III SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UNIFESSPA
14 E 15 DE SETEMBRO
PROPIT/UNIFESSPA

# APLICAÇÕES OLAP UTILIZANDO O BANCO DE DADOS VERTICA: UM ESTUDO DE CASO NO LABORATÓRIO DE COMPUTAÇÃO CIENTÍFICA DA UNIFESSPA

# OLAP APPLICATIONS USING THE DATA BASE VERTICA: A CASE STUDY IN THE UNIFESSPA SCIENTIFIC COMPUTING LABORATORY

Aylce Lorena Carvalho Freire[1] - Unifesspa
Fernanda Carla Lima Ferreira[2] - Unifesspa
Marcelo Santana Camacho[2] - Unifesspa
Vitor de Souza Castro[2] - Unifesspa

## 1. INTRODUÇÃO

Os Sistemas de Gerenciamento de Banco de Dados (SGBD) são de extrema importância nos ambientes corporativos, pois facilitam a administração da ampla quantidade de dados presentes. Logo, é imprescindível que a escolha do SGBD seja cautelosa, sempre visando atender as necessidades do ambiente no qual será inserido. Ferreira e Junior (2012), afirmam que a escolha de um SGBD, dentre a grande diversidade disponível no mercado, é uma tarefa difícil, devido à responsabilidade que essa ferramenta representa por gerenciar uma das maiores riquezas de uma organização, que são as informações.

No decorrer dos anos, surgiram diferentes exigências impostas aos SGBDs Relacionais. Dentre essas exigências estão as cargas de trabalho do tipo *OLAP* (*Online Analytical Processing*), compostas por consultas complexas, que resultam no acesso a grandes quantidades de informações (DOMINICO, 2013). *OLAP* é considerado uma categoria de *software* que permite analistas e gerentes obterem respostas dentro dos dados através de uma rápida, consistente e interativa forma de acesso a uma grande variedade de possíveis visões (PRIMAK, 2008). Ou seja, as aplicações *OLAP* servem para que se obtenha resultados de consultas complexas de diferentes ângulos, onde essa forma de visão ajuda na tomada de decisões.

Na atualidade, o amplo volume de dados armazenados (*Big Data*), é um grande desafio para as corporações, pois com o crescimento do mesmo, surge a necessidade de se utilizar recursos que gerencie de forma eficiente esses dados. Nas últimas décadas, os SGBD no Modelo Relacional, tem sido amplamente utilizado na gerência desses dados. Contudo, essa ênfase em *Big Data*, tem causado fatores que induzem as corporações à buscarem alternativas em outros modelos, um desses modelos é o *NoSQL* (*Not only Structured Query Language)*, que segundo Brito (2010), é uma geração de banco de dados no modelo não relacional, que apresenta uma boa escalabilidade do sistema sendo utilizado quando o modelo relacional não apresenta performance adequada. Entretanto, dentro do próprio modelo relacional, encontram-se alternativas que apresentam características capazes de suprir essas necessidades de gerenciar esse grande volume de dados de forma eficiente. Um exemplo, é o banco de dados relacional orientado a coluna, o *Vertica*.

Segundo Mullins (2016), *Vertica* é um SGBD relacional construído especificamente para lidar com cargas de trabalho analíticas modernas onde sua plataforma usa uma abordagem em *cluster* para armazenar dados grandes, oferecendo funcionalidade de consulta e análise de alto desempenho. Diante disso, alguns fatores motivaram o desenvolvimento deste trabalho, dentre eles, a necessidade de um SGBD para gerenciar o grande volume de dados do LCC, de forma que as consultas analíticas realizadas em aplicações *OLAP* fossem otimizadas.

Outro motivo para a realização deste trabalho, é que na atualidade é frequente encontrarmos trabalhos que atestem a comparação de desempenhos entre SGBDs. Porém, são poucos trabalhos que realizam a análise comparativa utilizando o banco de dados *Vertica*.

