Teoria

Estudo de caso: GOOGLE

Contribuições: Prof. Chris Lima

Professor:
Vitor Figueiredo

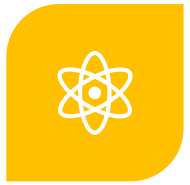
Disciplina
Sistemas Distribuídos

Versão:
2.0

O que vamos ver?

HISTÓRIA;

WEB CRAWLERS, INDEXAÇÃO E PAGERANKING;



MODELO FÍSICO;

VISÃO GERAL DA ARQUITETURA GOOGLE;

PARADIGMAS DE COMUNICAÇÃO;

ARMAZENAMENTO DE DADOS;

SERVIÇOS DE COMPUTAÇÃO DISTRIBUÍDA;

2

História

- Fundada em 1998 por Larry Page e Sergey Brin;
- Missão: "Organizar as informações do mundo e torná-las mundialmente acessíveis e úteis";
- Surgiu na universidade de Stanford, a partir de um projeto de Doutorado sobre PageRank (determinar a importância de páginas individuais);
- Google é um trocadilho para o termo googol (que significa 10^{100});

História

- A infraestrutura do Google provê serviços com requisitos extremamente elevados em termos de:
  - Escalabilidade;
  - Confiabilidade;
  - Desempenho;
  - Segurança;
  - Abertura de Sistema;
  - E várias outras características da disciplina;
- Exemplo (2020):
  - 100 bilhões de pesquisas/mês == 38.600 pesquisas por segundo;
  - Motor principal de busca nunca sofreu queda de energia;
  - Cada pesquisa retornava o resultado esperado na média de **0.2 segundos**

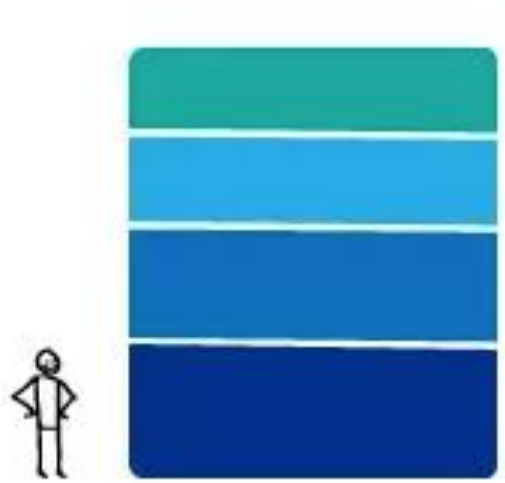
Indexação

- A Web é como se fosse uma enorme biblioteca com bilhões de livros, porém, sem um índice de pesquisa;
- O Google normalmente reúne milhares de páginas durante o processo de Crawling e então cria os chamados índices (indexes), a fim de saber exatamente como encontrar um conteúdo na Web;
- Quando um usuário pesquisa por algo, o algoritmo da Google relaciona os termos inseridos e pesquisa no índice para retornar as páginas apropriadas;

