Resumo. Este trabalho avaliou o desempenho em um ambiente de banco de dados distribuído analisando o tráfego dos dados que tramitam na rede local que abriga os *hosts*. Para tanto, foram realizadas requisições de dados no referido ambiente de testes, por meio da ferramenta JMeter, emulando, desta forma, usuários virtuais. Assim, o tráfego gerado permitiu avaliar o comportamento de um banco de dados distribuído levando em consideração todos os fatores que compuseram o planejamento dos experimentos.

1. Introdução

A busca por armazenamento de dados e informações tende a aumentar cada vez mais com o passar dos anos. Sobre isto, Junior e Oliveira (2011, p. 01) dizem que “o modo de trabalhar com a grande quantidade de dados que o ser humano e suas máquinas produzem diariamente torna-se cada vez mais complexo”

Pessoas conectadas a Internet buscam a todo o momento notícias, informações, realizam pesquisas e, com isso, o fluxo de dados limita-se por vários fatores. Dentre estes fatores destacam-se o limite máximo que pode ser alcançado nas redes. Outro fator importante é a disponibilidade dos dados para os requisitantes. Tais dados são obtidos a partir de bases de dados e estas precisam sempre estar disponíveis.

Há alguns anos, os bancos de dados (ou bases de dados) tornaram-se importantes para as empresas por armazenarem informações pertencentes a estas. No início, os sistemas de banco de dados eram centralizados e o acesso a estes eram feitos através de terminais. Com o passar do tempo, os computadores pessoais tiveram uma melhora significativa de velocidade e o custo destes foram reduzindo gradativamente. Assim, estes ocuparam o lugar dos terminais que eram conectados aos sistemas centralizados. A interface com os usuários ficaram por conta dos computadores pessoais, enquanto que os sistemas centralizados ficaram responsáveis por atender as solicitações dos sistemas clientes, ou seja, os sistemas centralizados tornaram-se sistemas servidores (SILBERSCHATZ, et al., 2006).

Problemas com hardwares podem ocorrer e os sistemas servidores podem deixar de atender as requisições dos usuários e, caso não exista uma opção que faça com que o serviço volte a funcionar rapidamente, tais requisições ficam prejudicadas.

Mesmo que esse tipo de banco de dados seja utilizado até os dias de hoje, devido aos problemas mencionados anteriormente e devido à grande necessidade de

armazenar um número cada vez maior de dados, gerenciar estes de forma centralizada tornou-se uma tarefa árdua a ser realizada. (SHIBAYAMA, 2004)

Uma alternativa utilizada atualmente para solucionar estes tipos de problemas é o uso de um banco de dados configurado como um sistema distribuído que, por ter a base em vários computadores, caso um deixe de funcionar, as requisições de dados serão supridas, pois os outros computadores ainda estarão ativos na rede.

Segundo um trabalho proposto por Shibayama (2004), os dados são distribuídos em vários computadores, sendo que cada nó tem autonomia dos dados armazenados em si próprios e, mesmo assim, consegue compartilhar estes com outros nós.

O presente trabalho justifica-se pelos apontamentos que evidenciaram quais dos fatores analisados (isto é, número de usuários, número de requisições e número de *hosts*) que compõem este ambiente de testes devem ser levados mais em conta quando se vai projetar um banco de dados distribuídos com as características propostas pelo ambiente de testes estudado.

2. Objetivos

O presente trabalho tem como objetivo avaliar o desempenho em um sistema de banco de dados distribuído utilizando computadores virtuais e reais, levando em consideração o tráfego de rede obtido através de requisições efetuadas por usuários simulados.

3. Metodologia

Para a montagem do ambiente de testes e realização dos experimentos foram utilizados os seguintes programas: VMware® *Workstation*, versão 8.0.4 *build*-744019, possibilitando a criação de máquinas virtuais (*Virtual Machine* ou VM) nas quais foram instalados os seguintes software: *Windows* XP *Professional* Versão 2002 *Service Pack* 3 e o SGBD SQL *Server* 2008 *Developer*,sendo que este último foi o banco de dados utilizado neste trabalho. Foi utilizado também o programa Putty, o qual possibilitou o acesso ao software nativo do switch, e a ferramenta JMeter, que permite que uma carga de trabalho seja gerada para que se possa obter os valores para a elaboração de uma análise das respostas a que se quer. No caso deste projeto, JMeter foi utilizado para simular uma certa quantidade de usuários que fizeram uma certa quantidade de requisições ao banco de dados

distribuído (BDD) desta experiência. Essa ferramenta possibilita fazer estes tipos de requisições, simulando usuários reais. Desta maneira foi possível realizar a análise de desempenho deste projeto.

