Processamento de Dados em Tempo Real (Streaming)

Centro Universitário Christus- UNICHRISTUS

Especialização em Ciência de Dados e Inteligência de Negócios (Big Data e BI)

Prof. Dr. Manoel Ribeiro

Unichristus
Centro Universitário Christus

Prof. Dr. Manoel Ribeiro

- Formação
  - Doutor em Computação Big Data, Machine Learning e Sistemas Distribuídos (UFC)
    - GPS2GR:Optimized Urban Green Routes based on GPS Trajectories
      - Temas: Trajectory Pattern Mining, Green Routes, Traffic-Light Scheduler
  - Mestre em Sistemas de apoio a decisão (UECE)
    - FastClass: Classificação Automática Fuzzy, enfase em Data mining; Análise de agrupamentos; Clustering; Análise de Componentes Principais; Fuzzy.
  - Publicações relevantes
    - GPS2GR:Optimized Urban Green Routes based on GPS Trajectories, 8th ACM SIGSPATIAL Workshop on GeoStreaming, 2017
    - LB-RLT Approach for Load Balancing Heterogeneous Storage Nodes. XXIX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos, 2016.
    - DMM: A Distributed Map-matching algorithm using the MapReduce Paradigm. Intelligent Transportation Systems Society Conference Management System, 2016.
  - Bacharel em Computação (UFC)
  - MBA em Finanças, Controladoria e Auditoria (FGV)
  - Especialista em Projetos (CETRED)

Prof. Dr. Manoel Ribeiro

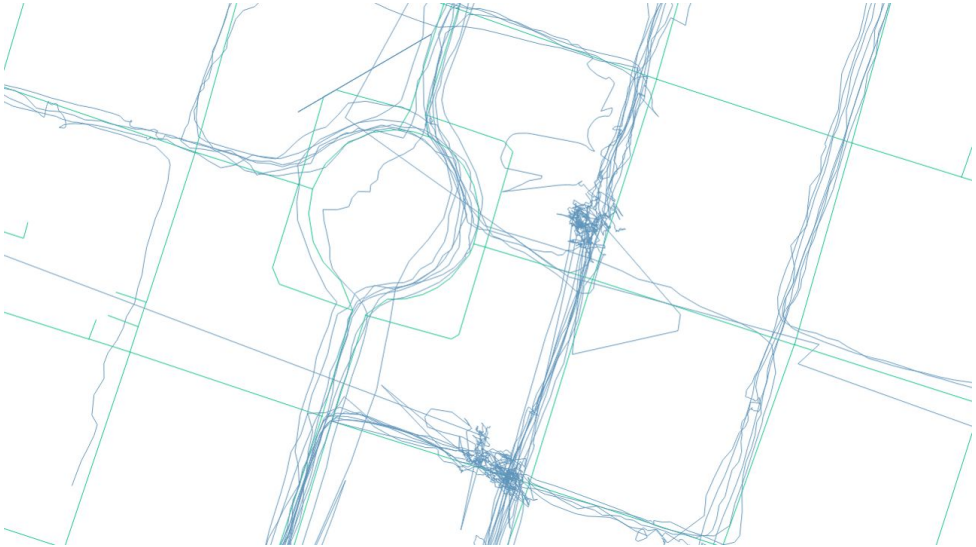
- Experiência
  - Foi executivo de TI por 25 anos no Grupo J.Macêdo e Grupo Marquise
    - Grupo J.Macêdo
      - Implantação do BI
      - Implantação do ERP SAP (SEM/BPS e BW)
      - Implantação da automação da força de venda
      - Desenvolvimento de sistema Inteligência de negócio - Navigator
      - Mudança de paradigma de formação de preço dos produtos
    - Grupo Marquise
      - Implantação ERP E-Business Suite (Oracle)
      - Implantação BI Cognos (IBM)
      - Terceirização de commodities de TIC
      - Terceirização de processos de negócios -ADP
  - Foi fundador e presidente do Grupo de Gestores de TIC do Ceará - GGTIC-CE
  - Foi sócio fundador da www.softium.com.br
  - Foi diretor de relacões institucionais do I3D.org.br

Prof. Dr. Manoel Ribeiro

- Atuação
  - Magistério Público Federal - Unilab
  - Professor de pós-graduação nas áreas de Data Science, BI e governança de TIC
  - Pesquisador associado no Instituto de Tecnologia da Informação e Comunicação (ITIC) com ênfase em IIoT, Big Data e Data Analytics
  - Possui quatro patentes em Sistemas Embarcados (INPI)
  - Consultoria em Data Science na **OPENCARE**
  - Empreendedor em IIoT com ênfase em:
    - **Data Logger** para sensores sem fio de longo alcance utilizando protocolo **LoRaWAN** (Mash) e com fio utilizando barramento **I2C** para uso industrial
    - Computação embarcada para acessibilidade

