Armazenamento de Dados Abertos com NoSQL: Um estudo de caso com Dados do Bolsa Família e NoSQL Cassandra

Jorge Luiz Andrade

Universidade de Brasília

jorgeluizandrade@outlook.com

4 de dezembro de 2017

Introdução

Bancos de Dados Relacionais
A aprendizagem de lógicas e métodos de programação é fundamental para cursos relacionados a computação. Entretanto, ela não é trivial.

Introdução

Introdução

- Modelagem gráfica usada como catalisador de ensino
  - Despertar ou aumentar interesse
  - Diversas universidades recorrem a computação gráfica

Introdução

Problema

Banco de Dados Relacionais podem não apresentar um desempenho satisfatório ao operar grandes volumes de dados

Introdução

Problema

Banco de Dados Relacionais podem não apresentar um desempenho satisfatório ao operar grandes volumes de dados

Hipótese

O uso de múltiplas máquinas em um ambiente Cassandra distribuído pode oferecer um melhora do desempenho que justifique sua utilização na análise de dados abertos.

Introdução

Objetivos

Comparar o desempenho de um banco Cassandra para inserções e consultas em diferentes tamanhos de *cluster* e de volumes de dados;

- Desenvolver uma aplicação para inserção e busca dos dados do Bolsa Família;
- Realizar testes de inserção e busca com diferentes configurações;
- Comparar o desempenho do Cassandra nas diferentes situações;

Bancos Relacionais

- Proposto em 1970 por Edgar Codd;
- Conjunto de relações entre tuplas;

Propriedades ACID

- Atomicidade;
- Consistência;
- Isolamento;
- Durabilidade;

Garantem a validade do esquema, mas sacrificam desempenho e disponibilidade.

Normalização

- **1ª Forma Normal**: Cada campo possui apenas valores atômicos;
- **2ª Forma Normal**: Cada atributo não-chave é dependente da totalidade da chave primária;
- **3ª Forma Normal**: Cada atributo não-chave não depente de um outro atributo não-chave;

NoSQL

- Termo utilizado pela primeira vez em 1998(Strozzi NoSQL)
- Google Bigtable(2006) e Amazon's Dynamo(2007)

Teorema CAP

- Proposto em 2000 por Eric Brewer, define limitações em sistemas distribuídos;
- Revisado em 2012;
- Consistência;
- Disponibilidade;
- Tolerância a partições

Chave-Valor

Consiste em uma tabela *hash*, com consultas a um valor a partir de uma chave;

- Berkeley DB;
- Amazon DynamoDB;

Documentos

Acesso à um documento de esquema flexível a partir de uma chave;

- CouchDB;
- MongoDB;

Resultados

Comparação do aumento do número de máquinas:

- Melhora média de 7,5% na inserção dos dados;
- Melhora média de 56,53% na busca dos dados;

Trabalhos futuros

- Isolamento da rede no ambiente utilizado;
- Comparação com outros bancos;
- Implementar diferentes modelagens no banco Cassandra;

Bibliografia