Por fim, a importância deste trabalho é justificada por diferentes fatores, tais como: assegurar que o novo banco de dados implantado no LCC será realmente melhor que o banco de dados já utilizado, para assim

[1]Graduanda do curso de Bacharel em Sistemas de Informação (FACEEL/UNIFESSPA). Email: aylcelorena@hotmail.com
[2]Doutora em Física pela Universidade Federal de Sergipe, Atualmente é Pró-Reitora de Pós-Graduação, Pesquisa e Inovação Tecnológica na UNIFESSPA. Email: fernandaferreira@unifesspa.edu.br
[2] Mestrando em Computação Aplicada - PPGEE/UFPA e membro do Laboratório de Computação Científica da UNIFESSPA. Email: marcelo@unifesspa.edu.br
[2] Mestre em Ciência da Computação, Chefe de Divisão de Sistemas de Informação (CTIC/UNIFESSPA) e Professor (FACEEL/UNIFESSPA). Email: vitor@unifesspa.edu.br

adotar o novo recurso de forma segura; possibilidade de oferecer aos pesquisadores e membros do LCC melhores resultados em suas consultas analíticas; o estudo poderá apoiar outras organizações e pesquisadores que pretendam utilizar o *Vertica*.

## 2. MATERIAIS E MÉTODOS

Para o desenvolvimento do trabalho foi utilizado um *notebook Dell* com Sistema Operacional (SO) *Windows, core i5*, com *1 terabyte* de *HD* e *8GB* de *RAM*, onde foram realizados os testes de desempenho. Os bancos de dados utilizados para testes, foram instalados na ferramenta de virtualização, *VirtualBox*, cujo ambas as máquinas possuíam o mesmo Sistema Operacional (SO), *Debian GNU/Linux*, com a *versão 8.0 (jessie) amd64 ou versão 7.0 (Wheezy) amd64*, pois é esse o SO compatível com a infraestrutura de TI do LCC, possuindo as mesmas configurações, ou seja, a mesma quantidade de memória, mesma quantidade de CPU, com a mesma versão e a mesma quantidade de *Swap*.

As ferramentas utilizadas foram o banco de dados *PostgreSQL* versão 9.4.4, o visualizador de consultas OLAP *Saiku Standalone*, pois são essas as versões utilizadas no LCC; o *Vertica Community Edition*, versão v8.0.0-3, pois é um banco de dados com versão gratuita e segundo Kyurkchiev e Kaloyanova (2013), oferece um ótimo desempenho em consultas analíticas; o aplicativo de virtualização *VirtualBox*, devido a sua capacidade de executar quantas máquinas for necessárias, com as configurações desejadas e a base de dados utilizada para testes é referente ao ENEM (Exame Nacional do Ensino Médio), na qual possui informações referentes a notas de alunos que estavam concluindo ou que já havia concluído o Ensino Médio e com informações dos anos de 2010 a 2014.

No trabalho foi adotado o método de Estudo de Caso. O objetivo do estudo de caso é relatar os fatos como sucederam, descrever situações ou fatos, proporcionar conhecimento acerca do fenômeno estudado e comprovar ou contrastar efeitos e relações presentes no caso (GUBA E LINCOLN, 1994 apud ARAÚJO et al. 2008).

