CQM-DW: Gerenciamento de Consultas em Data Warehouse Distribuído

Orlando P. S. Júnior[1], Edie C. Santana[2], Ernando O. Santana[1]

[1]Instituto Federal de Educação e Tecnologia de Mato Grosso (IFMT)
Núcleo Avançado de Campo Verde - São Vicente – MT – Brazil

[2]Faculdades Evangélicas Integradas Cantares de Salomão (FEICS) - Cuiabá-MT-Brazil
orlando.junior@svc.ifmt.edu.br, ediecs@gmail.br, ernando-1231@hotmail.com

Abstract. *The main objective of this paper is to present a solution to facilitate the management of queries in data warehouse distributed repositories stored in cloud. As result the research presents the module CQM-DW, which was developed in order to manage the execution of queries to the data warehouse distributed repositories, counting on a proposed partitioning algorithm queries, which is based on meta-information about the data organization and physical distribution.*

Resumo. *O objetivo principal deste artigo é apresentar uma solução para facilitar o gerenciamento de consultas em repositórios de Data Warehouse distribuídos armazenados em nuvem. Como resultado, a pesquisa apresenta o módulo CQM-DW, que foi desenvolvido com o intuito de gerenciar a execução de consultas a repositórios de DW distribuídos, contando com a proposta de um algoritmo de particionamento de consultas, que se baseia em meta-informações sobre a organização e distribuição física dos dados.*

1. Introdução

Data Warehouse (DW) são repositórios de dados voltados à análise histórica de informações e são compostos por informações oriundas de diferentes tipos de sistemas e aplicações. Tais informações são carregadas ao DW e transformadas de modo a refletir uma visão ampla e histórica acerca das informações operacionais das organizações (KIMBAL; ROSS, 2002). A sumarização e temporalidade das informações armazenadas, os repositórios de DW proporcionam informações estratégicas para as organizações, o que é fundamental para a constante melhoria de seus resultados.

Devido a esta importância, faz-se necessário que tais repositórios de dados, mesmo volumosos, estejam disponíveis para consulta contando com armazenamento de dados flexível com alta capacidade de vazão. O módulo CQM-DW executa consultas a repositórios de DW distribuídos e armazenados em nuvem, propondo a divisão das consultas através da análise das informações acerca da distribuição dos dados, possibilitando a execução paralela das subconsultas geradas com o intuito de diminuir o tempo de resposta das consultas em grandes massas de dados.

Na seção 2 apresentaremos os principais conceitos de DW e Computação em Nuvem (CN). Na seção 3 apresentaremos a arquitetura do módulo CQM-DW. Na seção 4 discutiremos os resultados dos experimentos realizados.

2. Data Warehouse e Computação em Nuvem

A Computação em Nuvem oferece um grande conjunto de recursos (*hardware* e/ou *software*) virtualizados facilmente utilizáveis. Tais recursos são entregues como serviços pela internet, onde são contratados conforme a demanda (ABADI, 2009).

Data Warehouse são repositórios de dados para análise histórica de informações. Tais informações são gerenciais, orientadas a um determinado assunto de interesse da organização e que integram informações operacionais de diferentes fontes. Devido à característica histórica, tais informações não são voláteis (KIMBAL; ROSS, 2002).

Diversas pesquisa foram realizadas com o intuito de melhorar o tempo de consultas em grandes repositórios de DW, com destaque o ParGRES e o ESQP. O ParGRES tem por objetivo prover paralelismo para a execução eficiente em aplicações OLAP sobre clusters de banco de dados (MATTOSO *et al.*, 2006). O ESQP é um algoritmo baseado em MapReduce realiza consultas OLAP em réplicas de dados (ZHAO *et al.*, 2010).

Ambos os trabalhos, ParGRES e ESQP, trabalham com réplicas de dados, já o módulo CQM-DW foi construído para trabalhar com dados distribuídos em nuvem, consultados conforme a necessidade, visando Escalabilidade e Otimização de Recursos de armazenamento e processamento, propondo uma maneira de particionar as consultas com base nas informações sobre a distribuição dos dados.

3. Arquitetura do módulo CQM-DW

A abordagem do CQM-DW é focada na CN, pois fornece um mecanismo gerenciar consultas em DW's distribuídos, armazenados em nuvem, onde o CQM-DW atua como um *middleware* que gerencia as subconsultas aos locais de armazenamento do DW.

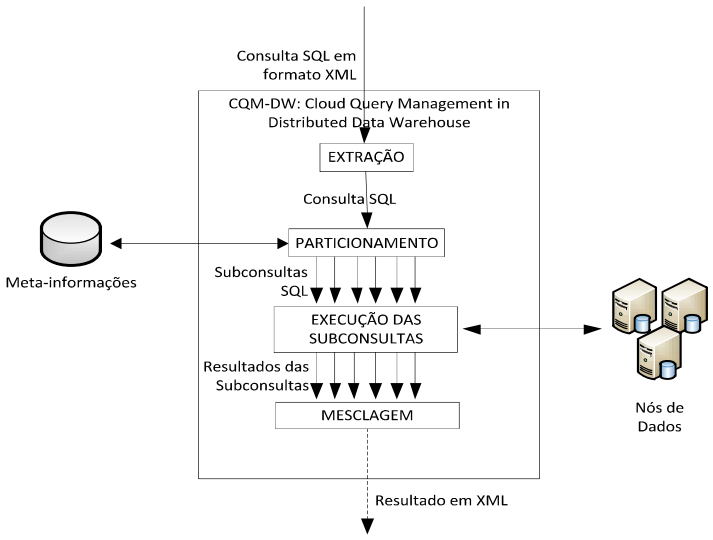

Figura 1 - Arquitetura CQM-DW

O CQM-DW foi desenvolvido propondo uma solução para o problema da localização dos dados, onde o particionamento das consultas é feito segundo o critério da distribuição dos dados. As subconsultas geradas são executadas paralelamente, visando oferecer maior escalabilidade em consultas em grandes repositórios de DW.

