

Processamento de dados em Big Data

Parte 1 - Motivação e conceitos

Luiz Henrique Zambom Santana, D.Sc.

INE | CTC



UNIVERSIDADE FEDERAL
DE SANTA CATARINA

Vamos nos apresentar?

- **Nome Dados gerais (idade, onde mora, filhos) - se quiser compartilhar ;)**
- **Qual é o projeto atual?**
- **Uma curiosidade/hobby**
- **O que já leu/viu sobre processamento de Big Data e o que espera**

Agenda

- Motivação
- Histórico
 - Demanda
 - IoT
 - Machine Learning
 - Oferta
 - Computação distribuída
 - Nuvem
 - Micro Serviços
 - Serverless
- Computação em larga escala
 - Map Reduce
 - Streaming
 - In-memory
 - Datalakes
 - NoSQL
 - Orquestração

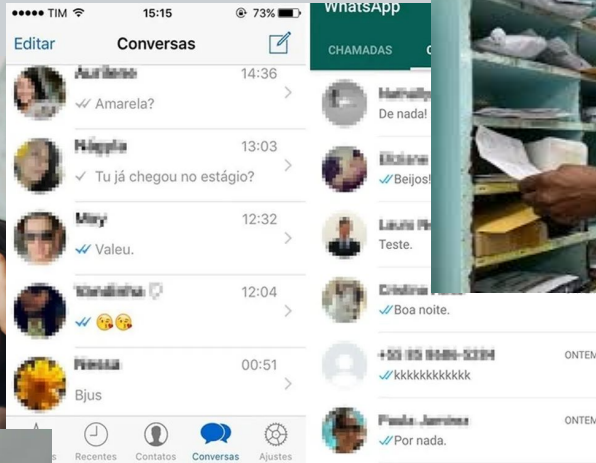
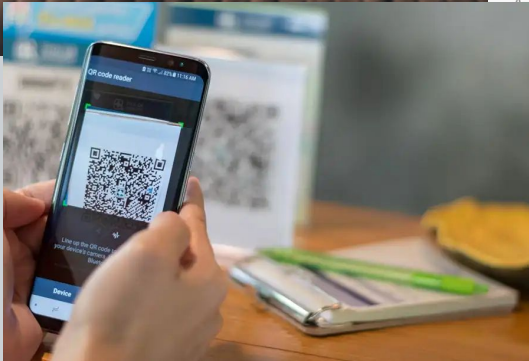
Motivação

De onde vem a Big Data?

Anos 1990...



Anos 2000...



Por volta de 2010...

How big is your data – really?
H/T to David Wellman @ Myriad Genetics

Byte of data:	one grain of rice
Kilobyte:	cup of rice
Megabyte:	8 bags of rice
Gigabyte:	3 container lorries
Terabyte:	2 container ships
Petabyte:	covers Manhattan
Exabyte:	covers the UK (3 times)
Zettabyte:	fills the Pacific ocean

© 2010 Pure Storage Inc.