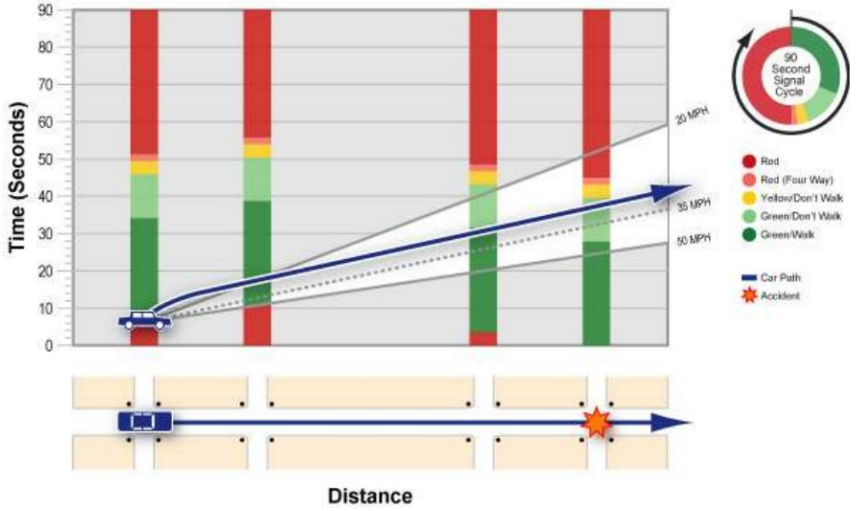
Prof. Dr. Manoel Ribeiro

- **DMM: A Distributed Map-matching algorithm using the MapReduce Paradigm**
  - Intelligent Transportation Systems Society Conference Management System, 2016.
  - Processamento em larga escala de trajetórias de GPS para descobertas de caminhos
  - Spark/Scala num cluster com 8 nós



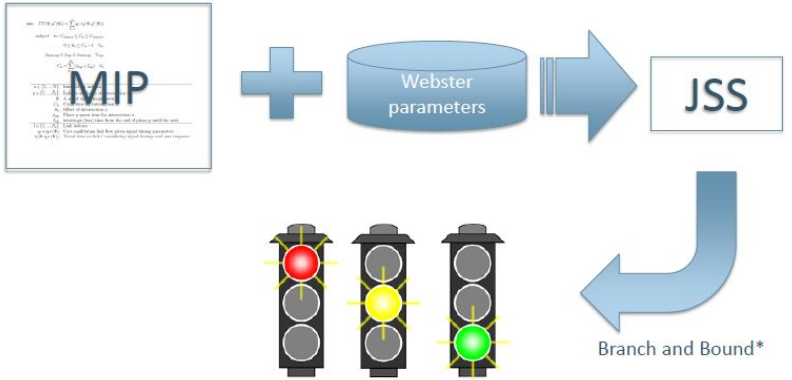
Prof. Dr. Manoel Ribeiro

- **GPS2GR:Optimized Urban Green Routes based on GPS Trajectories**
  - 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2017)
  - Processamento de BigData de trajetórias de veículos de uma grande cidade durante uma semana visando otimizar os semáforos para um padrão de deslocamentos diários
  - Pipeline/C#



Prof. Dr. Manoel Ribeiro

- **Optimization of urban semaphore times turning into JSSP**
  - 44th International Conference on Very Large Data Bases (VLDB 2018)
  - Processamento de BigData de semáforos e rotas frequentes
  - Google Optimization Tools



Prof. Dr. Manoel Ribeiro

- **Internet on the Forest - IoT**
  - ITIC/RNP/MCTI/EU
  - Sensores para captura de características específicas da região
  - Desafios intempéries, bateria, transmissão, armazenamento e análise
  - MongoDB/Sofia2

Prof. Dr. Manoel Ribeiro

- **Patentes de Invenção**
  - Dispositivo para monitoramento do consumo energético de equipamentos de computação
  - Sistema automatizado de acessibilidade e segurança pública para transporte urbano
  - Controle remoto universal para televisores com comando por voz
  - Etiqueta lavável para identificação de peças de roupas
  - ...

Metodologia

Aulas expositivas com discussões.

Práticas em laboratório.

Leituras.

Tarefas individuais.

Avaliação (em dupla).

Conteúdo da disciplina

- Dia 1 (sexta 18:00 às 22:00h)
  - Apresentação
  - Aula motivacional - Qual a importância do processamento de dados em tempo real?
- Dia 2 (sábado 8:00 às 18:00)
  - Fundamentos de sistema de processamento de tempo real
  - Fundamentos da arquitetura Apache Kafka, Apache ZooKeeper
  - Prática com Apache Kafka (tarde)

Conteúdo da disciplina

- Dia 3 (sexta 18:00 às 22:00h)
  - Fundamentos Apache Flume
  - Fundamentos Spark Streaming
  - Configuração Hadoop e Spark
- Dia 4 (sábado 8:00 às 18:00)
  - Fundamentos Introdução ao Apache Storm
  - Implementação de projeto prático/aplicado envolvendo Real Time Analytics
  - Avaliação