4. Desenvolvimento

Neste tipo de trabalho deve-se adotar o conjunto correto de métodos e técnicas para que o objetivo seja alcançado. Caso a metodologia utilizada nesse tipo de pesquisa não seja a correta, os resultados podem ser obtidos de forma inexata pois, fatores cruciais que influenciam a variável de resposta podem ser deixados de lado. Esta pesquisa tem como variável de resposta a quantidade de pacotes que trafegam em uma rede por segundo (ou PPS – Pacotes por Segundo), gerados através de requisições em um banco de dados distribuído.

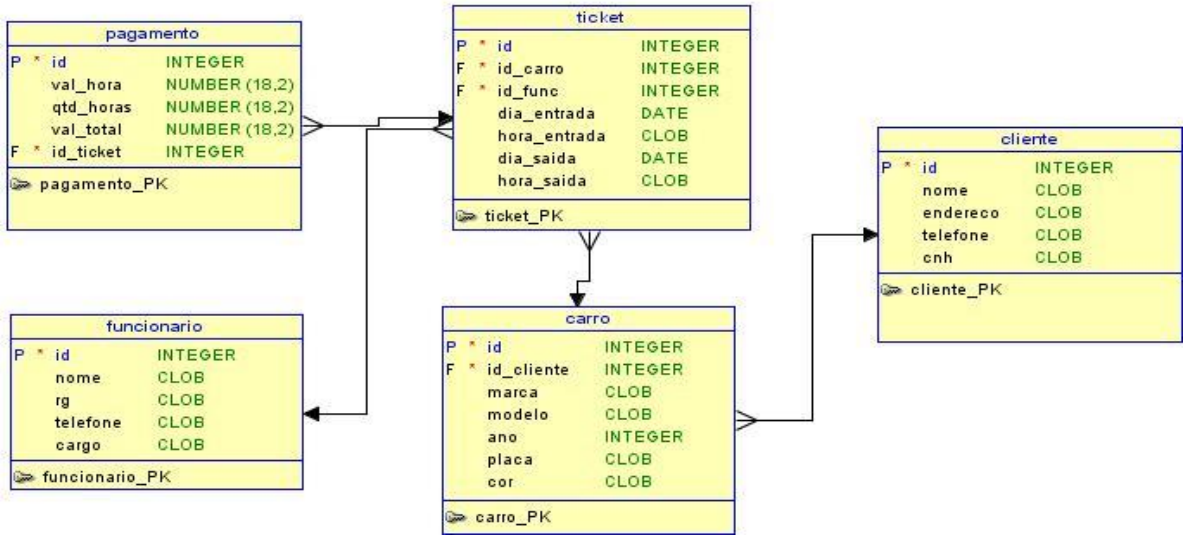

Figura 1 – Modelo de dados relacional
Fonte: Elaborado pelo autor

Para esta pesquisa foi utilizado o modelo de dados de um sistema de estacionamento, composto pelas seguintes tabelas: cliente, funcionário, pagamento, ticket e carro. A Figura 1 expõe o modelo lógico de dados implementado, dispondo como as tabelas se relacionam, bem como as respectivas colunas e, consequentemente, os tipos destas. Assim pode-se verificar como o banco de dados utilizado neste projeto foi elaborado, fisicamente.

O ambiente de testes deste experimento foi baseado em envio e recebimento de requisições a um banco de dados distribuídos. Para a parte de infra-estrutura de rede de computadores foi elaborada uma rede local situada na sala 3, contida