O *engine* principal é responsável por organizar os processos necessários para a execução de uma consulta ao DW. A Figura 1 apresenta a arquitetura do módulo CQM-DW, composto por Extração, Particionamento, Execução de Subconsultas e Mesclagem das Informações.

3.1. Extração e Particionamento

A consulta é extraída de uma estrutura XML, que contém informações que identificam o DW e o comando SQL a ser executado, além de informações de segurança de acesso. Em seguida ocorre o Particionamento da consulta, que a subdivide em subconsultas de acordo com as meta-informações. As meta-informações registram a maneira como os dados do DW foram distribuídos. A Figura 2 ilustra o processo de particionamento da consulta.

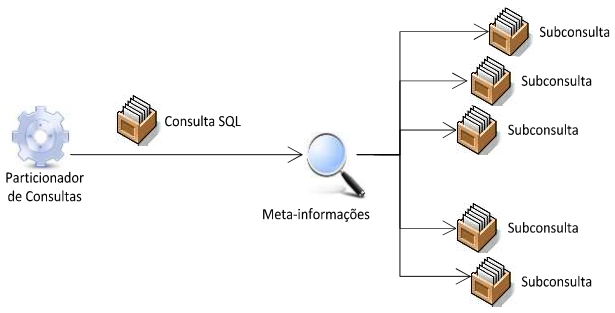

Figura 2 - Particionamento da Consulta

3.2. Execução das subconsultas e Mesclagem dos resultados

As subconsultas são enviadas ao servidor de dados adequado, de acordo com as meta-informações e executadas em paralelo, visando diminuir o tempo total de execução. Para a conclusão do processo, é necessário que todas as subconsultas tenham sido concluídas.

Após a execução das subconsultas, as informações resultantes sõ mescladas. Este processo se encarrega de concatenar e reordenar os dados resultantes das subconsultas, conforme a ordem requisitada na consulta extraída.

4. Resultados

Para verificação da performance do módulo, foram montados dois ambientes, o ambiente Solo, composto por um único servidor com cerca de 4 GB de memória RAM e processador de 2.3 Ghz, e o ambiente Heterogêneo, composto por quatro servidores com 1GB de RAM e processadores de 1.5 Ghz. Em ambos os casos os servidores estavam hospedados em Nuvem e o SGBD escolhido foi o PostGreSQL.

O Quadro 1 apresenta os resultados do tempo de execução, desde o recebimento da consulta até a entrega das informações solicitadas. São apresentados o tempo de execução em cada ambiente e a quantidade de registros recuperada em cada quadro de execução.

Quadro 1: Comparativo do Tempo de Execução (em segundos)

	Ambiente Solo	Ambiente Heterogêneo	Diferença de Tempo	Qtde. de Registros
1º Quadro	33,5869210	27,8605936	5,7263274	120.000
2º Quadro	66,3927974	53,0190325	13,3737649	240.000
3º Quadro	134,0876694	108,6882166	25,3994528	480.000

Constatou-se que o CQM-DW é capaz de diminuir o tempo das consultas. Tal resultado dá-se pelo fato de que as consultas são executadas paralelamente, o que resultou na diminuição do tempo de execução total mesmo em cenários com quantidades crescentes de dados.

O módulo CQM-DW foi desenvolvido em C#.Net, usando técnicas de programação em *thread,* com acesso a dados usando ADO.Net. A requisição de processamento de consultas e a consequente entrega dos dados é realizada via WebServices, característica que torna o módulo plenamente interoperável.

5. Conclusões

O CQM-DW é um modulo que proporciona a consulta em DW's distribuídos em nuvem, visando a diminuição do tempo total de execução de consultas em grandes repositórios de dados. O modulo propõe uma solução para o particionamento de consultas, baseado em informações sobre a distribuição dos dados, e execução paralela das subconsultas geradas.

Destacamos como possíveis trabalhos futuros a implementação da estrutura de acesso a dados usando meios mais velozes para a serialização e entrega dos dados e a construção de mecanismos de análises dos resultados das consultas para a proposição de modificações na distribuição dos dados.

O trabalho se encontra em andamento. Alunos do curso Tecnologia em Análise e Desenvolvimento de Sistemas do IFMT – campus São Vicente - estão desenvolvendo o módulo CQM-DW como trabalho de conclusão de curso e também como atividades de grupos de pesquisa. Espera-se com o desenvolvimento de novas caracteríscias, o módulo possa se apresentar mais robusto e funcional no atendimento às consultas aos DW's.

6. References

Kimbal, R, M. Ross. (2002). The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd ed., John Wiley & Sons, 2002.

Abadi, D. J. (2009). Data management in the cloud: Limitations and opportunities". IEEE Data Eng. Bull., vol. 32, no. 1, pp. 3-12, Mar. 2009

Mattoso, M. L. Q.; Zimbrão, G.; Lima, A. A.; Baião, F.; Braganholo, V.; Aveleda, A. A.; Miranda, B.; Almentero, B. K. ; Costa, M. N. (2005). ParGRES: Middleware para Processamento Paralelo de Consultas OLAP em Clusters de Banco de Dados. In: SBBD - Sessão de Demos, Uberlândia. pp.19-24.

ZHAO, J; HU, X; MENG, X. (2010). ESQP: An efficient SQL query processing for cloud data management. In: Proceedings of the second international workshop on Cloud data management. ACM, 2010. p. 1-8.