Repositório

https://github.com/antoniomralmeida/streaming

Entidades

DATA SCIENCE INSTITUTE

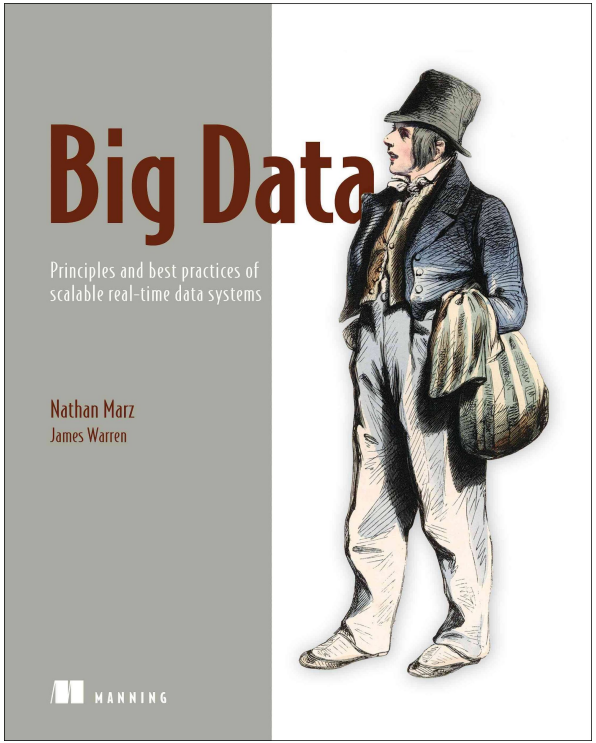
http://www.datascienceinstitute.org/

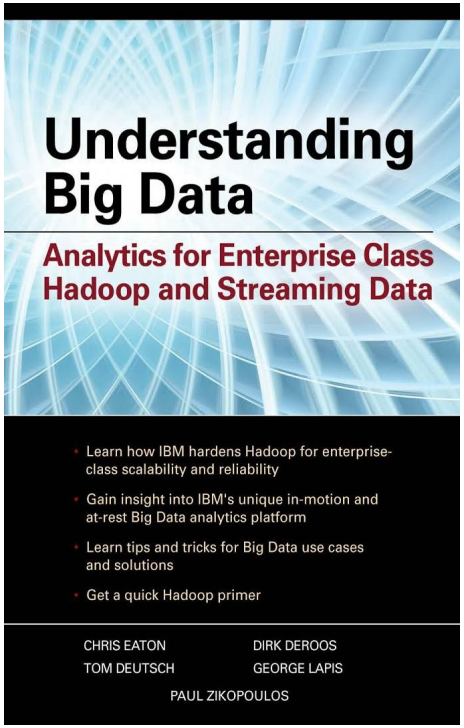
http://www.datascienceinstitute.com.br/

Pré-requisitos da disciplina

- Pré-requisitos da disciplina
  - Fundamentos de Rede
  - Fundamento de Sistemas Distribuídos
  - Bancos de Dados Relacional
  - Linguagem de Programação Java / Phyton

Bibliografia Básica






Contextualização

Contextualização

7,53 BILHÕES DE PESSOAS

2,2 BILHÕES DE USUÁRIOS (17%)
4 BILHÕES DE LIKES
300 MILHÕES DE FOTOS
83 MILHÕES FAKES

5 BILHÕES SMARTPHONES

Fonte: https://zephoria.com/top-15-valuable-facebook-statistics/, https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

Popularização dos Gadgets

2005

2013

Motor de buscas

- ★ O Google processa diariamente mais de 3 bilhões de pesquisas em todo o mundo
- ★ sendo desse total 15% totalmente inéditas.
- ★ Seu "motor" de pesquisa rastreia 20 bilhões de sites diariamente
- ★ armazenando 100 petabytes de informação.
- ★ Tudo em tempo real

Processamento de dados em tempo real (panorama atual de Big Data)

Processamento de dados em tempo real (panorama atual de Big Data

A tecnologia de Big Data evoluiu!

Se antes estava praticamente restrita ao armazenamento de dados, hoje assume protagonismo digital, sendo responsável por permitir análises de dados e respostas em tempo real, transformando o setor mundial de TI e ampliando a gama de serviços e produtos disponíveis aos usuários finais de internet.

Diretrizes de dados: prevendo as tendências de 2019

- Aplicativos preditivos: capacitando aplicativos por meio de aprendizado de máquina
- A tecnologia transforma: o papel do cientista de dados evolui
- Confiança no Analytics: o caminho para a adoção em nível corporativo

fonte: https://tdwi.org/articles/2019/01/03/adv-all-predicting-trends-of-2019.aspx

Os diversos significados de dados em tempo real

Extrapolando o entendimento de tempo real, muitas empresas precisam oferecer respostas imediatas, com acesso e análise de dados para reduzir a latência do negócio.

