PÓS-GRADUAÇÃO EM ENGENHARIA DE DADOS

IEC PUC MINAS

Disciplina:

BANCOS DE DADOS NoSQL

Professor: Augusto Zadra

3.2 – Web Crawler

INTRODUÇÃO

- Web Crawler é um algoritmo que percorre os hiperlinks os indexa.
- Grava as informações em bancos de dados e as utiliza para gerar *insights* ou classificar os dados encontrados.
- São conhecidos também como Spiders, Robots (bots), Wanderers.
- Possui muitas finalidades, como análise da Web, coleta de endereços de e-mail, embora a principal delas seja a descoberta e indexação de páginas para mecanismos de busca na Web.

INTRODUÇÃO

- Os principais utilitários do Web Crawler incluem:
  - ✓ Execução de mineração de dados que é a primeira etapa do processo de mineração de dados da Web.
  - ✓ Reunir páginas da Web baixando documentos automaticamente de um servidor Web.
  - ✓ Analisar documentos recuperados de um servidor Web e enviar dados de volta para um banco de dados de mecanismo de pesquisa.

INTRODUÇÃO

- ✓ Apoiar um motor de pesquisa.
- ✓ Mecanismos de busca, análise de dados, interações automatizadas da Web, espelhamento e validação de HTML/link onde quando um rastreador da Web visita uma página, ele lê o texto visível, os hiperlinks e o conteúdo das várias *tags* usadas no site, como meta *tags* com muitas palavras-chave.

PUC Minas

Bancos de dados NoSQL

Web Crawlers

- Então pensemos: Qual seria o banco de dados ideal para armazenarmos tantas informações variáveis?
- As coletas são armazenadas geralmente em bancos NoSQL.
- Veja as características:
  - ✓ Distribuído: o que aumenta a disponibilidade.

PUC Minas

Bancos de dados NoSQL

Web Crawlers

- ✓ Facilidade de desenvolvimento de novas APIs.
- ✓ Escalável: Resolvendo o problema da utilização da banda.
- ✓ Schemaless: possibilidade de fazermos a estrutura livre de travas possibilitando armazenar diversos documentos com configurações diferentes.

PUC Minas

Bancos de dados NoSQL

Hadoop

- Projeto Apache Hadoop:
- https://hadoop.apache.org/

- Este projeto é um Framework , ou seja, um conjunto de projetos de softwares relacionados voltados para uma infraestrutura de computação distribuída com foco em processamento de dados em larga escala.

PUC Minas

Bancos de dados NoSQL

Hadoop

- Termo mais conhecido no meio em relação aos componentes do Hadoop é ECOSSISTEMA HADOOP.
- Ele é composto de dois componentes básicos:
  - ✓ HDFS: Sistema de arquivos escalável e tolerante a falhas, baseado em Java com alta capacidade de armazenamento de forma distribuída.
  - ✓ Map Reduce : Mecanismo ou modelo de programação para processamento de dados.

