9ª MOSTRA ACADÊMICA UNIMEP
08 a 10/11 de 2011
Tema: "Ambiente e Sustentabilidade"

9º Congresso de Pós-Graduação

GERENCIAMENTO DE CONSULTAS EM DATA WAREHOUSE DISTRIBUÍDO EM NUVEM

Autor(es)

ORLANDO PEREIRA SANTANA JUNIOR

Orientador(es)

MARINA TERESA PIRES VIEIRA

1. Introdução

A informação é um bem de estimado valor para as organizações. Segundo Chen, Chen e Jiang (2010), fatores de velocidade, integridade e consistência são muito importantes quando se trata de acesso às informações.
Num cenário competitivo onde buscam-se a diminuição de custos, contar com a Computação em Nuvem se faz mais que importante.
A Computação em Nuvem vem com o propósito de dar uma nova visão para o fornecimento de serviços de TI.
Esta visão é baseada no modelo de serviços básicos oferecidos à sociedade, como água e luz, onde há um provisionamento conforme a demanda do usuário.
Visto que as organizações tendem gerar grandes volumes de informação, surge a necessidade de manter estas informações em repositórios consistentes e confiáveis com custo reduzido.
Tais repositórios de dados, também conhecidos como Data Warehouse, são de uso específico para a tomada de decisão nas organizações.
Forlani (2006) afirma que um Data Warehouse consolida dados oriundos de provedores de informação autônomos, heterogêneos e distribuídos, em uma única base de dados, normalmente volumosa.
As informaçãos contidas em um Data Warehouse são caracterizadas por serem orientadas a um assunto específico, integradas, históricas e não voláteis, além de organizadas em diferentes níveis de agregação (KIMBALL e ROSS, 2002; MOHANIA et al., 1998, WU e BUCHMANN, 1997).
Manter infraestrutura para o armazenamento e tratamento de um Data Warehouse, tendo em vista o constante crescimento e a crescente utilização de tais informações, pode se tornar dispendioso para as organizações. Segundo Sousa et al (2010) a melhor saída tanto para as pequenas quanto para as grandes organizações é a alocação de recursos de armazenamento e processamento conforme a demanda.
Segundo Abadi (2009) estima-se que 27% das bases de dados existentes no mercado são bases de dados voltadas à análise (OLAP) e ainda conta-se com a taxa de crescimento de 10.3% ao ano.
Na visão de Cao et al. (2010), o modelo de computação em nuvem tem surgido como escolha ideal para o armazenamento de dados com altas taxas de crescimento.
Possuir e manter uma infraestrutura que possa atender solicitações de variados níveis de complexidade requer um tratamento baseado na carga de trabalho mais onerosa, o que pode fazer com que as organizações mantenham grandes capacidades computacionais, de armazenamento e processamento, para atendimento de grandes cargas de trabalho em momentos isolados de mais utilização.
Segundo Chen, Chen e Jiang (2010), a necessidade de manter grandes infraestruturas que atendam diferentes níveis de cargas de trabalho, requer uma grande quantidade de hardware, o que pode se tornar dispendioso para as organizações.
Nesta visão, um Data Warehouse poderia ser espalhado por recursos de armazenamento presentes em uma nuvem computacional.
Estes recursos de armazenamento de dados são denominados Nodes de Dados, onde cada um desses recursos seria utilizado conforme a demanda, fazendo uso da otimização e alocação dinâmica dos recursos da nuvem.
Este trabalho pretende propor e implementar um módulo middleware, que possa efetuar consultas em Data Warehouse Distribuído em

Nodes de Dados hospedados em nuvem. Esse módulo será utilizado em conjunto com um Servidor OLAP-XML, em desenvolvimento em outro trabalho de mestrado.

2. Objetivos

O objetivo deste trabalho é a construção de um módulo middleware que poderá ser utilizado para auxiliar um Servidor OLAP no processo de consulta aos dados para o processamento das operações solicitadas pelo usuário, de modo que a distribuição e organização física dos repositórios de DW sejam transparentes para o Servidor OLAP.
Pretende-se identificar as informações necessárias para guiar o processo de particionamento de consultas de acordo com a forma como os dados do data warehouse estão espalhados na nuvem, e também gerenciar e executar as consultas que serão enviadas a cada Node de Dados.