As principais aplicações no setor informático são:
– Dados sob demanda
– Transmissão de dados
– Fluxos contínuos de dados
– Dados de acesso imediato
– Dados em tempo certo
– Dados em tempo quase real

Os setores organizacionais que mais demandam dados em tempo real

- Operações (53%);
- Serviços ao consumidor e suporte (51%);
- Vendas (45%);
- Gestão organizacional e sistemas de TI (43%);
- Marketing (43%);
- Finanças (32%);
- Comércio eletrônico (28%);
- P&D (19%).

Melhorias organizacionais esperadas com adoção de análise de dados em tempo real

- Maior capacidade de resposta e ação (53%);
- Automação da resposta em tempo real para aprovações e serviços ao consumidor (47%);
- Melhoria da tomada de decisões (45%);
- Incremento operacional e de BI (45%).

Frequência de atualização de dados observada nas organizações

- Diariamente (83%);
- Mensalmente (60%);
- Tempo real (51%); e
- Semanalmente (49%).

Quais os tipos de dados em tempo real utilizados para incrementar o crescimento das organizações?

– Dados de redes sociais (38%)

– Dados de eventos (36%)

– Dados desestruturados (34%)

– Logs e cliques (34%)

Percentual de empresas que esperam aplicar o levantamento e análise de dados em tempo real

- ★ Estão pensando a respeito (33%);
- ★ Estão implementando (14%); e
- ★ Já utilizam em uma ou mais aplicações (4%).

Estudo de caso: Panorama de utilização de Big Data na PSafe

- Qual a importância do processamento de dados em tempo real para o negócio da PSafe?
  - Rodrigo Souza: Processamento de dados em Tempo Real é fundamental para a segurança dos clientes da PSafe, assim como para o bom desempenho do aplicativo.
- Quais funções do PSafe Total Android necessitam tratamento de dados em tempo real e que desafios representam para as equipes de Big Data e BI?
  - RS: Tratamos os dados de segurança em tempo real, já que não pode haver demora na detecção e tratamento de uma ameaça. Também processamos em tempo real o uso de todas as funcionalidades, customizadas para o gosto do usuário brasileiro.

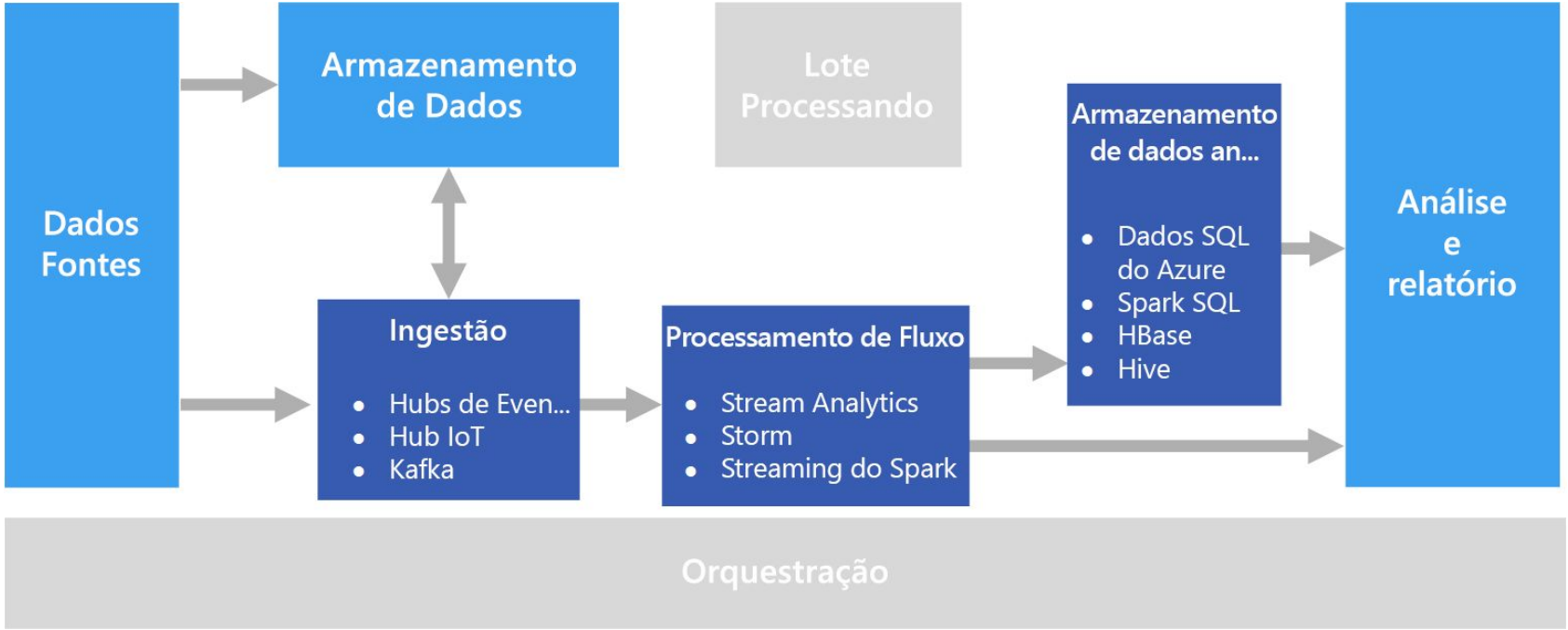
Estudo de caso: Processamento em tempo real na Microsoft

M: O processamento em tempo real lida com fluxos de dados que são capturados em tempo real e processados com latência mínima para gerar relatórios em tempo real (ou quase em tempo real) ou respostas automatizadas.

