memória para o SGBD, processador, espaço em disco e velocidade e acesso.
    b. Manter o sistema operativo atualizado e sob vigilância, nomeadamente, gerir os recursos disponíveis para o SGBD, ler os logs e efetuar tarefas de manutenção operacional de maneira a garantir a disponibilidade e fiabilidade do sistema de suporte do SGBD.
    c. Otimizar o SGBD de modo a o manter atualizado com as versões mais recentes.
    d. Configurar o SGBD não otimizado para as condições reais de operação: sistema operativo, hardware, ligações (https://pgtune.leopard.in.ua/# , https://wiki.postgresql.org/wiki/Tuning_Your_PostgreSQL_Server).
    e. Configurar e usar os logs do SGBD eficientemente na gestão operacional para determinar atempadamente falhas, picos de serviço etc...
2. Criar uma base de dados que inclua a informação disponibilizada pela Proteção Civil ANEPC.
    a. Especificar o tipo de questões a que a base de dados deve dar resposta.
    b. Implementar a base de dados em conformidade.
    c. Usar os dados reais disponíveis para popular a base de dados e registar o tempo de carregamento e espaço ocupado.
    d. Acrescentar mais dados para ter pelo menos uma tabela com 1 000 000 de registos inseridos.
    e. Instalar a extensão POSTGIS para processar as coordenadas geográficas disponibilizadas em datasets como objetos tipo POINT (converter latitude e longitude para um objeto do tipo POINT).
    f. Importar mapa com divisão administrativa de Portugal (https://download.geofabrik.de/europe/portugal.html). Os formatos de dados incluídos em ficheiros OSM são aqui detalhados (https://download.geofabrik.de/osm-data-in-gis-formats-free.pdf) para depois se cruzar esta informação com o POINT obtido da latitude e longitude.
    g. Compreender o que são índices e como estes são utilizados para melhorar o desempenho.
    h. Aprender a escolher onde utilizar os diferentes tipos de índices BTREE, GIN, GiST, SP-GiST, BRIN e HASH para obter ganhos de desempenho.
    i. Identificar as queries mais lentas (pelo menos uma dessas consultas deve envolver vários campos).
    j. Construir índices para reduzir o tempo de resposta. Registar o tempo de indexação e espaço ocupado.
    k. Determinar o impacto dos índices no carregamento dos dados e no tamanho da base de dados.

    l. Análise de resultados e conclusões. Avaliar os ganhos de desempenho, comparando os respetivos tempos de planeamento e execução.
3. Testar em ambiente real de utilização os tempos de acesso utilizando aplicações pré-existentes que permitam aceder à informação disponível e visualizar dados geográficos com o pgAdmin4, QGIS ou outra aplicação, por exemplo baseada no trabalho anterior. Não é necessário implementar uma aplicação.
4. Escrever um relatório sobre o trabalho realizado.

3. Competências a adquirir

1. Capacidade para identificar problemas de desempenho em base de dados.
2. Saber adequar a implementação da base de dados às necessidades dos utilizadores.
3. Conhecimento dos mecanismos disponíveis para otimizar o desempenho de base de dados.
    a. conhecer o que são e para que servem os índices;
    b. identificar os diferentes tipos de índices;
    c. ser capaz de usar índices para resolver problemas de desempenho;
    d. otimizar de consultas e planos de execução,

4. Datas Importantes

Data de divulgação: 7 de novembro de 2019, 22 de novembro de 2019 (enunciado disponibilizado).

Formação dos grupos: Caso os grupos sejam diferentes dos do primeiro trabalho, a informação sobre a constituição dos novos grupos deve ser enviada ao docente até 1 de dezembro de 2019.

Data de entrega: Dom, 19 de novembro de 2019 (cada dia de atraso implica menos um valor na nota).

Data da defesa: 5ªf, 19 de novembro de 2019 ou 6ªf 20 de novembro de 2019 (a efetuar no turno prático ou teórico).

5. Grupos

O trabalho é realizado por grupos de 3 alunos do mesmo turno prático, caso não seja possível e excecionalmente podem ser considerados elementos de outro turno.

6. Avaliação

O trabalho está cotado para 5 valores.