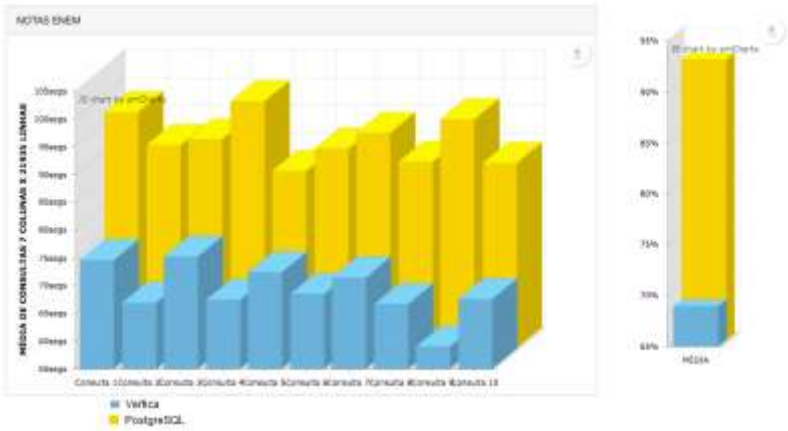
Diante disso, este estudo de caso foi realizado no LCC, da UNIFESSPA, onde foi implantado o banco de dados colunar Vertica, com o intuito de otimizar as consultas realizadas juntamente com o visualizador de consultas *OLAP*, o *Saiku.* Após a implantação do Vertica, foi realizada uma avaliação de desempenho entre ele e o banco de dados relacional orientado a linhas, o *PostgreSQL*, banco de dados já utilizado no laboratório. Essa avaliação de desempenho foi realizada por meio de consultas analíticas no *Saiku* e através das ferramentas de *Benchmark* chamada *DTM BD Stress*, no intuito de comprovar qual vai proporcionar o melhor desempenho em consultas analíticas em meio aos grandes volumes de dados no LCC.

O LCC é o Laboratório de Computação Científica da UNIFESSPA que auxilia pesquisadores que precisam de acesso a grandes bases de dados e estrutura computacional de alto desempenho, contribui com a Unifesspa, no desenvolvimento, implantação e aplicação de técnicas e modelos matemáticos e computacionais para a resolução de problemas científicos e tecnológicos dos diversos grupos de pesquisa da Instituição. (LCC, 2016).

## 3. RESULTADOS E DISCUSSÕES

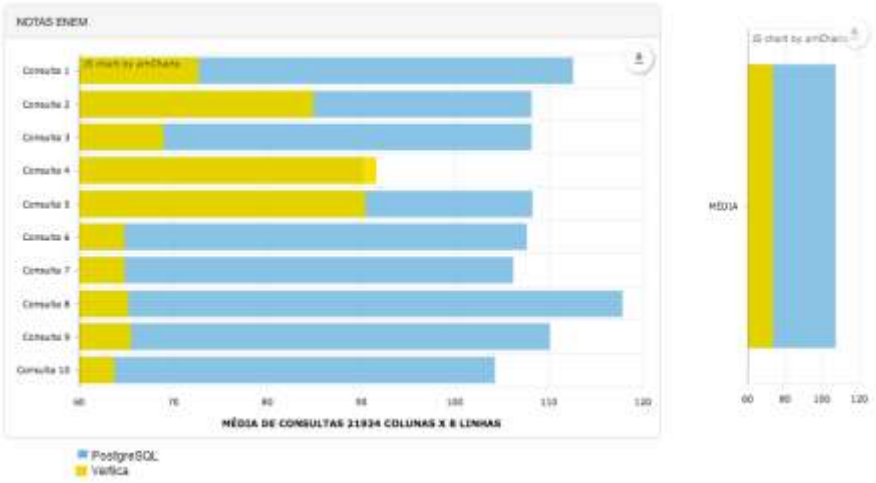
O resultado da análise comparativa foi dividido em duas etapas, a primeira etapa foi a realização de consultas analíticas no *Saiku*, onde foram executadas 10 consultas com 7 colunas por 21935 linhas e 10 consultas com 21934 colunas por 8 linhas, como mostram a figura 1 e figura 2. O resultado dessa primeira etapa mostra que o banco de dados *Vertica* tem um desempenho melhor em consultas analíticas que o banco de dados *PostgreSQL*, com a média de diferença de 23 segundos nas consultas com 7 colunas por 21935 linhas e a diferença de 34 segundas nas consultas com 21934 colunas por 8 linhas.

Figura 1: Resultado primeira Etapa, 7 colunas x 21935 linhas.




Fonte: Próprio Autor.

Figura 2:Resultado primeira Etapa, 21934 colunas x 8 linhas.




Fonte: Próprio Autor.