M: Por exemplo, uma solução de monitoramento de tráfego em tempo real pode usar dados de sensor para detectar grandes volumes de tráfego. Esses dados podem ser usados para atualizar um mapa dinamicamente para mostrar o congestionamento ou iniciar automaticamente faixas de ocupação alta ou outros sistemas de gerenciamento de tráfego.

https://docs.microsoft.com/pt-br/azure/architecture/data-guide/big-data/real-time-processing

Estudo de caso: Processamento em tempo real na Microsoft



https://docs.microsoft.com/pt-br/azure/architecture/data-guide/big-data/real-time-processing

Estudo de caso: Processamento em tempo real na Microsoft

M: Um dos grandes desafios das soluções de processamento em tempo real é ingerir, processar e armazenar mensagens em tempo real, especialmente em grandes volumes.

M: O processamento precisa ser feito de forma que não bloqueie o pipeline de ingestão. O armazenamento de dados precisa dar suporte a gravações de alto volume.

M: Outro desafio é conseguir tomar decisões com base nos dados rapidamente, como a geração de alertas em tempo real ou apresentação dos dados em um painel em tempo real (ou quase em tempo real).

https://docs.microsoft.com/pt-br/azure/architecture/data-guide/big-data/real-time-processing

Estudo de caso: Bovespa em tempo real via Streaming

Desde 2007, quando a BM&FBovespa lançou o Acesso Direto ao Mercado (Direct Market Access — DMA, em inglês), os clientes finais passaram a ter a opção de enviar ofertas diretamente ao ambiente de negociação eletrônica por meio de uma corretora.

Para atender a esse modelo, as corretoras, então, investiram em infraestrutura, software de negociação eletrônica e negociação por algoritmo. Muitos clientes finais investiram em software e até mesmo no desenvolvimento de sistemas proprietários para a tomada de decisão ou acompanhamento de mercado.

https://blog.cedrotech.com/saiba-como-acompanhar-dados-bovespa-em-tempo-real-via-streaming/

Estudo de caso: Internet das Coisas (IoT) e processamento em tempo real (ESP) na SAS

SAS: Historicamente, a análise de dados é realizada após os dados terem parado de se movimentar e estarem armazenados. No entanto, recentemente, percebemos um aumento na demanda do 'processamento em tempo real' (Event Stream Processing – ESP), na qual as informações são continuamente analisadas ainda em movimento.

SAS: O ESP captura o valor dos dados no instante em que eles são gerados, antes que eles se percam no espaço de tempo entre a criação e armazenamento. Essa tecnologia também detecta padrões significativos, além de desvios dos mesmos que irão indicar uma ação imediata.

https://www.sas.com/pt_br/insights/articles/data-management/internet-das-coisas-processamento-em-tempo-real-esp.html

Impactos das inovações em serviços no mercado brasileiro de música: o caso da tecnologia streaming.

Estudo analisa os impactos dos serviços de streaming interativos de música, ou webcast on demand, nos modelos de negócio adotados no mercado brasileiro da música gravada e nos seguintes agentes produtores dessa cadeia produtiva: a indústria fonográfica (titulares de direitos), os artistas (interpretes), músicos e compositores.

A inserção de novos agentes nesse mercado, como Spotify, Deezer e Apple Music, acarretou o surgimento de modelos de negócio com base em novos formatos de consumo que vêm se difundindo no Brasil. A música gravada passou a ser consumida sob a forma de serviço gerando recuperação financeira aos agentes da indústria fonográfica e novos desafios aos músicos e compositores.

http://www.ie.ufrj.br/images/defesas/pped2017/leonardo_de_moraes_morel__86f6b.pdf

Big Data - Valor

Data is the new Oil!

- "Dados são o novo Petróleo"

Perry Rotella
Contributor

FOLLOW

full bio →

Opinions expressed by Forbes Contributors are their own.

TECH 4/02/2012 @ 11:09AM | 10,791 views

Is Data The New Oil?

+ Comment Now + Follow Comments

Recently, on a CNBC Squawk Box segment, "The Pulse of Silicon Valley," host Joe Kernan posed the question, "What is the next really big thing?" to Ann Winblad, the legendary investor and senior partner at Hummer-Winblad. Her response: "Data is the new oil."