3. Desenvolvimento

Em conformidade com as tendências acerca do consumo de recursos de forma otimizada e compartilhada, e com vistas na diminuição de custos e melhor aproveitamento dos recursos, surge o seguinte desafio: como repassar a solicitação do usuário, que foi formulada por meio de uma consulta, à camada responsável pelo armazenamento dos dados, tendo em vista a transparência da distribuição dos dados, sendo que a organização e localização dos dados são desconhecidas para o usuário solicitante?
Para a solução do desafio acima está sendo desenvolvido um módulo middleware que poderá ser utilizado com um Servidor OLAP-XML, onde esta junção proporcionará uma alternativa ao consumo de informações para a tomada de decisão.
A arquitetura do Servidor OLAP em conjunto com o middleware é apresentada na figura 1, demonstrando que as solicitações do usuário ao Servidor OLAP serão repassadas ao middleware e este consultará a camada responsável pelo armazenamento dos dados para a obtenção das informações necessárias.
Este trabalho se dedicará ao desenvolvimento do midlleware Módulo de Gerenciamento de Consultas em Data Warehouse Distribuído em Nuvem, visando atender diversos níveis de cargas de trabalho e também promover a otimização de recursos por meio de uma infraestrutura em nuvem.

Módulo Middleware de Gerenciamento de Consultas

O processamento de consultas ao Data Warehouse distribuído será realizado por meio do particionamento de uma consulta principal em subconsultas.
O processo de particionamento será dirigido por um conjunto de informações acerca da organização e distribuição física dos dados nos recursos de armazenamento da nuvem.
Tais informações são denominadas meta-informações.
Conforme a tabela 1 as meta-informações preliminares são: Nome do Node de Dados, Endereço IP, Tecnologia de SGBD usada, Atributos de Tempo que restringem as informações armazenadas e informações de acesso e segurança.
Em outras palavras, as meta-informações serão responsáveis por direcionar a forma de consultar os dados do Data Warehouse distribuído, fornecendo parâmetros de localização e organização dos dados dentro da infraestrutura de armazenamento da nuvem.
Com o intuito de minimizar o tempo de execução das subconsultas, o módulo middleware propõe a execução paralela das subconsultas geradas, contando com recursos de processamento escaláveis conforme a demanda.
O resultado de cada subconsulta será mesclado a um único resultado, produzindo então o resultado da consulta principal.
A figura 2 apresenta a arquitetura do módulo middleware proposto, a seguir são comentados seus módulos constituintes.

A. Extrator

O primeiro módulo do midleware é o Extrator, que tem a função de extrair informações de conteúdos XML enviados pelo Módulo 1 - XML-DW.
Este módulo recebe a Consulta SQL Original em formato XML, de forma que do conteúdo XML seja extraído o comando SQL referente à consulta desejada e também as demais informações inerentes ao DW que se deseja consultar.
Contidas na Consulta SQL Original, empacotada em XML, a informação “name”, que identificará qual será o DW a ser consultado, em conjunto com as informações de segurança “user”, “password”, são suficientes para a conexão ao DW.

São extraídos da Consulta SQL Original em XML, o comando SQL que se deseja executar no conjunto de bases de dados componentes do DW em questão, a identificação do DW e informações de segurança suficientes para a conexão aos Nodes de Dados do DW, com isto passa-se a identificar o conjunto de meta-informações que estarão aptas a auxiliarem no processo de consulta ao DW distribuído.

