UNIVERSIDADE FEEVALE

MOISÉS FELIPE LEHNEN

##### PROPOSTA DE REPLICAÇÃO DE DADOS EM AMBIENTE HETEROGÊNEO

(Título Provisório)

###### Anteprojeto de Trabalho de Conclusão

Novo Hamburgo, setembro de 2010.

MOISÉS FELIPE LEHNEN

moises.felipe@gmail.com

##### PROPOSTA DE REPLICAÇÃO DE DADOS EM AMBIENTE HETEROGÊNEO

(Título Provisório)

Universidade Feevale

Instituto de Ciências Exatas e Tecnológicas

###### Curso de Sistemas de Informação

###### Anteprojeto de Trabalho de Conclusão

Professor orientador: Juliano Varella de Carvalho

Novo Hamburgo, setembro de 2010.

# RESUMO

O desejo de integrar e proporcionar acesso aos dados de um empreendimento é para muitos, uma das principais motivações para o uso de um sistema de banco de dados. A replicação de dados por sua vez, é um processo de compartilhar informações entre diferentes bases de dados. O uso da replicação de dados é bastante comum em atividades como migração de sistema legado, ou na integração de sistemas. Essa tarefa pode tornar-se um processo bastante crítico em um ambiente com um número crescente de bases de dados distintas e heterogêneas, e por isto é importante que este processo não afete a desempenho do ambiente ou sistema. Uma replicação mal planejada pode acarretar em riscos para o negócio de uma empresa ou organização. Em algumas situações podemos ter a ocorrência de uma base de dados central, que pode ser caracterizada como uma base com informações sumarizadas, onde podem ser encontrados registros replicados a partir de outras bases de dados ou registros de comum interesse para as mesmas. Essa base central pode se tornar uma fonte rica em conhecimento para o negócio da empresa, conhecimento este que pode ser desconhecido para qualquer gestor ou colaborador da organização. Desta forma, este trabalho tem como objetivo propor uma solução aos problemas encontrados na replicação de dados no ambiente de uma empresa.

Palavras-chave: Base de dados. Replicação de dados. Informação. Compartilhar.

SUMÁRIO

MOTIVAÇÃO...........................................................................................................................5

OBJETIVOS ..............................................................................................................................8

METODOLOGIA......................................................................................................................9

CRONOGRAMA ....................................................................................................................10

BIBLIOGRAFIA ....................................................................................................................11

# MOTIVAÇÃO

Devido à rápida expansão da empresa[[1]](#footnote-1) no território nacional, existe a necessidade de compartilhar informações a partir de diversas bases de dados. A empresa em questão atua no segmento de varejo e possui diversas unidades de venda por todo o país. Em cada unidade comercial existe uma base de dados local, as informações dessa base são replicadas de modo assíncrono para um banco de dados central localizado na matriz da empresa, onde são tomadas as decisões administrativas. Manutenções em dados referentes a produtos e outros cadastros básicos são efetuadas na matriz da empresa e replicadas para a base de cada unidade. Este processo de replicação é gerenciado por uma camada de software executada diretamente no servidor. Esta aplicação fica residente na memória do servidor, e constantemente verifica, por um intervalo definido de três minutos, atualizações ocorridas em uma base de dados de origem desde a última verificação e submete essas atualizações a uma ou mais bases de dados destino.

“Um banco de dados é uma coleção estruturada de dados relacionados a alguns fenômenos reais que estamos tentando modelar.” (ÖZSU, 2001, p. 29). Ainda segundo o autor, uma das motivações importantes por trás do uso de sistemas de bancos de dados é o desejo de integrar os dados operacionais de um empreendimento e proporcionar acesso centralizado e, portanto controlado a esses dados.

“Replicação é um processo de cópia e manutenção de objetos do banco de dados, como tabelas, nos múltiplos bancos de dados que compõem um sistema distribuído. Alterações aplicadas em um site específico são capturadas e armazenadas localmente antes de serem enviadas e aplicadas a cada uma das localizações remotas. (URBANO, 2003, p.1, tradução nossa). Ainda segundo Urbano(2003), a replicação significa que os mesmos dados estão disponíveis em vários locais, para isso utiliza-se uma tecnologia de bancos de dados distribuídos para compartilhar os dados em múltiplos sites, porém um banco de dados replicado não é exatamente um banco de dados distribuído, neste, os dados estão disponíveis em vários locais, mas uma tabela em particular encontra-se em apenas um local. “Os computadores de um sistema de banco de dados distribuídos recebem diversos nomes, como sites ou nós, dependendo do contexto no qual são inseridos.” (SILBERCHATZ, 1999, p.557).