The Economist
FEBRUARY 27TH - MARCH 5TH 2010
Economist.com

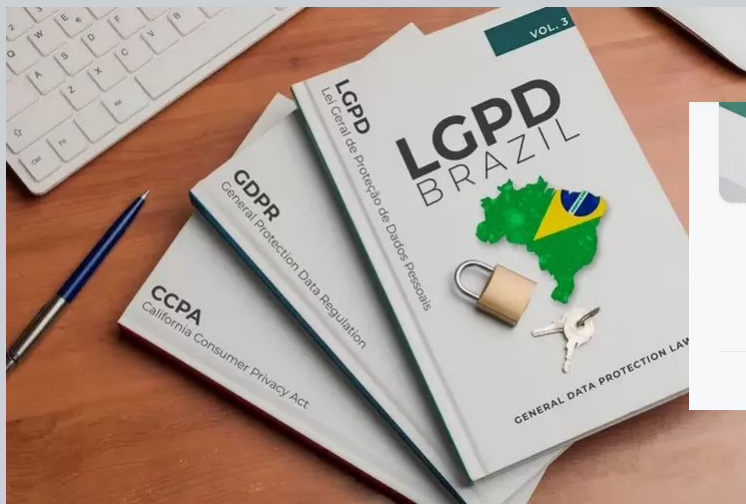
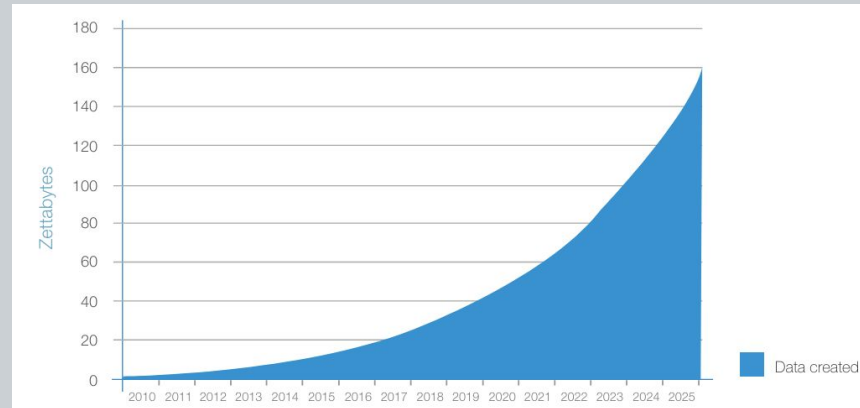
Gordon Brown's pitch
What went wrong at RBS
Genetically modified crops blossom
The EU woos Russia
The right to eat cats and dogs

The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

64.00
9 770013 061110

Atualmente



Pix bate recorde com 152,7 milhões de transações em um único dia

Valor médio foi de quase R\$ 500

Publicado em 08/09/2023 - 13:48 Por Bruno de Freitas Moura* - Repórter da Agência Brasil - Rio de Janeiro
Atualizado em 08/09/2023 - 14:54

Quanto de dados é Big Data?



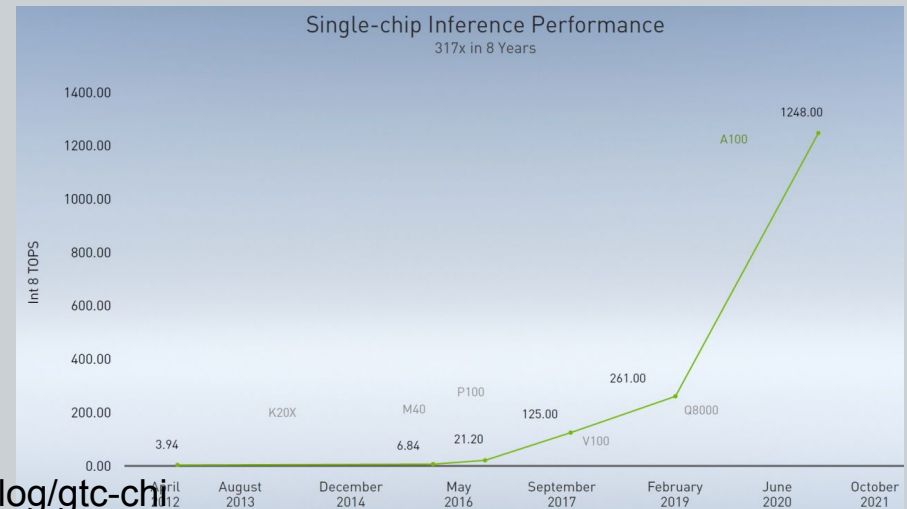
Atualmente

Demanda

- IoT
- Machine Learning
- Problemas de larga escala:
 - ESG
 - Mudanças climáticas
 - Segurança
 - ...

Oferta

- Computação distribuída
- Nuvem
- Micro Serviços
- Serverless
- ...



<https://blogs.nvidia.com/blog/gtc-china-keynote-dally-ai/>

No início era o Map Reduce...

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper.

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

As a reaction to this complexity, we designed a new

<https://static.googleusercontent.com/media/research.google.com/en/archive/mapreduce-osdi04.pdf>

Exemplo de como você pode utilizar imagem e texto no mesmo slide.

Evite diminuir o tamanho da fonte. Crie novos slides, se necessário.