dentro das dependências da Faculdade de Tecnologia de Lins “Prof Antonio Seabra”, sendo utilizados 8 computadores para a realização dos testes e um *notebook* para efetuar o bombardeio de requisições, através do JMeter. Em sete destes oito computadores, e no *notebook*, foi instalado o programa VMware® *Workstation*, versão 8.0.4 *build*-744019. Assim foi possível criar uma máquina virtual (VM) em cada uma destas máquinas físicas. em sete destas oito máquinas virtuais foram instalados o sistema operacional *Windows* XP *Professional* Versão 2002 *Service Pack* 3 e o SGBD Microsoft SQL *Server* 2008 *Developer*. Desta maneira, o BDD foi composto por sete *hosts*, sendo que um *host* representava o publicador e os outros seis representavam os subscritores. No oitavo *host*, ou seja, no *notebook*, também foi instalado o *Windows* XP como sistema operacional, e esta VM ficou disponibilizada para efetuar as requisições ao BDD, com a utilização da ferramenta JMeter. O nono computador foi utilizado para a visualização das portas do switch que compôs a rede deste experimento, através do programa Putty, com o qual foi possível acessar, em modo texto, o software do switch.

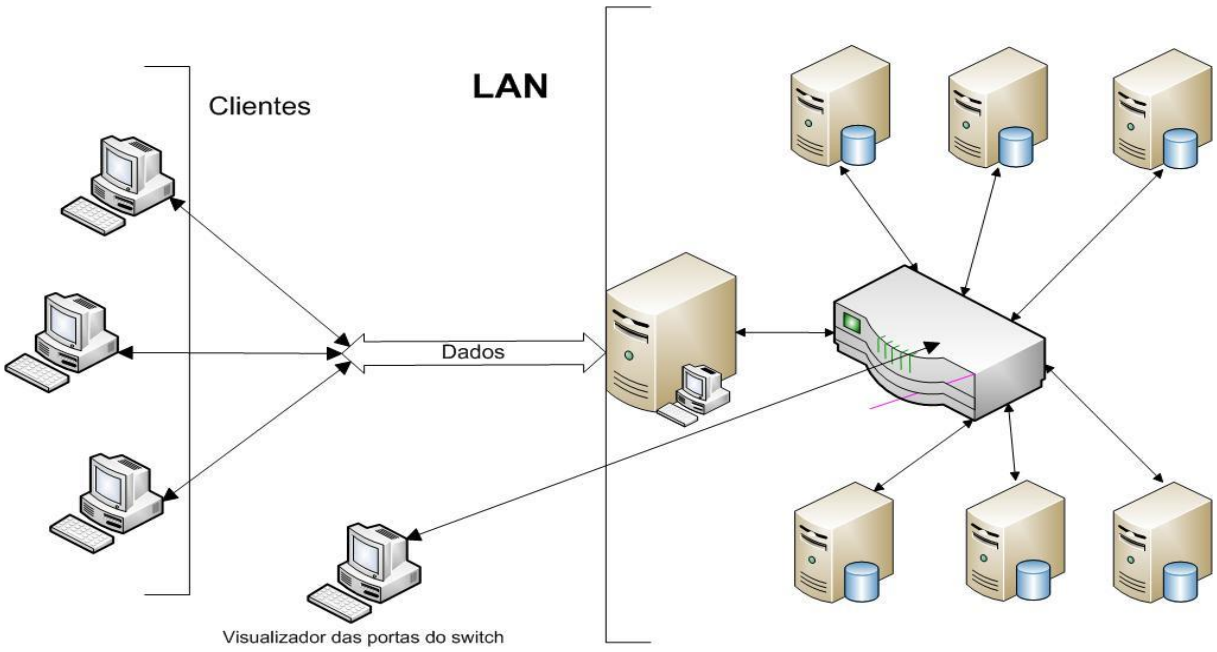

Figura 2 – Ambiente de testes
Fonte: Elaborado pelo autor

A Figura 2 representa o ambiente de testes, de uma maneira geral, utilizado nesta avaliação. De forma a compreender esta figura, descreve-se a mesma da seguinte maneira: os clientes foram simulados através da utilização do JMeter, já as requisições que chegavam ao BDD, representadas pela seta nomeada “Dados”,

poderiam vir de qualquer lugar, de forma global ou local, a este sistema de banco de dados distribuídos; o visualizador foi utilizado para verificar o tráfego nas portas do switch. Por fim, os sete *hosts*, compostos por suas respectivas máquinas virtuais formaram, desta forma, o banco de dados distribuídos.