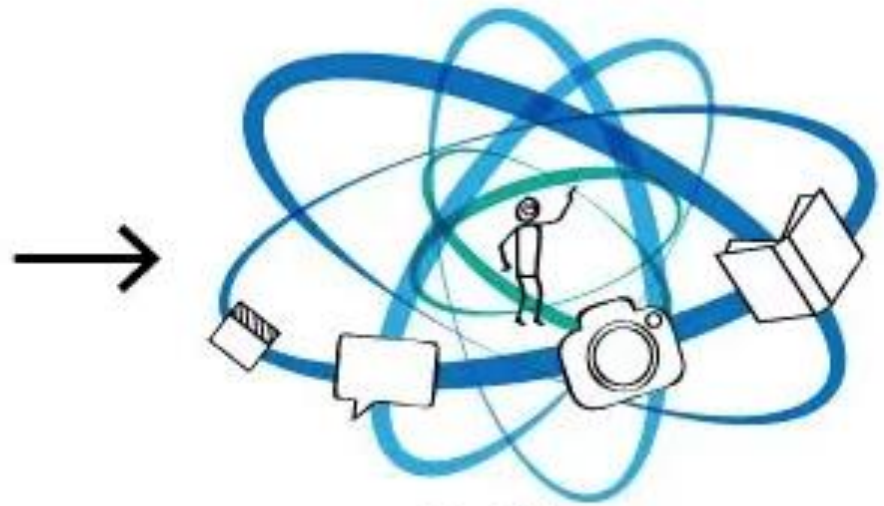
Web Crawlers

- A tarefa de um Web Crawler (ou Web Spider) é localizar e recuperar todo tipo de conteúdo na Web, mais especificamente na "Surface Web";
- No caso do Google, um crawler é chamado de **GoogleBot**, que tem a função de ler recursivamente uma dada página Web, rastreando e colhendo todos os links apontados por esta página sucessivamente; Essa técnica é conhecida como **Deep Searching**;
- **Caffeine**- o mecanismo de Crawling da Google lançado em 2009. Maior velocidade e melhor indexação de dados. Caffeine Foi atualizado em 2017 para ficar ainda mais eficiente.

Web Crawlers – Google Caffeine

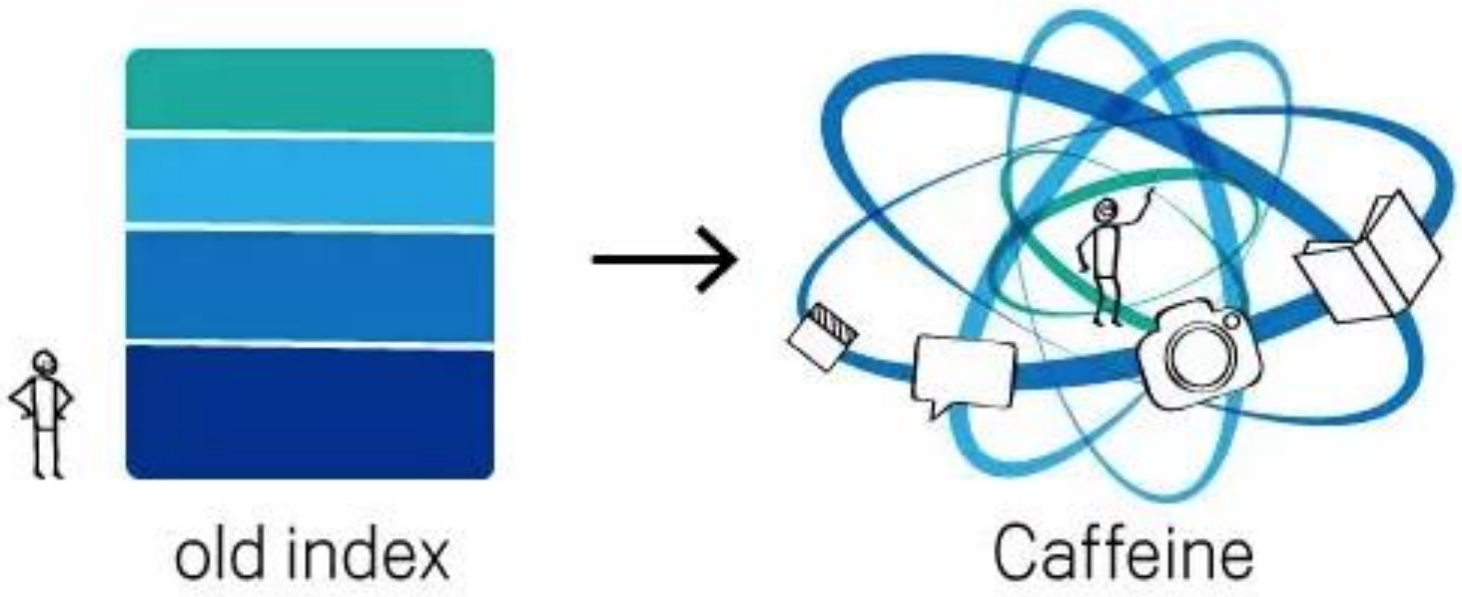


old index



Caffeine

Web Crawlers – Google Caffeine



Em 2013, eles indexaram mais de 30 trilhões de páginas na web - e em 4 anos, esse número cresceu para mais de 130 trilhões de páginas.