B. Particionador

A consulta SQL recebida pelo Particionador é então particionada em subconsultas.
A divisão da consulta em subconsultas visa facilitar a execução paralela destas subconsultas, de modo que as subconsultas possam ser direcionadas paralelamente aos respectivos Nodes de Dados que contenham as porções de dados necessárias.
De acordo com as meta-informações mostradas na tabela 1, uma consulta que abrange os anos de 2000 a 2003, será dividida em 3 outras subconsultas, a exemplo, de acordo com as meta-informações do Data Warehouse.
A primeira subconsulta deverá ser direcionada ao Node 1, que responde pelos anos de 2000 a 2001, a segunda subconsulta gerada, deverá ser direcionada aos Nodes 2 e 3, pois os mesmos possuem as informações dos anos de 2002, sendo que o Node 2 atende pelas informações referentes aos meses compreendidos entre 1 e 6 do ano de 2002, e o Node 3 atende pelas informações referentes aos meses compreendidos entre 7 e 12 do mesmo ano.
Para finalizar a execução, uma última subconsulta deverá ser encaminhada ao Node 4, por conter parte dos dados do ano de 2003, sendo este ano ainda contido na consulta SQL inicial como critério de filtro.

C. Executor

Os Nodes de Dados, como componentes importantes, são servidores de banco de dados hospedados em lugares distintos podendo ou não estar geograficamente distribuídos, sob características de hardware e software diferentes.
Esses Nodes de Dados são pontos de apoio de armazenamento do Data Warehouse.
O Executor envia as subconsultas ao Nodes de Dados adequados. Estes efetuam o processamento das subconsultas e entrega o resultado ao módulo de Mesclagem.
Cada Node de Dados que compõe a infraestrutura em nuvem do Data Warehouse distribuído assume o armazenamento dos dados bem como o processamento de consultas relacionadas àquela porção de dados.

E. Mesclagem
De acordo com a Consulta SQL inicial, as subconsultas geradas pelo processo de Particionamento de Consultas, ao serem executadas, seus resultados são mesclados e entregues ao Servidor OLAP.
Esta fase se encarregará de agrupar, de forma correta, os resultados das subconsultas geradas, de modo que o resultado das subconsultas representará o resultado desejado da Consulta SQL recebida.
Cabe ressaltar que os resultados serão empacotados em formato XML para facilitar a comunicação com o Servidor OLAP.

4. Resultado e Discussão

Os módulos componentes da arquitetura da figura 2 foram parcialmente implementados, e atualmente, o módulo middleware está realizando consultas em Data Warehouse distribuído em recursos de armazenamento locais.
Nos primeiros testes foram criados 2 Nodes de Dados com tecnologia MS SQL Server com bases de dados semelhantes em termos de estruturas e dados.
A Extração das informações do XML está ocorrendo corretamente.
Algumas informações estão sendo adicionadas às meta-informações para que o Particionador efetue a divisão corretamente.
O Executor está enviando as consultas aos Nodes de Dados, embora ainda resta a implementação do envio assíncrono.
A Mesclagem dos resultados está ocorrendo como planejado.
Como continuidade do trabalho está sendo discutido um padrão para as consultas SQL que o módulo receberá, com vistas a diminuir a complexidade no tratamento da consulta SQL.

5. Considerações Finais

O trabalho apresentando encontra-se em processo de desenvolvimento e pretende-se nas próximas fases ampliar as meta-informações, melhorar o processo de particionamento de consultas e implementar a execução paralela de subconsultas.
O módulo middleware, quando usado em conjunto com outra ferramenta que objetive a extração de informações, possibilitará a consulta a bases de dados de DW distribuídas em recursos de armazenamento em nuvem.
O principal objetivo do módulo middleware é a transparência na distribuição e organização física dos dados.
O módulo middleware será dotado de serviços web, que possibilitarão a leitura e entrega de dados em diversas plataformas devido ao uso do protocolo SOAP.

Referências Bibliográficas

Armbrust, M., Fox, A., Grif?th, R., Joseph, A. D., Katz, R. H., Konwinski, A., Lee, G., Patterson, D. A., Rabkin, A., Stoica, I., and Zaharia, M. (2009). Above the clouds: A berkeley view of cloud computing. Technical report, EECS Department, University of California, Berkeley.

Agrawal, D., Das, S., and Abbadi, A. E. (2010b). Big data and cloud computing: New wine or just new bottles? PVLDB, 3(2):1647–1648.

Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., and Brandic, I. (2009). Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Gener. Comput. Syst., 25(6):599–616.

Chen C., Chen G., Jiang D., Providing Scalable Database Services on the Cloud. Proceeding WISE'10 Proceedings of the 11th international conference on Web information systems engineering

Forlani, D. T. Fragmentação Vertical de Dados em Data Warehouses no Sistema WebD2W. 2006. Dissertação (Programa de Pós-Graduação em Ciência da Computação) – Faculdade de Ciência da Computação, Universidade Estadual de Maringá, 2006.

Inmon, W. H., Building the Data Warehouse. John Wiley & Sons, Inc, 2. ed, 1996. 401p.

Kimball, R., Ross, M. The Data Warehouse Toolkit: o guia completo para modelagem multidimensional. John Wiley & Sons, Inc., 2. ed, 2002.

Mohania, M.; Samtani, S.; Roddick, J.; Kambayshi, Y. Advances and Research Directions in Data Warehousing Technology. The Australian Journal of Information Systems, 1999.

Wu, M.C.; Buchmanm, A.P. Research Issues in Data Warehousing. In: The German Database Conference, Uml, Germany. Proceedings… p. 61-82, 1997.

Abadi , D. J., “Data Management in the Cloud: Limitations and Opportunities,” in Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2009,pp. 3–12.

Cao, Y., Chen, C., Guo, F., Jiang, D., Lin, Y., Ooi, B.C., Vo, H.T., Wu, S., Xu, Q.: ES: A cloud data storage system for supporting both OLTP and OLAP. In ICDE(2011) 291-302.

Anexos

Table 1 - Meta-informações do DW

Nome	Endereço	Porta	Tecnologia	Atributos de Tempo	Usuário
NODE 1	187.33.1.44	1433	SQL Server	Ano=2000 a 2001 Mês=Todos	User
NODE 2	187.33.1.45	1433	SQL Server	Ano=2002 Mês=1 a 6	User
NODE 3	187.33.1.46	5535	PostgreSQL	Ano=2002 Mês=7 a 12	User
NODE 4	187.33.1.47	6550	MYSQL	Ano=2003 Mês=Todos	User