Entre os benefícios da utilização da replicação de dados destacam-se o desempenho, e a disponibilidade. Segundo Özsu (2001) o uso da replicação de dados pode ser desejável por motivos de desempenho, confiabilidade e disponibilidade. Para Date(2004), a replicação pode ser desejável por duas razões. Primeiramente, ela pode prover um ganho de desempenho, visto que as aplicações podem executar localmente ao invés de terem de se comunicar com um site remoto; e em segundo lugar, a replicação pode oferecer uma melhor disponibilidade. Ainda segundo Date(2004), a maior desvantagem da replicação é quando um objeto replicado é alterado, pois todas as cópias do objeto precisam ser atualizadas. “Em geral, a replicação aumenta o desempenho nas operações *read* e aumenta a disponibilidade dos dados para as transações de somente leitura. Entretanto, as transações de atualização geram um grande *overhead*.” (SILBERCHATZ, 1999, p.590).

A principal razão da escolha da empresa por adotar esse modelo de replicação é a centralização. No contexto da organização pode-se identificar a seguinte situação: o cadastro de produtos precisa estar centralizado na matriz, já o faturamento é feito pela unidade, e este não pode ser comprometido em função de alguma falha de comunicação entre a unidade e a matriz. Na matriz é utilizado o DBMS (*Database Management System*) Oracle, porém em função do custo, nas unidades é utilizado o DBMS PostgreSQL, o que caracteriza um ambiente heterogêneo. “Se todos os servidores (ou DBMS locais) utilizam o mesmo software e todos os usuários (clientes) utilizam o mesmo software, o DDBMS (*Distributed Database Management System*) é chamado de homogêneo, caso contrário é chamado de heterogêneo.” (ELMASRI, 2003, p.815, tradução nossa).

Atualmente a replicação entre as bases de dados é realizada através de um software desenvolvido pela própria empresa, que identifica as atualizações ocorridas em uma determinada origem aplicando-as aos demais destinos. Inicialmente este sistema atendia as necessidades da empresa, porém, com a expansão da própria houve um aumento significativo no volume de dados e conseqüentemente no volume de replicações, o que tornou a utilização do software bastante instável, com demora excessiva na atualização das réplicas e sobrecarga no servidor da base de dados central, sendo necessário muitas vezes efetuar uma atualização em lote manualmente em cada base de dados fora do horário de expediente. Também não há como definir prioridades na replicação, pois muitas vezes existem dados críticos que devem ser replicados de modo instantâneo e apesar do replicador verificar as atualizações ocorridas na base de dados de origem em um curto intervalo de tempo, a atualização das réplicas pode ocorrer com dias de atraso, tornando-se mais um agravante do ambiente atual. Além disso, não há um controle de falha na atualização de dados em nenhuma das bases se dados. Não há como verificar se uma atualização foi efetuada com sucesso.

A transparência de rede é uma importante propriedade de bancos de dados distribuídos, é a capacidade de ocultar detalhes da implementação ou distribuição das bases de dados. Para Silberchatz (1999), a transparência de rede é definida como o grau de desconhecimento que os usuários do sistema podem manter em relação ao armazenamento dos dados dentro de um sistema distribuído. Para Özsu (2001), em sistemas de bancos de dados centralizados, o único recurso disponível que precisa ser isolado dos usuários são os dados. Ainda segundo o autor, em um ambiente de bancos de dados distribuídos a rede é um recurso que deve ser administrado da mesma forma, o usuário deve ser protegido contra detalhes operacionais da rede, se possível é desejável, até mesmo, ocultar a existência da rede.

Com base nos conceitos citados, o objetivo deste trabalho é propor uma solução para os problemas no processo de replicação de dados existente, proporcionando melhor escalabilidade para o mesmo, estando ciente de que a empresa encontra-se em uma crescente expansão no mercado nacional, o que representa um aumento no número de bases de dados no cenário de replicação.

OBJETIVOS

Objetivo geral

Propor solução para o problema no processo de replicação de dados da empresa.

Objetivos específicos

* Descrever o ambiente de armazenamento de dados atual da empresa;
* Descrever os problemas da situação real;
* Caracterizar replicação de dados;
* Investigar técnicas de replicação de dados;
* Procurar ferramentas de replicação disponíveis no mercado;
* Avaliar utilização de ferramenta existente, ou desenvolvimento de ferramenta própria.
* Detalhar a solução encontrada;
* Analisar os resultados da nova ferramenta.
* Demonstrar os resultados obtidos a um especialista da área na empresa;

METODOLOGIA

O desenvolvimento deste trabalho ocorrerá em duas etapas, o Trabalho de Conclusão

I e o Trabalho de Conclusão II. A primeira etapa consiste em pesquisar conceitos, definições, algoritmos e técnicas de replicação de dados. Para isso serão utilizados livros, revistas, artigos acadêmicos e outras fontes disponíveis. As informações coletadas nessa etapa subsidiarão a parte teórica destinada ao Trabalho de Conclusão I e a parte prática a ser realizada no Trabalho de Conclusão II.