PageRanking

- Mesmo que grande parte das páginas presentes na Web sejam indexadas, como descobrir quais páginas são as mais relevantes/importantes?
- Nos algoritmos de **PageRanking**, páginas são marcadas como mais importantes se são referenciadas por muitos outros sites (links). A importância da página que referencia também é medida;
- Outros fatores também são levados em considerações pelo algoritmo Page Rank:
  - Palavras-chave buscadas;
  - Importância do termo no texto;
  - Fontes minúsculas ou maiúsculas;
  - Entre vários outros fatores;

Modelo Físico

- Até o momento, falamos do Google apenas como uma ferramenta de busca. Mas não devemos nos esquecer que vários outros serviços, como serviços do tipo SaaS também são oferecidos pela Google, são eles:

Modelo Físico

Application	*Description*
Gmail	Mail system with messages hosted by Google but desktop-like message management.
Google Docs	Web-based office suite supporting shared editing of documents held on Google servers.
Google Sites	Wiki-like web sites with shared editing facilities.
Google Talk	Supports instant text messaging and Voice over IP.
Google Calendar	Web-based calendar with all data hosted on Google servers.
Google Wave	Collaboration tool integrating email, instant messaging, wikis and social networks.
Google News	Fully automated news aggregator site.
Google Maps	Scalable web-based world map including high-resolution imagery and unlimited user-generated overlays.
Google Earth	Scalable near-3D view of the globe with unlimited user-generated overlays.
Google App Engine	Google distributed infrastructure made available to outside parties as a service (platform as a service).

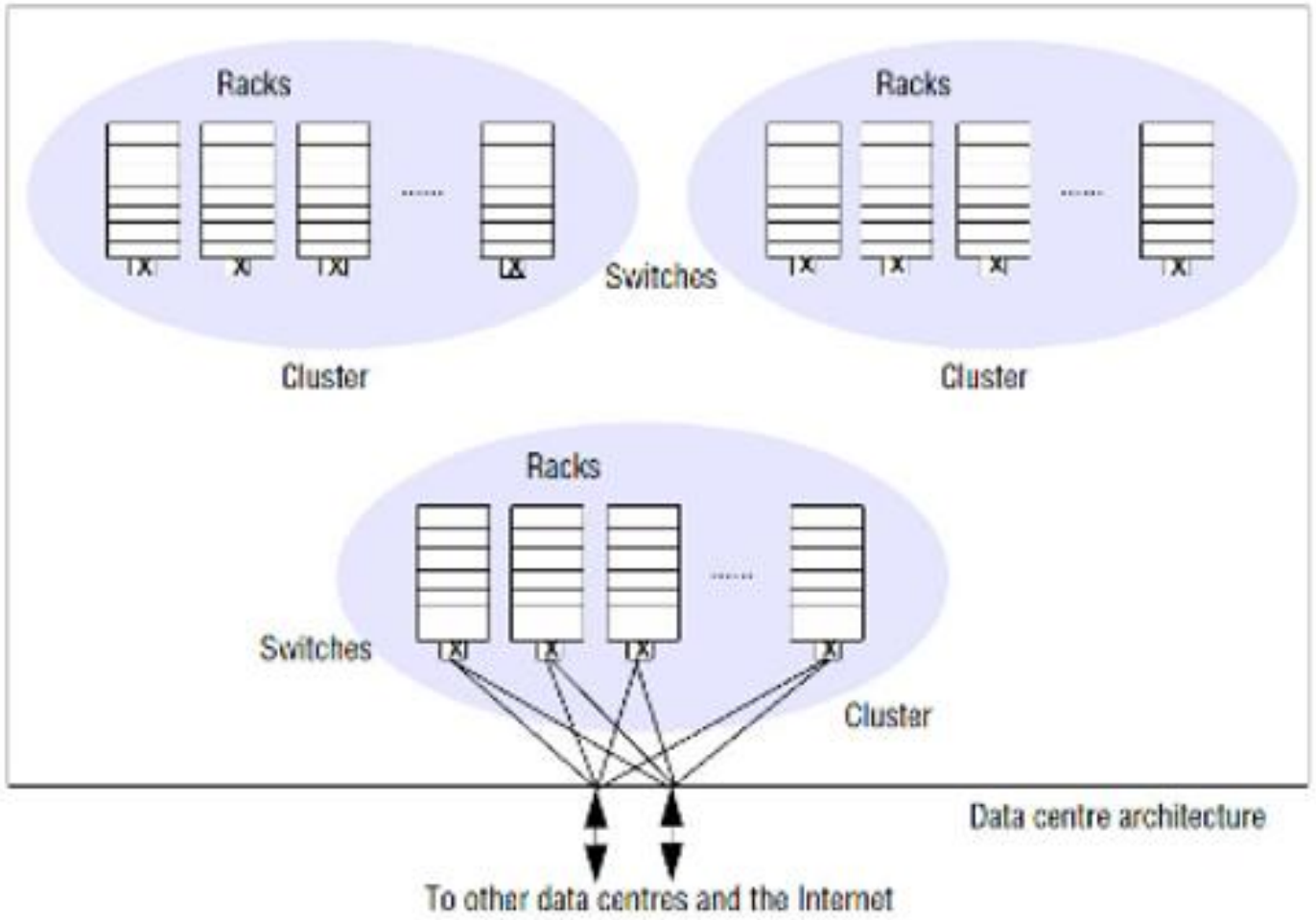
E onde a Google coloca tudo isso???