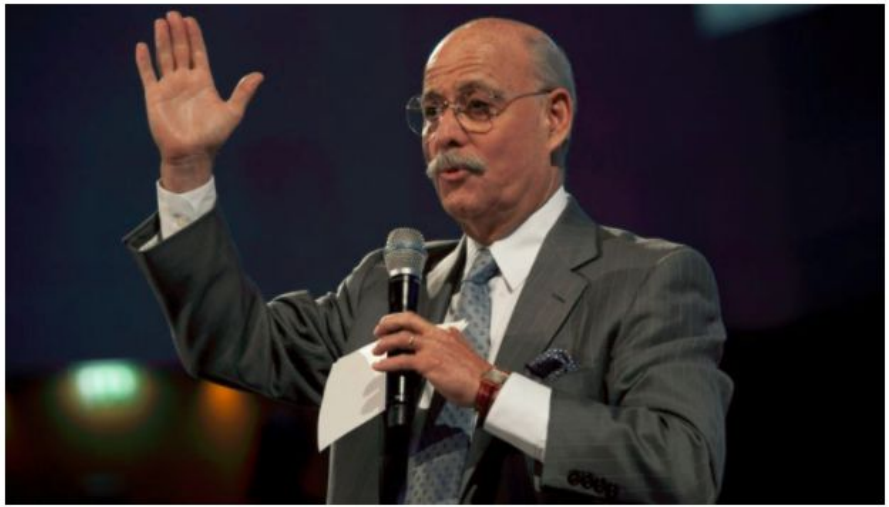
- Como petróleo, precisam ser refinados !

A nova economia do compartilhamento dos dados

Jeremy Rifkin: The sharing economy is the future of the society

The economist explained to the Forum PA audience why the future of human race is in danger and how the only choice available to public and private organizations is a change towards the sharing economy

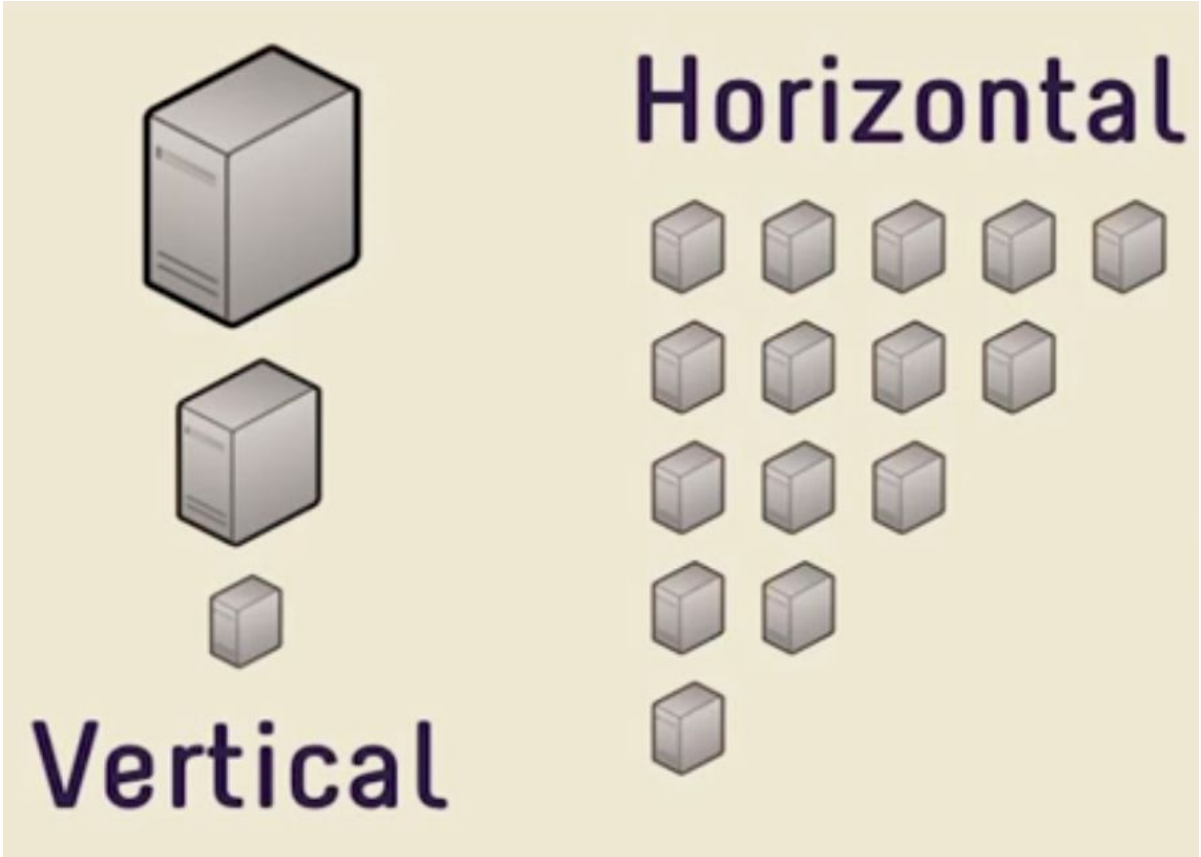
Barbara Bosco



Como se processa alto tão desafiador?