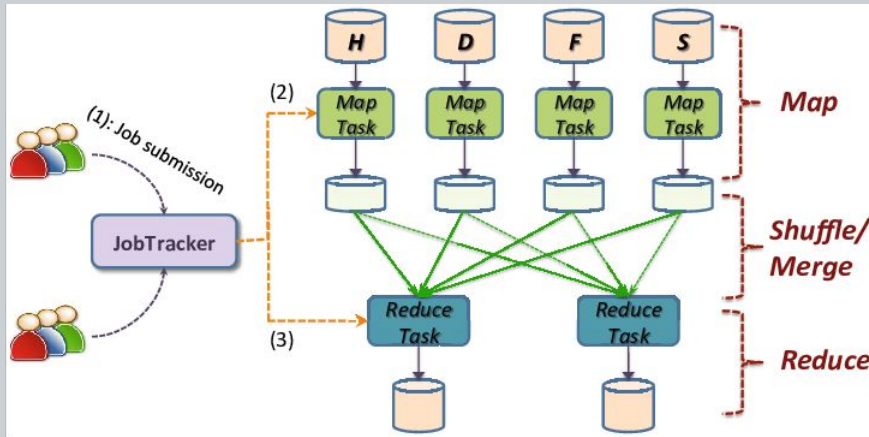
As imagens utilizadas neste arquivo são de uso exclusivo da UFSC.

Use esse espaço para escrever.

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam.

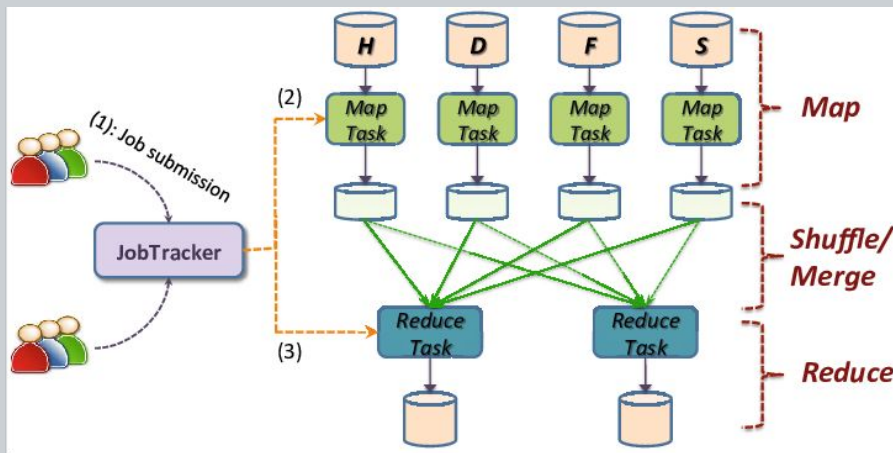
Como funciona o Map Reduce...

Apache Hadoop - conceitos



- Em o Hadoop 2010 foi promovido a top level project da Fundação Apache.
- Foi um dos primeiros frameworks open-source que implementavam o MapReduce.
- Além disso, ele trouxe um ecossistema completo de execução com o Hive, Pig e o Zookeeper.

In-memory data processing



- As primeiras versões do Hadoop tinham um problema:
 - A sincronização das etapas era feita em disco (HDFS);
 - A prioridade do Google era a **vazão** do seu processamento em background.

Apache Spark - conceitos

An Architecture for Fast and General Data Processing on Large Clusters

by

Matei Alexandru Zaharia

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy
in

Computer Science

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Em 2013 foi criado o Spark, ele resolveu o problema da escrita em disco usando uma abstração de memória chamada RDD. Com isso novos casos de uso se tornaram possíveis:

- Streaming
- Processamento de grafos
- Machine Learning

<https://escholarship.org/content/qt19k949h3/qt19k949h3.pdf>

Streaming

Throughput (vazão) vs. Latência



Streaming

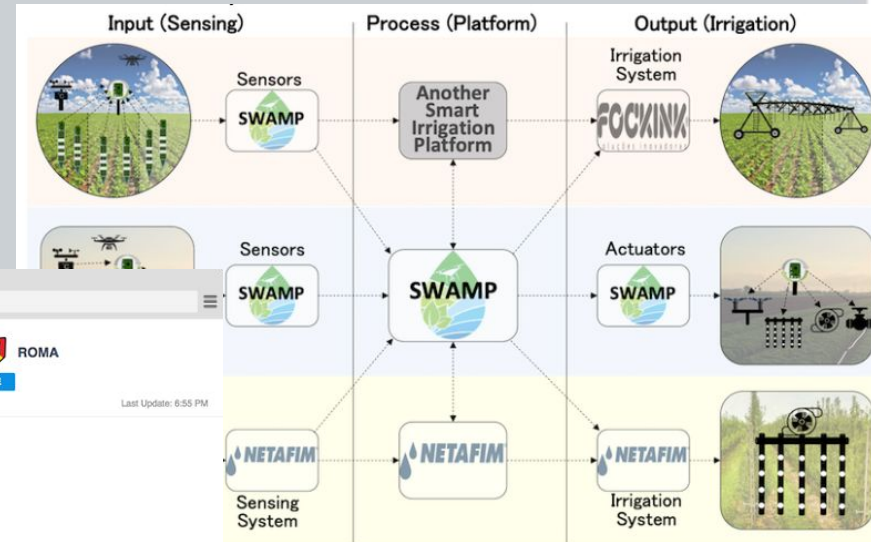
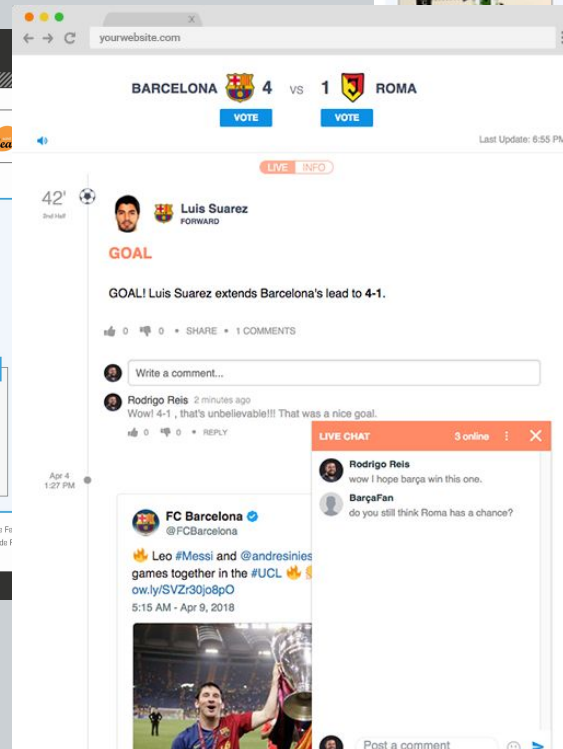
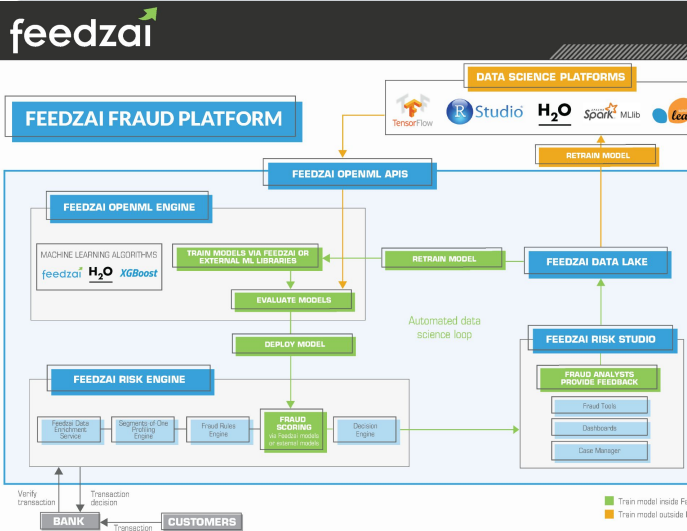
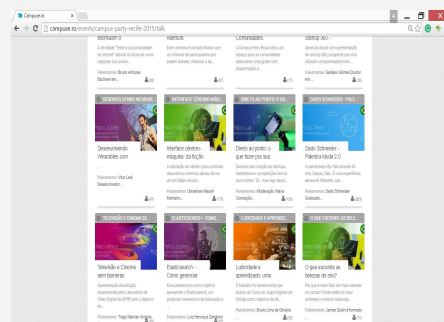