A segunda etapa foi a realização de testes de stress na ferramenta de *Benchmark DTM BD Stress*. Na ferramenta foi executado um script que realiza consultas na base de dados do ENEM. Neste teste é definido que o *script* seja executado a mesma quantidade de vezes tanto no banco de dados do *PostgreSQL* quanto no banco de dados *Vertica*, ou seja, é realizada 110 execuções, como mostra na tabela 1. No teste é realizado o cálculo da velocidade média de iterações por segundos das execuções; a duração média, máxima e mínima em segundos de cada iteração; o tempo em que cada banco de dados demora em segundos para se conectar e o tempo de se desconectar. Podemos observar que o *Vertica* tem um desempenho melhor que o *PostgreSQL* em quase todos os requisitos mostrados na tabela 1, perdendo a penas no tempo de conexão.

Tabela 1:Resultado segunda Etapa. Testes realizados na ferramenta *DTM DB Stress*.

| - | Quantidade de execução | Velocidade média de Iterações por segundo | Duração média, segundos | Duração máxima, segundos | Duração mínima, segundos | Tempo de conexão, segundos | Tempo de desconexão | Duração total, segundos |
|---|---|---|---|---|---|---|---|---|
| **Vertica** | 110 | 0.1432 | 6.9854 | 17.7810 | 5.8430 | 0.0930 | 0.0000 | 768.3940 |
| **PostgreSQL** | 110 | 0.0800 | 12.5042 | 251.0690 | 9.5300 | 0.0160 | 0.0930 | 1375.4660 |

## 4. CONCLUSÃO

Ressaltamos nesta conclusão que os testes realizados neste trabalho foram executados de forma igual em ambos os bancos de dados em estudos, sempre utilizando as mesmas quantidades de execuções, mesmo script, mesmas configurações e levando em conta que o processo de execução das consultas concorrem com outros processos do SO, Internet, etc.

Em todos os testes realizados, se tratando de execuções de consultas analíticas, o banco de dados Colunar obteve um desempenho melhor em relação ao banco de dados *PostgreSQL*. Na primeira etapa conclui-se que o *PostgreSQL* é entre 34,19% e 45,55% mais lento que o *Vertica*. Na etapa 2 o *PostgreSQL* mostra que é 79% mais lento que o *Vertica*. Podemos observar também através dos resultados que quanto mais colunas executadas, melhor é o desempenho do banco Colunar. Assim sendo, o *Vertica* é um banco de dados indicado para ser utilizado no LCC, de modo que seria de extrema importância essa otimização das consultas realizadas no *Saiku*.

## AGRADECIMENTOS

Agradecemos ao Laboratório de Computação Científica da UNIFESSPA por possibilitar o desenvolvimento da pesquisa.

## REFERÊNCIAS

PRIMAK, Fábio Vinícius. **Decisões com bi (business intelligence)**. Fabio Vinicius Primak, 2008.

BRITO, Ricardo W. **Bancos de dados NoSQL x SGBDs relacionais: análise comparativa**. Faculdade Farias Brito e Universidade de Fortaleza, 2010.

DOMINICO, Simone. **Tuning: um estudo sobre a otimização de desempenho de sistemas gerenciadores de banco de dados relacionais sob carga de trabalho de suporte a decisão**. Universidade Tecnológica Federal Do Paraná, Guarapuava, 2013.

FERREIRA, Erick Rodrigues; JÚNIOR, Sergio M. Trad. **Análise de desempenho de Bancos de Dados**. Universidade Presidente Antônio Carlos, Barbacena, 2012.

KYURKCHIEV, Hristo; KALOYANOVA, Kalinka. **Performance Study of Analytical Queries of Oracle and Vertica**. In: Proceedings of the 7th ISGT International Conference. 2013. p. 127-139.

LCC. **Página Laboratório de Computação Científica**. Disponível em < https://lcc.unifesspa.edu.br> Acessado em: 07/04/2017.

MULLINS, Craig s. **Explorando a Plataforma HPE Vertica Analytics**. Disponível em: <http://searchdatamanagement.techtarget.com/feature/Exploring-the-HPE-Vertica-Analytics-Platform> Acessado em: 04/04/2017.