Escalabilidade Horizontal

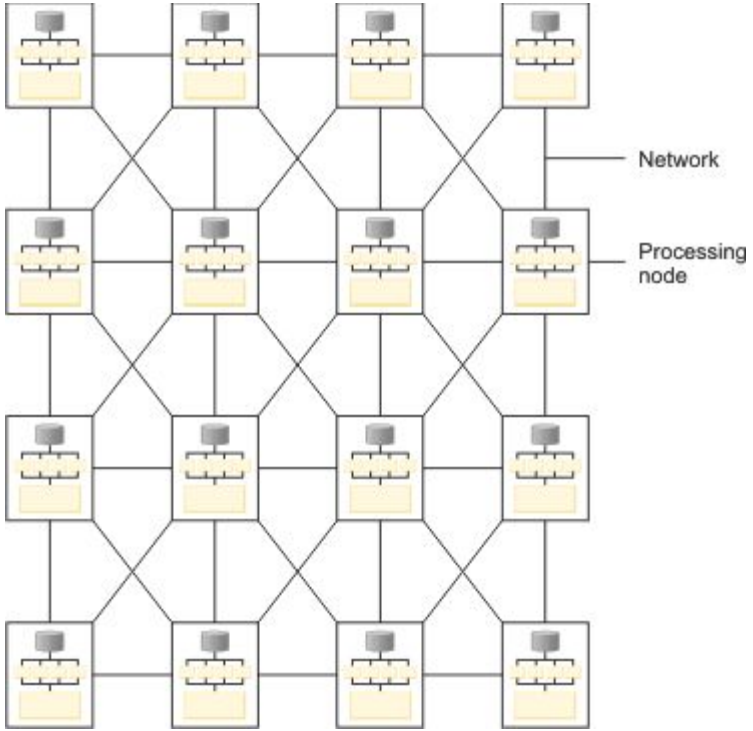
Novo paradigma da computação moderna



MASSIVE PARALLEL PROCESS - MPP

Remodelagem do conceito de Sistemas Distribuídos com gestão e recuperação de falhas: Cluster e Grid

HDFS - Modelo eventualmente consistente



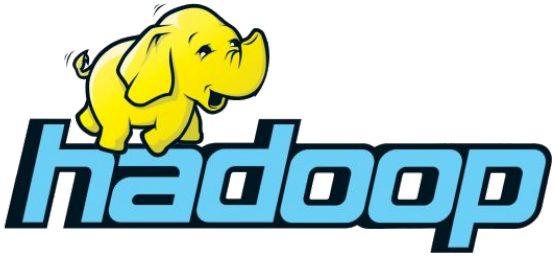
Novo paradigma para resolução de problemas complexos

“Dividir e Conquistar” é uma técnica de projeto de algoritmos que consiste em resolver um problema a partir da solução de “sub-problemas menores” do mesmo tipo.

Dividir para conquistar (“Divide et impera” ou “Divide et Vinces”) é um clássico nas estratégias de guerra para enfraquecer e subjugar os povos. O termo, embora já era conhecida na Antiguidade, foi cunhado por Júlio César em seu livro” De Bello Gallico” (Guerra das Gálias), que explicou como a vitória romana na guerra gaulesa era essencialmente uma política de “dividir” seus inimigos, aliar com tribos individuais durante suas disputas com adversários locais.

NoSQL

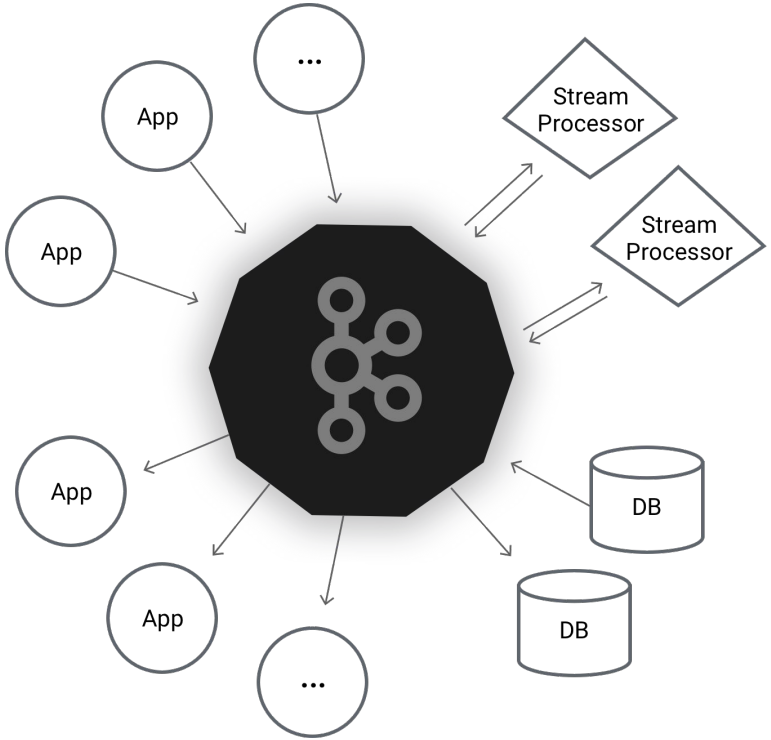
Novo paradigma para de banco de dados de estrutura mais simples e altamente escalar


hadoop


APACHE
Spark
TM

APACHE kafka®
A distributed streaming platform

O Kafka® é usado para construir pipelines de dados e aplicativos de streaming em tempo real.
É escalável horizontalmente, tolerante a falhas, perversamente rápida e é executado em produção em milhares de empresas.