Os componentes avaliados são: o grau de otimização das pesquisas, as estratégias desenvolvidas para melhorar o desempenho da base de dados, o desempenho antes e depois de aplicar os índices e também outras opções que considerou para a otimização.

O trabalho será defendido pelo grupo na 14ª semana de aulas no turno prático respetivo ou no turno teórico (sujeito a confirmação quando for entregue o trabalho). Alunos que não participem na apresentação do trabalho terão zero valores. Cada aluno pode ter uma classificação diferente da dos colegas de grupo, refletindo deste modo o seu desempenho no trabalho e na discussão do mesmo. Todos os trabalhos serão demonstrados e defendidos perante o docente da disciplina em sessões de defesa do trabalho específicas para cada grupo. As defesas dos trabalhos têm duração aproximada de 15 a 20 minutos. Na apresentação será fornecido pelo docente o trabalho recebido de cada grupo. Ficará a cargo dos alunos a sua instalação de modo a terem o trabalho pronto a demonstrar na sala e na hora marcada para a defesa.

7. Ferramentas de desenvolvimento

O sistema de gestão de base de dados a usar para a realização do trabalho é o Postgres. A base de dados deve possuir pelo menos uma tabela com 1Mb de registos inseridos.

Postgres, SGBD (https://www.postgres.com).
pgAdmin, aplicação para gerir o SGBD e interagir com a base de dados (https://www.pgadmin.org/download/).

psql aplicação de linha de comando para aceder e à base de dados.

DBSchema usada para gerar dados aleatórios (https://www.dbschema.com).

Postgis, extensão do Postgres para processar informação geográfica (https://computingforgeeks.com/how-to-install-postgis-on-centos-7/).

pgBadger é um utilitário que permite analisar e configurar os logs do Postgres (http://pgbadger.darold.net/), a sua utilização permite identificar rapidamente quais são as queries mais lentas.

QGIS ferramenta de interrogação e visualização de objetos GIS (mapas, polígonos, linhas e pontos) (https://www.qgis.org/en/site/).

8. Tarefas a efetuar

- Otimizar SGBD.
- Implementar a base de dados.
- Popular a base de dados. Guardar estatísticas sobre tempos, e tamanhos das tabelas e base de dados.
- Interrogar a base de dados. Recolher estatísticas sobre tempos de execução.
- Reformular a base de dados com índices e outras estratégias incluindo as coordenadas geográficas.
- Interrogar a base de dados. Recolher estatísticas sobre tempos de execução.
- Escrever as conclusões no relatório.
- Na apresentação do trabalho mostrar os resultados obtidos com testes de execução com e sem otimização.

9. Documentação a entregar

Código em SQL para reconstruir a base de dados sem dados no SGBD e o relatório.

Submeter através da página da disciplina no Moodle dois ficheiros **x_y_z_bd.zip** com a base de dados sem dados e **x_y_z_rel.zip** com o relatório (substituir x, y e z pelos números dos elementos do grupo e submeter dentro do prazo estabelecido.

10. Relatório

Deve ser elaborado um relatório detalhado abordando, pelo menos, os seguintes tópicos:

Introdução.

Desenho e população da Base de dados.

Mecanismos de otimização de base de dados.

a. Opções para melhorar o desempenho do SGBD.
b. Estratégias adotadas para melhorar o desempenho da base de dados.

Problemas e soluções para tabelas muito grandes e queries muito complexas.

a. Índices, comparar os índices e selecionar o mais adequado.
b. Outras estratégias.

Testes

a. Apresentar as operações efetuadas com o detalhe devido de modo a permitir comparar o antes e o depois.
b. Comparar as diferentes estratégias.
c. Analisar o desempenho.
d. Apresentar os resultados.

Conclusões e reflexão crítica sobre os resultados.