O switch utilizado no controle de tráfego da rede e visualização de pacotes que trafegavam na rede foi um DLink DGS – 3627, tendo como taxa de transferência de dados o equivalente à 1 Gigabits por segundos (GBPS). Este switch possui vinte e quatro portas gerenciáveis, lembrado que para esta avaliação, apenas nove portas foram utilizadas, e o dispositivo não possuía conexão com a Internet.

Quadro 1 – Qualificações dos *hosts*

Computador	VM	Switch (porta)	Função	IP (Protocolo de Internet)
LAB6-M02	servidor1	1	Publicador	10.1.1.1/192.168.1.1
LAB6-M01	cliente-01	2	Subscritor 01	10.1.1.10
LAB6-M04	cliente-02	3	Subscritor 02	10.1.1.20
LAB6-M03	cliente-03	4	Subscritor 03	10.1.1.30
LAB6-M06	cliente-04	5	Subscritor 04	10.1.1.40
LAB6-M05	cliente-05	6	Subscritor 05	10.1.1.50
LAB6-M08	cliente-06	7	Subscritor 06	10.1.1.60
Gustavo-Notebok	JMeter	9	Cliente Virtual-Jmeter	192.168.1.150
LAB6-M07	----------	8	Visualizar Tráfego-Switch	10.90.90.10

Fonte: Elaborado pelo autor

Com relação às máquinas utilizadas, para um melhor entendimento destas, virtuais ou física (esta última foi utilizada para visualizar o tráfego de dados do switch), que compuseram esta avaliação, é disposto o Quadro 1, contendo o nome do computador físico, o nome da máquina virtual (VM), a porta a que este acessa o switch, qual a atribuição do computador na avaliação de desempenho e o IP atribuído a este computador.

No trabalho foi utilizado o planejamento fatorial completo, que leva em consideração os fatores escolhidos e os níveis referentes a estes. O projeto fatorial utilizado aqui foi o 2^3, o que quer dizer que foram escolhidos 3 (três) fatores, sendo que cada um destes fatores possuem 2 (dois) níveis. Assim, foram efetuados 8 experimentos, pois $2^3 = 8$.

A medida de desempenho adotada neste projeto foi a *throughput*, que “é a taxa a qual os pedidos podem ser atendidos pelo sistema.“ (JAIN, 1991, p.53). O *throughput* é medido, neste projeto, em pacotes por segundo (pps), ou seja, a variável de resposta que deve ser adquirida neste experimento é a quantidade

média de pacotes que trafegam na rede elaborada, por segundo. Para tal, foi dividido um total de pacotes trafegados por um dado intervalo de tempo, em segundos.

Quadro 2 – Fatores e níveis para avaliação

Fatores	Nível 1	Nível 2
Número de *hosts* (subscritores)	3	6
Número de requisições de dados	8100	16200
Número de usuários	45	90

Fonte: Elaborado pelo autor

Para esta avaliação foram considerados três fatores e cada fator assumiu dois níveis, como exposto no Quadro 2.

Quadro 3- Combinações entre fatores e níveis

	a	**b**	**c**
Experimentos	**Núm. Usuários**	**Num. Requisições**	**Núm. *Hosts***
1	45	8100	3
2	45	8100	6
3	45	16200	3
4	45	16200	6
5	90	8100	3
6	90	8100	6
7	90	16200	3
8	90	16200	6

Fonte: Elaborado pelo autor

O quadro 3 demonstra as combinações entre os fatores e níveis que resultaram nos oito experimentos realizados neste trabalho, levando em conta a descrição dos fatores e níveis evidenciados no Quadro 2. O Quadro 4 representa os níveis, fatores e respostas de um projeto fatorial 2^3. Assim, o experimento 1 (um) é composto da seguinte maneira: xa = 1, xb = 1 e xc = 1. Desta maneira se verifica que tal experimento foi composto pelos fatores número de usuários, número de requisições de dados e número de *hosts* em seu primeiro nível, ou seja, 45 usuários, 8100

requisições e 3 *hosts*. A variável y1 representa a resposta deste primeiro experimento.

