UNIVERSIDADE FEEVALE

CLAUDIOMIRO MACHADO BRITES

Banco de Dados NoSQL – Conceitos e Aplicabilidade

(Título Provisório)

Anteprojeto de Trabalho de Conclusão

Novo Hamburgo
2012

CLAUDIOMIRO MACHADO BRITES

Banco de Dados NoSQL – Conceitos e Aplicabilidade

(Título Provisório)

Anteprojeto de Trabalho de Conclusão de Curso, apresentado como requisito parcial à obtenção do grau de Bacharel em Sistemas de Informação pela Universidade Feevale

Orientador: Edvar Bergmann Araujo

Novo Hamburgo
2012

RESUMO

O crescimento da quantidade de dados e informações na web é perceptível nos dias de hoje. Entretanto, se as aplicações possuem um grande volume de dados é possível que se tenha problemas com infra-estrutura. O NoSQL surgiu com o propósito de ajudar na solução desse problema, mostrando uma abordagem diferente de persistência de dados, baseada em disponibilidade, desempenho e escalabilidade dos dados. Grandes empresas mundiais já utilizam esta tecnologia, tais como Google, Twitter, Facebook, entre outras. Porém a diversidade de bancos NoSQL e variedade de bancos de cada categoria, dificulta a escolha de qual ferramenta utilizar e de quando utilizá-la, Esse trabalho tem como objetivo estudar os bancos de dados NoSQL, suas características, vantagens, desvantagens e aplicações práticas, com o intuito de fornecer mais subsídios para a escolha de qual produto utilizar em determinadas situações. Para tanto pretende-se identificar alguns cenários de utilização de bancos NoSQL e colocar tais cenários em prática, verificando o comportamento e os resultados obtidos em cada um deles.

Palavras-chave: NoSQL. Banco de dados. Sistema distribuído. Escalabilidade. Disponibilidade.

SUMÁRIO

MOTIVAÇÃO

NoSQL é um termo genérico para uma classe de banco de dados não-relacional que apresenta uma alternativa aos bancos de dados relacionais. Ao invés de oferecer a propriedade ACID (*Atomicity*, *Consistency*, *Isolation*, *Durability*), oferece a propriedade BASE (*Basically Available*, *Soft state*, *Eventual consistency*). Esta propriedade significa: basicamente estar disponível, ou seja, o sistema parece estar funcionando o tempo todo; em estado leve, pois o sistema não precisa estar consistente o tempo todo; e eventualmente consistente, o sistema torna-se consistente no momento devido (BRITO, 2010, p.5).

Os bancos de dados NoSQL apresentam características interessantes como alta performance, escalabilidade, replicação, suporte à dados estruturados e sub colunas. Caracterizam-se também por facilitar o trabalho da equipe de desenvolvimento de software, já que não impõem a estrutura de dados rígida, imposta pelos bancos relacionais. Mais especificamente essa classe surgiu da necessidade de uma melhor performance e fácil escalabilidade em sistemas web, que precisam manter grandes volumes de dados e um escalonamento frequente (IMASTER, 2010).

A escalabilidade pode ser alcançada em um banco de dados relacional. Entretanto, essa tarefa pode ter um custo elevado e se tornar complexa. Quando é necessário o aumento de infraestrutura para um banco de dados é normal usar como primeiro recurso distribuição vertical de servidores (*scale up*), ou seja, quanto mais dados, maior o poder de processamento como memória, processador e disco. Essa pode ser uma solução excelente se tratando de curto prazo. Contudo, no futuro o problema da escalabilidade pode voltar a aparecer, uma vez que o hardware possui limitações físicas e tecnológicas (ESCALABILIDADE, 2010).

Outra solução é adotar a distribuição horizontal (*scale out*), isto é, quanto mais dados, mais servidores, não necessariamente de alta capacidade de processamento. Entretanto, isso pode acarretar em outro problema, pois não é uma tarefa fácil adicionar mais um servidor em um sistema distribuído, já que o processo normalmente é de alta complexidade, demorado e de custo elevado (ESCALABILIDADE, 2010).

Outro fator que contribui pela alternativa de um banco de dados NoSQL ao invés de um banco relacional está relacionado à estrutura de dados pré-definida. Enquanto um banco relacional exige essa estrutura, os bancos NoSQL lidam de forma eficiente com dados desestruturados, tais como processamento de arquivos texto, e-mail, multimídia, e meios de

comunicação social (LEAVITT, 2010). Por isso pode ser conveniente usar um banco de dados NoSQL em casos onde existam muitas modificações na estrutura dos dados.