11. Referências

Lista de referências sobre procedimentos de otimização para base de dados Postgres:

a. Como analisar os planos de execução das várias operações efetuadas sobre as tabelas consultadas (https://www.postgresql.org/docs/current/runtime-config-query.html e https://github.com/kryonix/pg_cuckoo).
b. Como comparar, com e sem índices nos diferentes blocos de informação que podem ser extraídos a partir das operações utilizando a instrução EXPLAIN (https://www.postgresql.org/docs/9.2/using-explain.html) e também EXPLAIN ANALYZE (https://www.postgresql.org/docs/current/sql-analyze.html) para recolher dados.
c. As queries muito complexas, por exemplo que envolvam um número elevado de JOINS, podem degradar o tempo de resposta.
d. As operações de VACUUM (https://www.postgresql.org/docs/12/sql-vacuum.html) são fundamentais para a manutenção da operacionalidade da base de dados.
e. Se necessário depois de indexar a tabela, pode reordenar os registos da mesma no ficheiro onde está guardada de acordo com a ordem definida pelo índice (https://www.postgresql.org/docs/12/sql-cluster.html).
f. Como calcular tamanho ocupado por tabelas e base de dados (https://wiki.postgresql.org/wiki/Disk_Usage).
g. Como criar índices (https://www.postgresql.org/docs/current/sql-createindex.html).
h. Como efetuar a partição de tabelas (https://www.postgresql.org/docs/12/ddl-partitioning.html).
i. Algumas das operações SELECT, INSERT, DELETE e UPDATE podem precisar de ser preparadas previamente utilizando PREPARE e serem depois executadas com EXECUTE (https://www.postgresql.org/docs/12/sql-prepare.html).
j. Postgres SQL Prático (https://pt.wikibooks.org/wiki/Categoria:Livro/PostgreSQL_Pr%C3%A1tico).
k. Planeamento de queries (EXPLAIN), mostra o plano de execução de uma operação (https://www.postgresql.org/docs/current/sql-explain.html).
l. Recolha de dados estatísticos sobre uma operação (utilizada com EXPLAIN) (https://www.postgresql.org/docs/12/sql-analyze.html).
m. Ferramenta para facilitar a interpretação da operação EXPLAIN ANALYZE (https://explain.depesz.com/).
n. Postgres Explain Viewer (PEV) é uma ferramenta para proporcionar uma leitura mais fácil de planos de execução em Postgres (http://tatiyants.com/pev/#/plans/new , https://www.postgresql.org/docs/12/ddl-partitioning.html).
o. Como detectar queries mais lentas (https://www.cybertec-postgresql.com/en/3-ways-to-detect-slow-queries-in-postgresql/)
p. Para aumentar o desempenho de determinadas operações pode ser necessário implementá-las com PREPARE (https://www.postgresql.org/docs/12/sql-prepare.html).
q. As "*materialized views*" podem melhorar o desempenho (https://www.postgresql.org/docs/current/rules-materializedviews.html).
r. No caso em que seja necessário dividir tabelas em tabelas mais pequenas para melhorar o desempenho (https://severalnines.com/database-blog/guide-partitioning-data-postgresql).
s. Para otimizar operações é necessário também fazer um uso criterioso de VACUUM (https://www.postgresql.org/docs/12/routine-vacuuming.html).
t. Pgtune é uma ferramenta online para otimizar a performance do Postgres (https://pgtune.leopard.in.ua/#/).
u. Para determinar o tempo que determinada operação demora pode em psql usar o comando \timing para ligar/desligar o controlo temporal.
v. Comandos psql adicionais:
   i. Bases de dados \l \l+,
   ii. Tabelas na base de dados \dt \dt+,
   iii. Campos de uma tabela \d nome_tabela,
   iv. Esquemas disponíveis \dn \dn+,
   v. Funções disponíveis \df \df+
   vi. Índices disponíveis \di \di+
   vii. Vistas disponíveis \dv \dv+
   viii. Utilizadores e privilégios de acesso \du \du+
   ix. versão do SGBD SELECT version();
   x. Guardar histórico comandos \s ficheiro_historico
   xi. Executar script \i script.sql
   xii. Ajuda sobre um comando SQL \h comando_sql
   xiii. Ativar/desactivar tempo de execução \timing
   xiv. Editar comando e depois executar no psql \e
   xv. Determinar o espaço ocupado \d \d+
   xvi. Criar função \ef nome_funcao
   xvii. Executar o comando anterior \g
   xviii. Lista de comandos disponíveis \?