Quadro 4 – Níveis, fatores e respostas de um projeto 2^3

	xa	xb	xc	
Experimentos	**Núm. Usuários**	**Num. Requisições**	**Núm. *Hosts***	**Respostas (Y)**
1	1	1	1	**y1**
2	1	1	-1	**y2**
3	1	-1	1	**y3**
4	1	-1	-1	**y4**
5	-1	1	1	**y5**
6	-1	1	-1	**y6**
7	-1	-1	1	**y7**
8	-1	-1	-1	**y8**

Fonte: Elaborado pelo autor

Quadro 5 – Quadro de combinações dos níveis dos fatores

	Quadro de combinações							
Experimentos	**qa**	**qb**	**qc**	**qab**	**qac**	**qbc**	**qabc**	**Resposta (Y)**
1	1	1	1	1	1	1	1	**y1**
2	1	1	-1	1	-1	-1	-1	**y2**
3	1	-1	1	-1	1	-1	-1	**y3**
4	1	-1	-1	-1	-1	1	1	**y4**
5	-1	1	1	-1	-1	1	-1	**y5**
6	-1	1	-1	-1	1	-1	1	**y6**
7	-1	-1	1	1	-1	-1	1	**y7**
8	-1	-1	-1	1	1	1	-1	**y8**

Fonte: Elaborado pelo autor

Partindo do Quadro 4 elaborou-se o Quadro 5, que representa as combinações dos níveis de cada fator. Isso se faz necessário para o entendimento dos fatores no ambiente de experimento.

5. Resultados

Para esta avaliação de desempenho foram realizadas 10 repetições para cada um dos oito experimentos. Com isto foi obtido um total de 80 testes. Para o conjunto de testes de cada experimento foi estabelecido o intervalo de confiança (IC) dos

resultados obtidos. A partir destes verificou-se que as variações dos intervalos foram baixos, garantindo a integridade dos experimentos.

Gráfico 1- Comparativo entre os valores médios de cada experimento
Fonte: Elaborado pelo autor

O Gráfico 1 foi construído com as médias obtidas dos valores do conjunto de testes de cada experimento. Neste também é possível visualizar os ICs de cada experimento, os quais estão localizados no topo de cada barra.

Através dos resultados obtidos, é feita uma análise de quanto cada fator influenciou na variável de resposta.

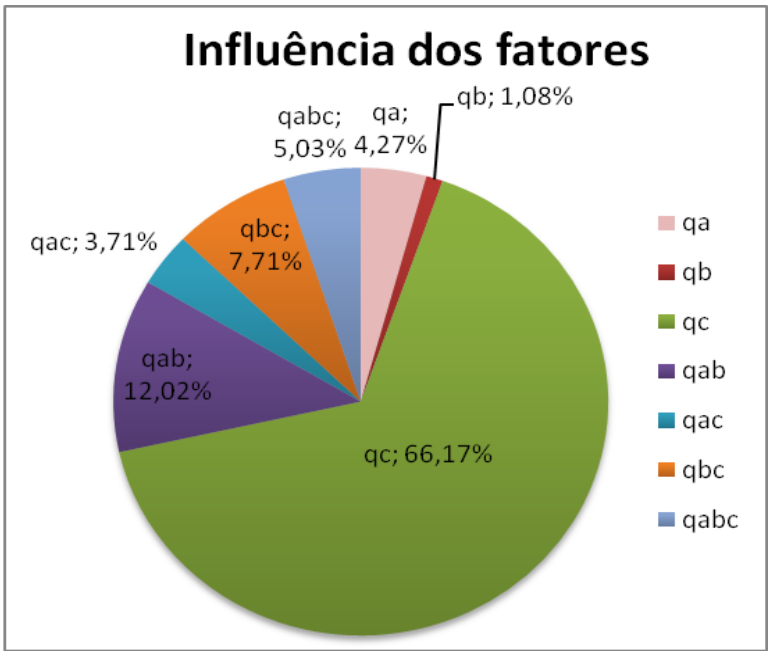

Gráfico 2 – Influência dos fatores
Fonte: Elaborado pelo autor

Observa-se no Gráfico 2 que o fator que teve mais influência na variável de resposta foi o número de *hosts* (qc). Verifica-se que este fator influenciou em quase 2/3 do

valor da variável de resposta, com 66,17%. Essa alta porcentagem de influência na variável de resposta faz sentido pois, em um ambiente de banco de dados distribuídos o qual foi configurado com sincronização de mesclagem, o que gera tráfego na rede deste tipo de banco são as máquinas (*hosts*) que o compõem. Desta forma, quanto maior o número de *hosts* que compõem este tipo de ambiente, maior o número de sincronizações e replicações de dados se fazem necessárias. Já o fator número de requisições (qb), quando analisado de forma isolada, foi o que menos influêncou a variável de resposta, com 1,08% do valor total. O fator número de usuários (qa) influênciou 4,27% do total da variável de resposta.