A tecnologia NoSQL não tem como propósito a substituição total do modelo de dados relacional, mas sim de atender determinadas necessidades que se fazem mais adequadas. Uma prova disto é a possibilidade de acessar um banco de dados NoSQL e um banco de dados relacional em uma mesma aplicação (LEAVITT, 2010).

Brito (2010, p.3) classifica os bancos de dados NoSQL em quatro categorias:

> Quanto ao modelo de dados, existem quatro categorias básicas: os sistemas baseados em armazenamento chave-valor, como é o caso do *Amazon Dynamo*; os sistemas orientados a documentos, entre os quais temos o *CouchDB* e o *MongoDB*; os sistemas orientados a coluna, que tem como exemplos o *Cassandra* e o *BigTable*; e os sistemas baseados em grafos, como são os casos do *Neo4j* e do *InfoGrid*.

"Banco de dados de armazenamento chave-valor é tal como define o nome, que são recuperados a partir da chave relacionada. Estes sistemas podem armazenar dados estruturados ou não estruturados" (LEAVITT, 2010, p.13, tradução nossa).

Um exemplo desta categoria é o Dynamo, que surgiu da necessidade da Amazon.com de uma plataforma escalável e distribuída para o seu sistema de comércio eletrônico. Porém a Amazon, após um determinado período, não conseguiu uma evolução considerável além dos serviços essenciais de plataforma de comércio eletrônico, pois a concepção e implementação da solução haviam ficado muito complexas, já que e o mesmo teve a participação na construção de inúmeros arquitetos, e em vários serviços da plataforma (VOGELS, 2012).

Outro exemplo da Amazon é o *SimpleDB*, que é oferecido como serviço na *Amazon Web Service Solutions*, plataforma de computação em nuvens da Amazon, que consiste em um banco de dados com baixa complexidade operacional para desenvolvimento de aplicações web, porém com limitações, como 10GB por coleção. E em janeiro de 2012 foi disponibilizado o *DynamoDB*, que não impõe limite de armazenamento, com menor latência (AMAZON, 2012).

No armazenamento orientado à coluna, em vez de conjuntos de informações em uma tabela de colunas e linhas de tamanho uniforme no banco de dados relacional, os dados são organizados por uma trio (linha, coluna e *timestamp*), onde linhas e colunas são identificadas como chaves e o *timestamp* permite diferenciar múltiplas versões de um mesmo dado. Outro

conceito associado a esse modelo é o de família de colunas, que é usado com o intuito de agrupar colunas que armazenam os mesmos tipos de dados. (LÓSCIO, 2011).

Após seis anos de criação, o *Facebook* possui cerca de 3,5 bilhões de conteúdos (links, posts, etc) compartilhados por semana. Para evitar problemas com a escalabilidade e disponibilidade dos dados, a empresa desenvolveu o Cassandra, que inicialmente foi criado para otimização do sistema de busca do *Facebook*. Atualmente o Cassandra é utilizado para dar suporte à replicação, detecção de falhas, armazenamento em cache dentre outras funcionalidades. Em janeiro de 2009 o Cassandra tornou-se um projeto da *Apache Software Foundation*, vindo a ser utilizado por outras empresas como *Cisco*, *Digg*, *Twitter*, *Cloudkick* e *Reddit* (LÓSCIO, 2011).

Outro exemplo é o Bigtable, um banco de dados proprietário construído sobre o *Google File System*, sistema de armazenamento distribuído projetado para gerenciar um volume de dados muito grande. O Bigtable é usado por mais de 60 produtos e projetos do Google como indexação de páginas web, *Google Analytics*, *Google Finance*, *Orkut*, *Writely*, *Google Earth*, entre outros. Embora não tenha distribuição fora do Google, a empresa oferece acesso pelo *Google App Engine* para seus usuários (GOOGLE, 2008).

O modelo orientado a grafos possui três componentes básicos: os nós (são os vértices do grafo), os relacionamentos (são as arestas) e as propriedades (ou atributos) dos nós e relacionamentos. Neste caso, o banco de dados pode ser visto como um multigrafo rotulado e direcionado, onde cada par de nós pode ser conectado por mais de uma aresta (LÓSCIO, 2011). Esse banco de dados é mais adequado para informações mais complexas e altamente inter-relacionadas, como por exemplo aplicações de redes sociais. A relação entre amigos é dinâmica, e nessas aplicações evoluem rapidamente, o que constitui um desafio a um banco de dados relacional. O Neo4j é um exemplo de banco de dados de grafo, que inclusive mantém o benefício de transações do banco relacional (NEO4J, 2012).