Modelo Físico – Filosofia

- Filosofia utilizada pela Google: ao invés de utilizar computadores gigantescos, utiliza-se um número muito alto de computadores comuns para produzir um ambiente de processamento e armazenamento mais efetivo.
- Um típico PC da arquitetura Google:
  - 2 Tb de disco;
  - 16 Gb de RAM;
  - Sistema Operacional Linux;

Modelo Físico – Organização

- Os computadores são organizados em Racks de 40 a 80 computadores;
- Cada Rack possui um Switch Ethernet;
- Racks são organizados em Clusters (ou Containers), que são tratados como unidades mínimas de gerenciamento pela Google;
- Cada Cluster possui 30 ou mais Racks e 2 Switches com links de alta largura de banda para conexão com o mundo externo;
- Clusters são hospedados em Datacenters da Google ao redor do mundo.
- O primeiro Datacenter da Google no Brasil foi instalado em São Paulo em set/2017.



Modelo Físico – Organização

Modelo Físico – Organização

- A Google mantém segredo de quantas máquinas possui; Segundo Coulouris et al. (2013), estima-se:
  - 1 PC: 2 Terabytes;
  - 1 Rack de 80 PC's: 160 Terabytes;
  - 1 Cluster de 30 Racks: 4.8 Petabytes;
  - Estima-se que a Google possua 200 Clusters: **960 Petabytes** (Capacidade de armazenamento perto de 1 Exabyte);

Visão Geral da Arquitetura Google

Princípios Chaves do Google

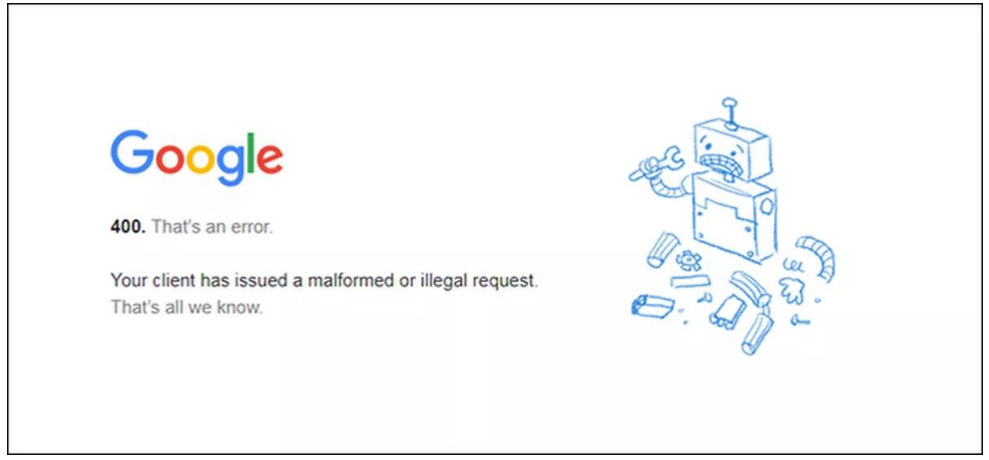
- **Princípio 1**: Simplicidade; Um componente de software deve fazer apenas uma coisa, e fazê-lo da melhor forma possível;
- **Princípio 2**: Forte ênfase em desempenho de operações primitivas -> *Envio de pacotes pela rede, acesso a disco, bloqueio e liberação de regiões críticas, entre outras*;
- **Princípio 3**: Teste rigorosos de softwares. Se baseia no seguinte lema: "Se não foi encontrado bug no software, ele não foi testado apropriadamente"

Visão Geral da Arquitetura Google

- **Escalabilidade**
  - Primeira e mais importante exigência adotada pela Google;
  - O Google é um sistema distribuído considerado ULS (Ultra-Large Scale), ou seja, um sistema que possui uma grande quantidade de Hardware e Software heterogêneos.
  - Usado por diferentes tipos de usuários e possui um volume de dados muito grande;