O Flume é um serviço distribuído, confiável e disponível para coletar, agregar e mover com eficiência grandes quantidades de dados de log. Tem uma arquitetura simples e flexível baseada em fluxos de dados de fluxo contínuo. Ele é robusto e tolerante a falhas, com mecanismos de confiabilidade ajustáveis e muitos mecanismos de failover e recuperação. Ele usa um modelo de dados extensível simples que permite a aplicação analítica on-line.

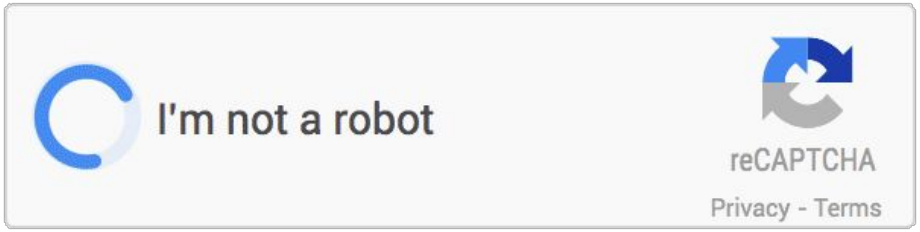
O Apache Storm é um sistema de computação distribuído em tempo real gratuito e de código aberto. O Storm facilita o processamento confiável de fluxos de dados ilimitados, fazendo o processamento em tempo real do que o Hadoop fez para o processamento em lote. Storm é simples, pode ser usado com qualquer linguagem de programação e é muito divertido de usar!

O Spark Streaming traz a API integrada ao ambiente do Apache Spark para o processamento de fluxo, permitindo que você grave tarefas de fluxo contínuo da mesma forma que grava tarefas em lote. Suporta Java, Scala e Python.

Qual o benefício disso tudo?

Bots Cybercrime e prova de humanidade

- reCaptcha - ferramenta de prova de humanidade que utiliza de reconhecimento de imagem ou ontologia para reconhecimento humano
- Pergunta:
  - Quanto custaria hoje para combater manualmente as tentativas de acesso indevido via Bots?



  - Estimativas falam em bilhões de dólares de economia para muitas empresas

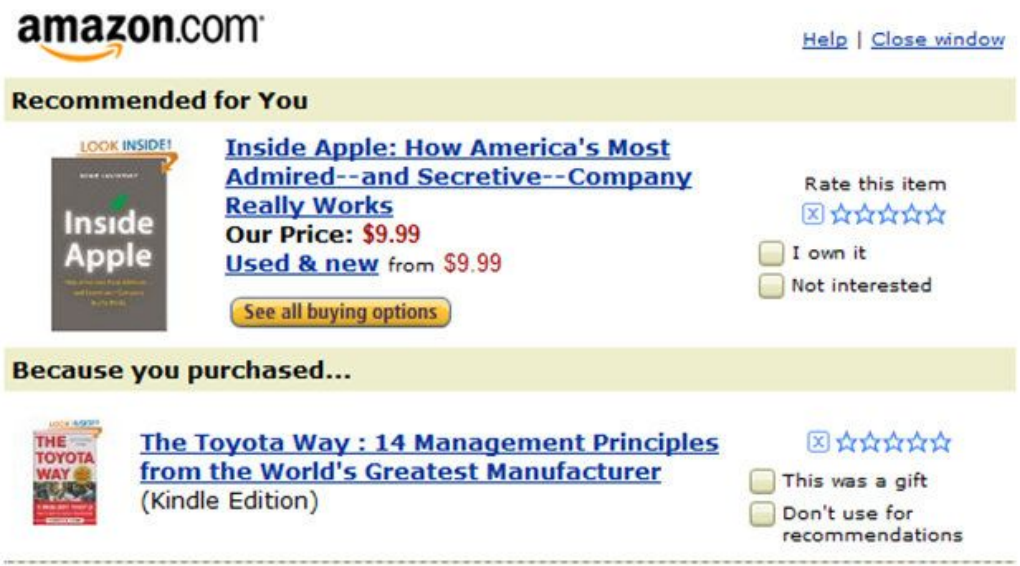
Captcha e ReCaptcha foram criados por Von Ahn

Sistema de recomendação

- O objetivo dos sistemas de recomendação (SR) é gerar recomendações válidas para um conjunto de usuários, de itens que possam interessá-los

Google AdSense

NETFLIX



Fim