Armazenamento orientado a documento organiza dados em coleções de documentos, ao invés de tabelas estruturadas com tamanho uniforme de campos para cada registro. Com estas bases de dados, os usuários podem adicionar qualquer número de campos de qualquer comprimento em um documento (LEAVITT, 2010, p.13).

MongoDB é um banco de dados orientado à documento, que armazena registros em um formato semelhante ao formato de objeto *JSON*(*JavaScript Object Notation*), com capacidade de armazenar e consultar atributos aninhados (WARDEN, 2011, p.6). Esse

formato é identificado como BSON(*Binary JSON*), que é uma serialização binária codificada de *JSON*. O MongoDB é um projeto de banco de dados de código aberto construído para escalabilidade e facilidade de utilização. "*Sharding* é a abordagem do MongoDB para escalar, que permite que você adicione mais máquinas para lidar com o aumento do tamanho e carga de dados sem afetar a sua aplicação" (CHODOROW; DIROLF, 2010, p.143).

Outro banco de dados NoSQL orientado à documento é o CouchDB, também da *Apache Software Foundation*. É um banco de dados de código aberto, escalável, escrito em *Erlang* e acessível de qualquer navegador (LEAVITT,2010,p.13).

O movimento NoSQL está transformando a forma com que as empresas lidam com seus dados, novas estruturas de dados, arquiteturas distribuídas e utilização intensiva da memória RAM.

A diversidade de bancos NoSQL e variedade de bancos de cada categoria, dificulta a escolha de qual solução utilizar e de quando utilizar. Sendo assim, esse trabalho tem como objetivo estudar os bancos de dados NoSQL, suas características, vantagens, desvantagens e aplicações práticas, objetivando fornecer mais subsídios para a escolha de qual produto utilizar em determinadas situações. Para tanto pretende-se identificar alguns cenários de utilização de bancos NoSQL e colocar tais cenários em prática, verificando o comportamento e os resultados obtidos em cada um deles.

OBJETIVOS

Objetivo geral

O objetivo principal deste trabalho é estudar e validar a aplicabilidade de bancos de dados NoSQL em cenários práticos, visando contribuir com informações a respeito das situações em que estes bancos apresentam bons resultados e podem ser uma alternativa interessante em relação aos bancos de dados relacionais.

Objetivos específicos

- Apresentar conceitos importantes dos bancos de dados NoSQL;
- Explorar os principais bancos de dados NoSQL;
- Promover experimentos e simulações;
- Identificar diferenças entre bancos de mesma categoria, como por exemplo MongoDB e CouchDB;
- Verificar a aplicabilidade dos bancos de dados NoSQL;
- Definir e implementar cases/cenários de aplicação;
- Validar resultados obtidos.

METODOLOGIA

A proposta desse trabalho está sustentada numa pesquisa de natureza prática de objetivo exploratório seguido de estudos de casos, com a finalidade de proporcionar mais subsídios para a escolha de base de dados de sistemas computacionais. O inicio se dará à partir de pesquisas bibliográficas de livros, tais como de Warden (2011), Chorodow (2010, 2011), Hewitt (2011), Anderson (2010); artigos relevantes sobre o assunto, como os de Leavitt (2010), Stonebraker (2010), Cattell (2010); além de trabalhos de mestrado e doutorado da área sobre o assunto. Essa leitura permitirá uma visão clara e objetiva dos conceitos relacionados na formação do embasamento teórico, e consequentemente na construção dos conhecimentos necessários para identificar a aplicabilidade da tecnologia em questão.

Em seguida será feita a escolha de bancos de dados NoSQL para instalar e executar experimentos para identificar características de cada um deles. Considerando que existem quatro categorias de bancos de dados NoSQL, será adotada a estratégia de escolher o banco de dados de maior popularidade dentro de cada categoria. E a outra estratégia, será o de conter no mínimo um banco de dados de cada categoria.

Após estes experimentos citados, será escolhido um banco de dados NoSQL, onde serão realizadas simulações através de uma aplicação, *Javascript* ou outra linguagem a ser definida. Através destes testes será possível verificar diferenças, benefícios, deficiências, vantagens e desvantagens entre os bancos de dados, e identificar cenários em que o banco de dado NoSQL escolhido possa ser mais eficiente em comparação ao banco de dados relacional.

A validação dos resultados será levado em consideração o nível de complexidade na instalação do banco de dados, operação no ambiente, capacidade de escalabilidade e performance de inserção e recuperação de dados.