Ainda na primeira etapa deste trabalho serão identificados os problemas do ambiente atual, através de entrevistas com os responsáveis pelo processo. Além disso, será realizada uma pesquisa sobre ferramentas disponíveis no mercado, bem como a instalação e preparação de um ambiente de testes das ferramentas e algoritmos selecionados para a continuidade do trabalho.

Na segunda etapa deste trabalho será desenvolvida a parte prática, como testes de técnicas e ferramentas de replicação, bem como avaliar se os resultados obtidos são pertinentes às necessidades da empresa. Ainda nesta etapa, serão demonstrados os resultados obtidos nos testes para algum responsável da área na empresa. Na ocorrência de resultados satisfatórios, será realizada uma proposta para a utilização da solução encontrada.

A finalização do Trabalho de Conclusão I ocorrerá no final do segundo semestre de

2010 e o Trabalho de Conclusão II têm a previsão de entrega para o final do primeiro semestre de 2011.

# CRONOGRAMA

Trabalho de Conclusão I

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Etapa | Meses | | | |
| Ago | Set | Out | Nov |
| Elaboração do Anteprojeto |  |  |  |  |
| Entrega do Anteprojeto |  |  |  |  |
| Pesquisar Conceitos |  |  |  |  |
| Identificar problemas |  |  |  |  |
| Pesquisar ferramentas |  |  |  |  |
| Preparação do ambiente de testes |  |  |  |  |
| Elaboração do TCI |  |  |  |  |
| Entrega do TCI |  |  |  |  |

Trabalho de Conclusão II

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Etapa | Meses | | | |
| Mar | Abr | Mai | Jun |
| Teste de técnicas e ferramentas |  |  |  |  |
| Avaliação dos resultados |  |  |  |  |
| Demonstrar resultados à especialista |  |  |  |  |
| Propor solução para o problema |  |  |  |  |
| Elaboração do TCII |  |  |  |  |
| Entrega do TCII |  |  |  |  |

# BIBLIOGRAFIA

COULOURIS, George; DOLLIMORE, Jean; KINDBERG, Tim. **SISTEMAS DISTRIBUÍDOS CONCEITOS E PROJETO**. Porto Alegre: Bookman, 2007.

DATE, C.J.. **AN INTRODUCTION TO DATABASE SYSTEMS**. Pearson Education, Inc. 2004.

ELMASRI, Ramez; NAVATHE, Shamkant B.. **FUNDALMENTS OF DATABASE SYSTEMS**. Pearson Education, Inc. 2004.

FAYYAD, Usama: **FROM DATA MINING TO KNOWLEDGE DISCOVERY IN DATABASES**. 1997. Disponível em: <http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf> Acesso em: 29 ago. 2010.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **DATA MINING UM GUIA PRÁTICO**. Rio de Janeiro: Elsevier: 2005.

ÖZSU, M. Tamer; VALDURIEZ, Patrick. **PRINCÍPIOS DE SISTEMAS DE BANCOS DE DADOS DISTRIBUÍDOS**. Rio de Janeiro: Campus, 2001.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. **METODOLOGIA DO TRABALHO CIENTÍFICO**. Novo Hamburgo: Feevale, 2009.

SILBERCHATZ, Abraham; KORTH, Henry F.; SUDARSHAN, S.. **SISTEMA DE BANCO DE DADOS.** São Paulo: Pearson Makron Books, 1999.

TANENBAUM, Andrew S.; STEEN, Maarten Van. **DISTRIBUTED SYSTEMS PRINCIPLES AND PARADIGMS**. Pearson Education, Inc., 2007.

URBANO, Randy. **ORACLE DATABASE ADVANCED REPLICATION.** Oracle Corporation, 2003.

WITTEN, Ian H.; FRANK, Eibe; **DATA MINING PRACTICAL MACHINE LEARNING TOOLS AND TECNIQUES**. San Franciso: Elsevier, 2005.

1. A empresa em questão não autorizou a divulgação do seu nome, porém, está ciente deste estudo e autorizou o uso das informações contidas nas suas bases de dados. [↑](#footnote-ref-1)