Visão Geral da Arquitetura Google

- Confiabilidade
  - Mecanismo de busca 24/7 no ar;
  - 99.9% de disponibilidade para clientes Google Corporativo (Clientes que pagam).
  - Você já presenciou um serviço do google fora do ar?
  - **Caso: em 1 de Setembro de 2009, o serviço do Gmail ficou 1:40min fora do ar por causa de problemas ocorridos em cascata em um serviço de manutenção de rotina;**



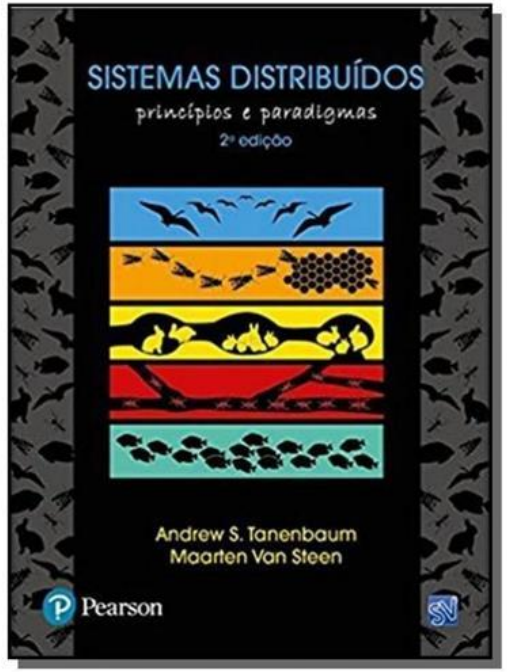
Visão Geral da Arquitetura Google

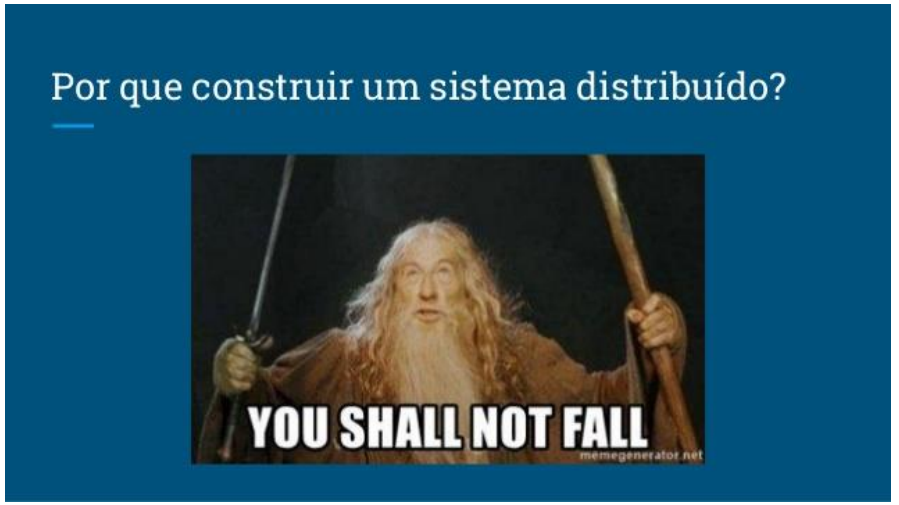
- Desempenho
- Qual o principal motivo da Google sempre fornecer seus serviços com alto desempenho?
  - Quanto melhor o desempenho, mais usuários usarão o sistema, mais os anúncios vendidos pela Google serão visualizados, maior será a renda.
  - Busca de alto desempenho: 0.2s.

Visão Geral da Arquitetura Google

- Abertura
- Google é uma empresa que encoraja e valoriza a inovação, por isso ela é sempre uma das pioneiras na criação de novos serviços e aplicações.
  - Possível graças a infraestrutura extensível (abertura do sistema).
  - Oferece total suporte para novos apps.
  - Segredo está na camada de Middleware: Permite que os serviços de busca/nuvem evoluam constantemente

Fim


SISTEMAS DISTRIBUÍDOS
princípios e paradigmas
2ª edição
Andrew S. Tanenbaum
Maarten Van Steen
Pearson


Por que construir um sistema distribuído?
YOU SHALL NOT FALL
memegenerator.net


THANK YOU