CRONOGRAMA

Trabalho de Conclusão I

Etapa	Meses			
	Mar	Abr	Mai	Jun
Levantamento bibliográfico				
Redação do Anteprojeto				
Estudar os bancos de dados NoSQL				
Escolher bancos de dados de cada categoria				
Fazer experimentos nos bancos de dados NoSQL				
Redação TC I				

Trabalho de Conclusão II

Etapa	Meses			
	Ago	Set	Out	Nov
Verificar aplicabilidades dos bancos NoSQL				
Definir cenário para a aplicação no banco orientado a documento				
Desenvolvimento da aplicação para o cenário escolhido				
Validar resultados obtidos				
Redação TC II				

BIBLIOGRAFIA

AMAZON. **Amazon DynamoDB (beta)**. Disponível em: <http://aws.amazon.com/pt/dynamodb/>. Acesso em: 25 mar 2012.

AMAZON. **Amazon SimpleDB (beta)**. Disponível em: <http://aws.amazon.com/pt/simpledb/>. Acesso em: 25 mar 2012.

ANDERSON, J. Chris; LEHNARDT, Jan; SLATER, Noah. **CouchDB. The Definitive Guide**. USA: O'Reilly Media, jan 2010.

BRITO, Ricardo W. **Bancos de Dados NoSQL x SGBDs Relacionais: Análise Comparativa**. Disponível em: <http://www.infobrasil.inf.br/userfiles/27-05-S4-1-68840-Bancos%20de%20Dados%20NoSQL.pdf>. Acesso em: 31 mar 2012.

BSON. **Binary JSON**. Disponível em: <http://bsonspec.org/>. Acesso em: 25 mar 2012.

CATTELL, Rick. Scalable SQL and NoSQL Data Stores. **ACM SIGMOD Record**, New York, NY, USA, v.39, n.4, dez. 2010.

CHORODOW, Kristina. **Scaling MongoDB**. USA: O'Reilly Media, fev 2011.

CHORODOW, Kristina; DIROLF, Michael. **MongoDB. The Definitive Guide**. USA: O'Reilly Media, set 2010.

ESCALABILIDADE. **Guia Rápido Para os Serviços da Amazon Web Services (AWS)**. Disponível em: <http://escalabilidade.com/2010/03/25/guia-rapido-para-os-servicos-da-amazon-web-services-aws/>. Acesso em: 01 mai 2012.

ESCALABILIDADE. **Introdução ao NoSQL parte I**. Disponível em: <http://escalabilidade.com/2010/03/08/introducao-ao-nosql-parte-i/>. Acesso em: 25 abr 2012.

GOOGLE, Inc. Bigtable: A Distributed Storage System for Structured Data. **ACM Transactions on Computer Systems (TOCS)**,New York, NY, USA, v.26 n.2, p.1-26, jun. 2008.

HEWITT, Eben. **Cassandra. The Definitive Guide**. USA: O'Reilly Media, nov 2011.

IMASTER. **NoSQL - você realmente sabe do que estamos falando?** Disponível em: <http://imasters.com.br/artigo/17043/banco-de-dados/nosql-voce-realmente-sabe-do-que-estamos-falando>. Acesso em: 03 mar 2012.

JSON. **Javascript Object Notation**. Disponível em: <http://www.json.org//>. Acesso em: 25 março 2012.

LEAVITT, Neal. Will NoSQL Databases Live Up to their Promise? **IEEE Computer Society Press**, Los Alamitos, CA, USA, v. 43, n. 2, p. 12-14, fev. 2010.

LÓSCIO, Bernadette Farias. **NoSQL no desenvolvimento de aplicações Web colaborativas** Disponível em: <http://www.addlabs.uff.br/sbsc_site/SBSC2011_NoSQL.pdf>. Acesso em: 25 mar 2012.

NEO4J. **Neo4j: The World's Leading Graph Database**. Disponível em: <http://neo4j.org/>. Acesso em: 25 mar 2012.

PROCELLI, Alexandre. **O que é NoSQL**. São Paulo: Java Magazine 86. ed., p. 21-31, jun 2011.

STONEBRAKER, Michael. SQL databases v. NoSQL databases, **Communications of the ACM**, New York, NY, USA, v.53 n.4, abr. 2010.

VOGELS, Werner. **All Things Distributed - Amazon DynamoDB – a Fast and Scalable NoSQL Database Service Designed for Internet Scale Applications**. Disponível em: <http://www.allthingsdistributed.com/2012/01/amazon-dynamodb.html>Acesso em: 28 abr 2012.

WARDEN, Pete. **Big Data Glossary**. USA: O'Reilly Media, 2011.