PUC Minas

Bancos de dados NoSQL

Hadoop

Map Reduce

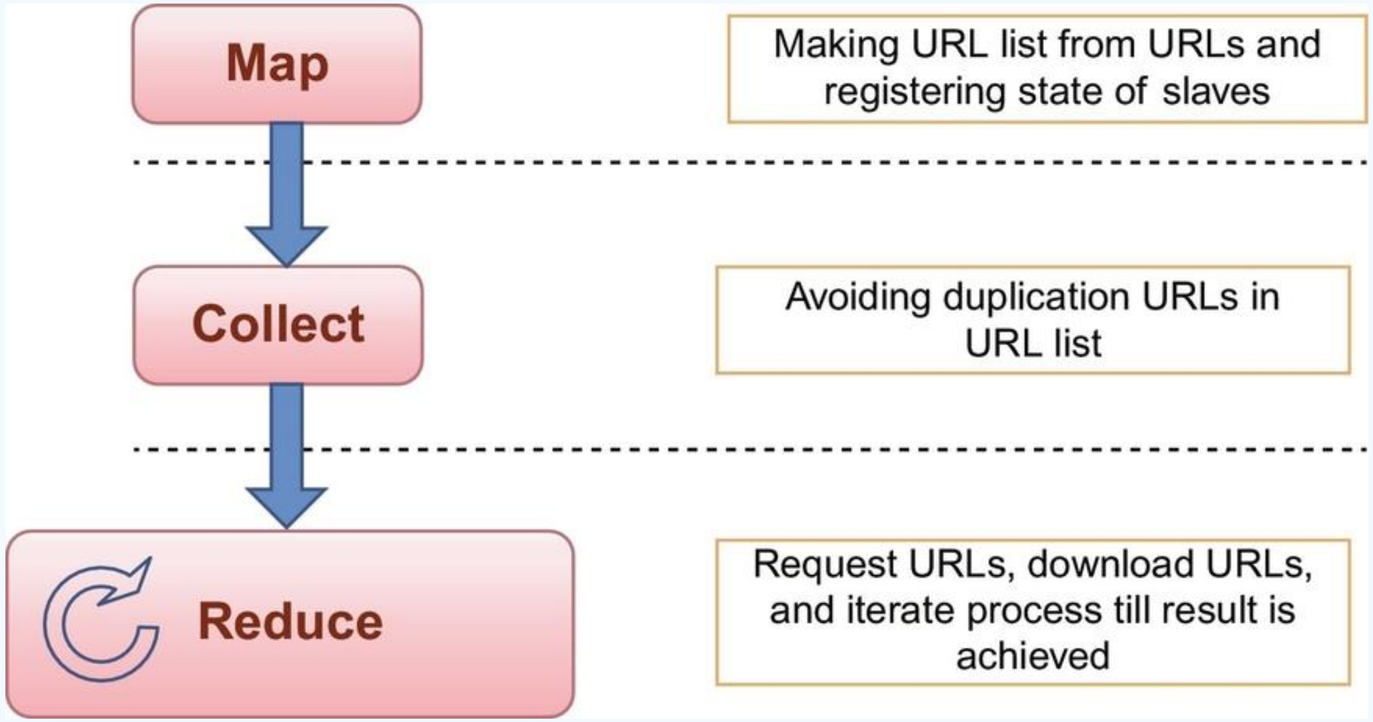
- **Map Reduce** na prática, é um processo disparado pelo Hadoop que transforma os dados.
- Ele pega um grande conjunto de dados e gera uma nova massa bem menor já consolidada que é gerada através de um conjunto de programas.
- Este modelo é utilizado para processamentos em larga escala e existem vários projetos de fiscalização inclusive do portal de dados abertos.

PUC Minas

Bancos de dados NoSQL

Web Crawlers

- Voltando então ao assunto dos Crawlers, imagine então todo este volume de dados de scanner de sites e armazenamento destas informações.
- **Quanto isto seria custoso?**



PUC Minas

Bancos de dados NoSQL

Web Crawlers

- O modelo do Map Reduce é o ideal para este tipo de acesso.
- Mas a pergunta é qual o banco de dados NoSQL está relacionado com a nossa conversa?

- **Cassandra** é um banco de dados NoSQL que faz parte do framework Hadoop.
- Ele é um par do banco de dados MongoDB sendo classificado como banco de dados de chave/valor.

PUC Minas

REFERÊNCIAS BIBLIOGRÁFICAS

DIANA, M. de;; GEROSA, M. A. NOSQL na Web 2.0: Um estudo comparativo de bancos não-relacionais para armazenamento de dados web 2.0. In: Workshop de Teses e Dissertações em BD - WTDB, 9., 2010, Belo Horizonte. Anais... Belo Horizonte: SBC, 2010.

LI, Y.;; MANOHARAN, S. A performance comparison of SQL and NoSQL databases. In: IEEE Pacific Rim Conference on Communications, Computers and Signal Processing - PACRIM, 14., 2013, Victoria, B.C., Canadá. Proceedings... IEEE, 2013.

LÓSCIO, B. F.;; OLIVEIRA, H. R. de;; PONTES, J. C. de S. NoSQL no desenvolvimento de aplicações web colaborativas. In: Simpósio Brasileiro de Sistemas Colaborativos – SBSC, 8., 2011, Paraty (RJ). Anais... Paraty: SBC, 2011.

VIEIRA, M. R. et al. Bancos de Dados NoSQL: conceitos, ferramentas, linguagens e estudos de casos no contexto de Big Data. In: Simpósio Brasileiro de Bancos de Dados, 27., 2012, São Paulo. Anais... São Paulo: SBC, 2012.