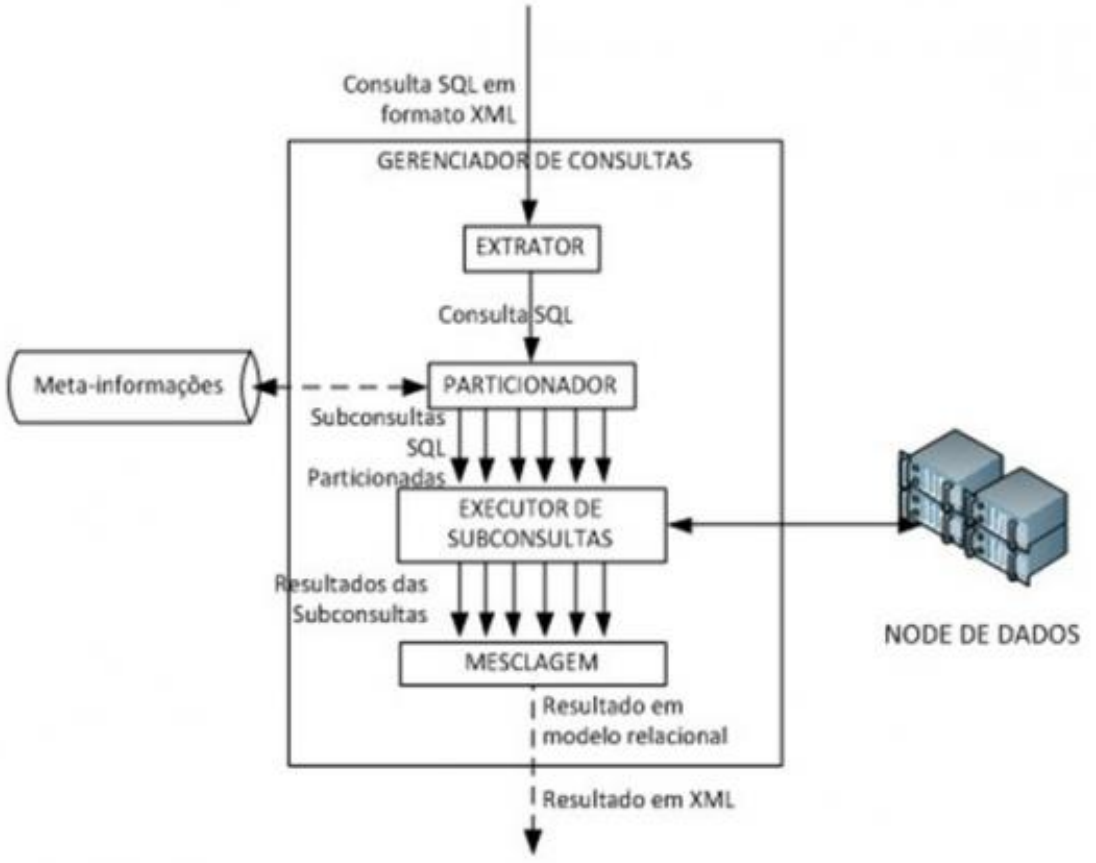

Figure 2 - Detalhamento do Middleware

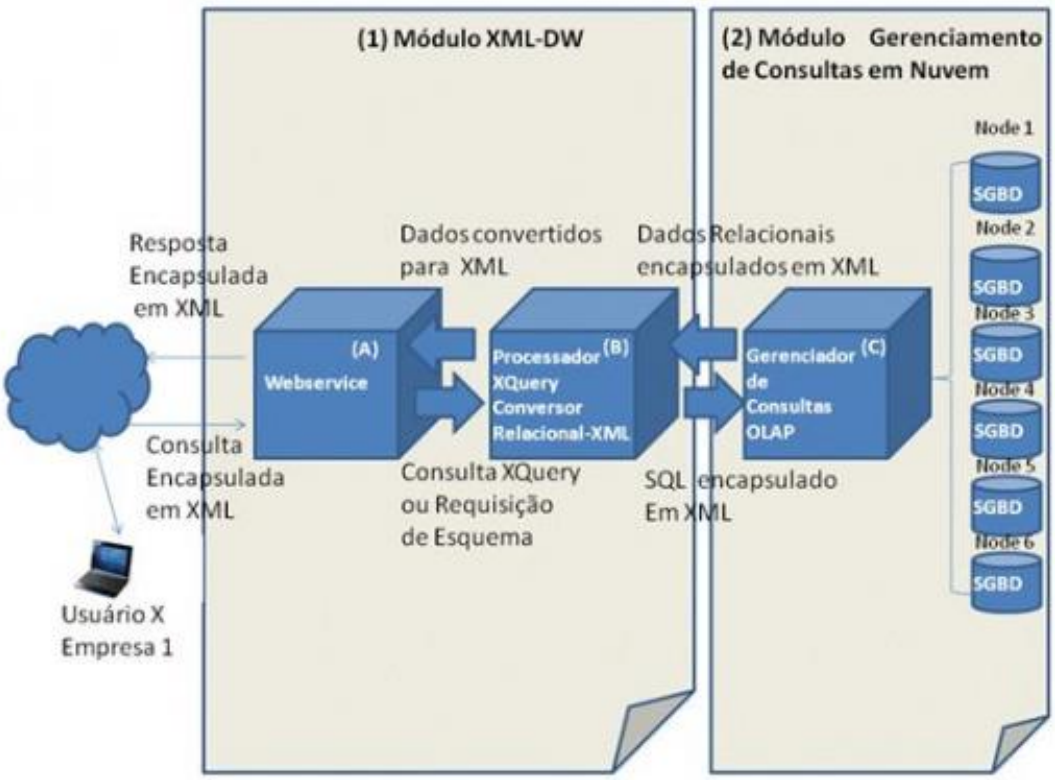

Figure 1 - Arquitetura do Servidor OLAP