O segundo fator que mais influênciou a variável de resposta foi a interação entre os fatores número de usuários (qa) e número de requisições (qb), formando o fator (qab). Isso demonstra que após certo instante, o valor de pacotes por segundo é influenciado pela carga vinda do ambiente externo ao BDD, ou seja, sofre influência da carga que é formada pela quantidade de requisições enviadas pelos usuários deste bd.

6. Considerações Finais

Estudos se fazem necessários em todas as áreas de tecnologia, ou para se criar algo novo, ou para aperfeiçoar o que já foi criado ou até mesmo para reinventar algo que já existe. A área de bancos de dados distribuídos é relativamente nova no mundo da computação e necessita de muitas pesquisas para aperfeiçoá-la. Com as suas várias opções de configurações, tais como replicação mesclada ou replicação instantânea, neste ambiente de armazenamento de dados existe a necessidade de serem realizadas várias análises e experimentos que demonstrem a quem for implantar e gerir um BDD, quais as melhores opções de configuração para o ambiente que se quer montar.

Este trabalho focou em demonstrar quais dos 3 fatores (número de usuários, número de requisições e número de *hosts*) e as variações de seus níveis influenciariam mais no tráfego de dados em um ambiente de banco de dados distribuído. Desta maneira, avaliou-se os resultados obtidos e viu-se que o fator número de *hosts* é o fator que teve mais influência no tráfego de dados, com um percentual de 66,17% do total. Percebe-se que este é o fator mais influente tanto quando analisa-se os fatores de modo individual ou quando analisa-se os fatores quando ocorre interações entre eles. Entretanto, o segundo fator que mais

influenciou no tráfego foi a interação entre os fatores número de usuários com número de requisições, com 12,02% do total. Na sequência, o fator quem mais influenciou a transição de pacotes pela rede do BDD foi a interação entre os fatores número de usuários e número de *hosts*, obtendo um valor de 7,71%.

Desta forma conclui-se que ao se projetar um banco de dados distribuído, deve-se perceber que o fator número de *hosts* influenciará e muito no tráfego de pacotes, ao contrário do que se possa pensar, que os fatores número de usuários e número de requisições poderiam ser os que mais influenciariam em um ambiente deste. Porém, a combinação destes dois últimos fatores citados é o segundo fator que mais influencia. Então, uma elaboração projetual deve focar em uma boa estruturação na rede deste tipo de ambiente para que não haja sobrecarga neste local, pois existe uma troca muito intensa de dados entre os próprios bancos, para que estes contenham os mesmos dados. Só então se deve pensar nas requisições vindas do ambiente externo, emitidas pelos usuários.

Vale lembrar que o ambiente de testes utilizado neste trabalho fez uso de máquinas virtuais e, com isso, o desempenho fica bem reduzido se comparado a um mesmo ambiente, só que fazendo uso de máquinas físicas.

Fontes Consultadas

ELMASRI, R.; NAVATHE, S. B. **Sistemas de Banco de Dados.** Tradução Marília Guimarães Pinheiro, Cláudio César Canhette, Glenda Cristina Valim Melo, Claudia Vicei Amadeu e Rinaldo Macedo Morais. 4. ed. São Paulo: Pearson Addison Wesley, 2005.

JAIN, R**. Art of Computer Systems Performance Analysis Techniques For Experimental Design Measurements Simulation And Modeling.** New York: Wiley, 1991.

JUNIOR, M. F; OLIVEIRA, R. L. **Uso da técnica de deduplicação para armazenamento de dados biológicos em storage.** 2011. Monografia (Bacharelado em Ciência da Computação) – Departamento de Ciência da Computação, Universidade de Brasilia, Brasilia.

SHIBAYAMA, E. T. **Estudo Comparativo entre Bancos de Dados Distribuídos.** 2004. TCC (Bacharelado em Ciência da Computação) – Departamento de Computação, Universidade Estadual de Londrina, Londrina.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. **Sistema de Banco de Dados.** Tradução Daniel Vieira. 5. ed. Rio de Janeiro: Elsevier , 2006.