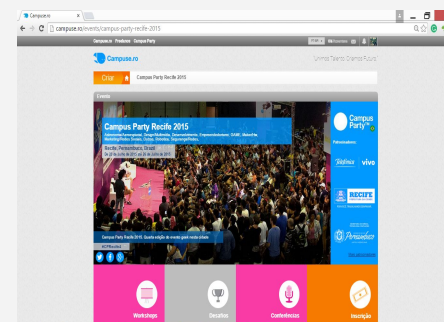
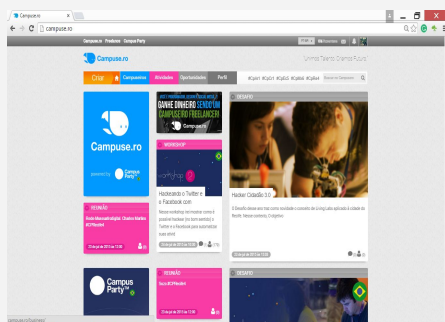
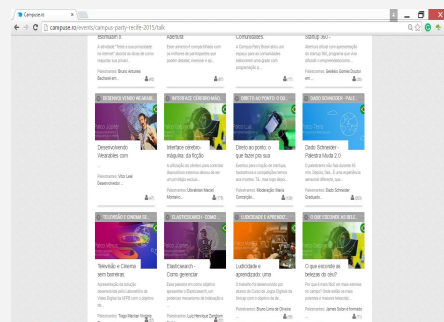
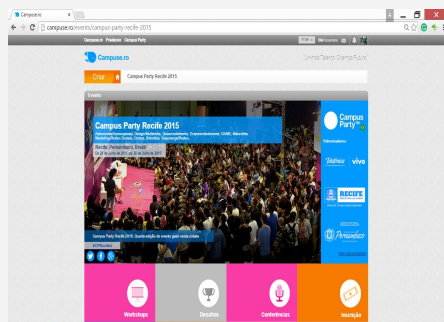
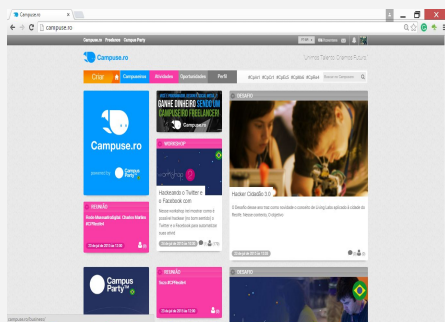
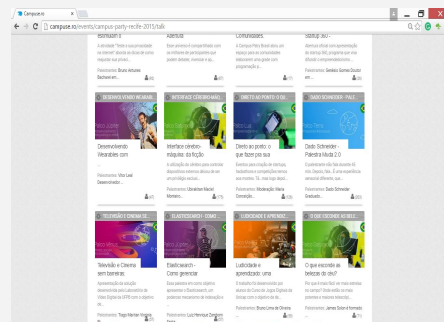
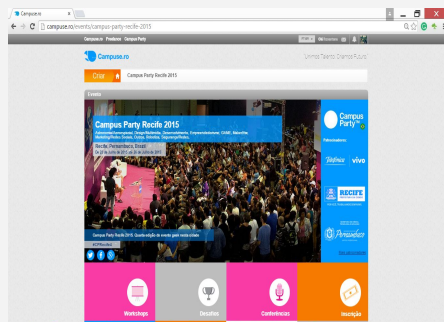
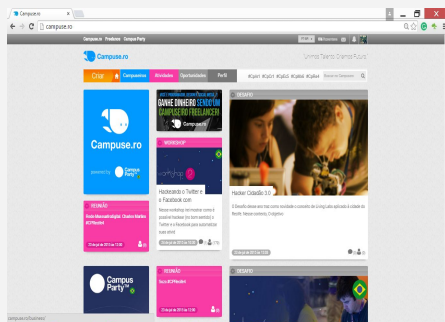


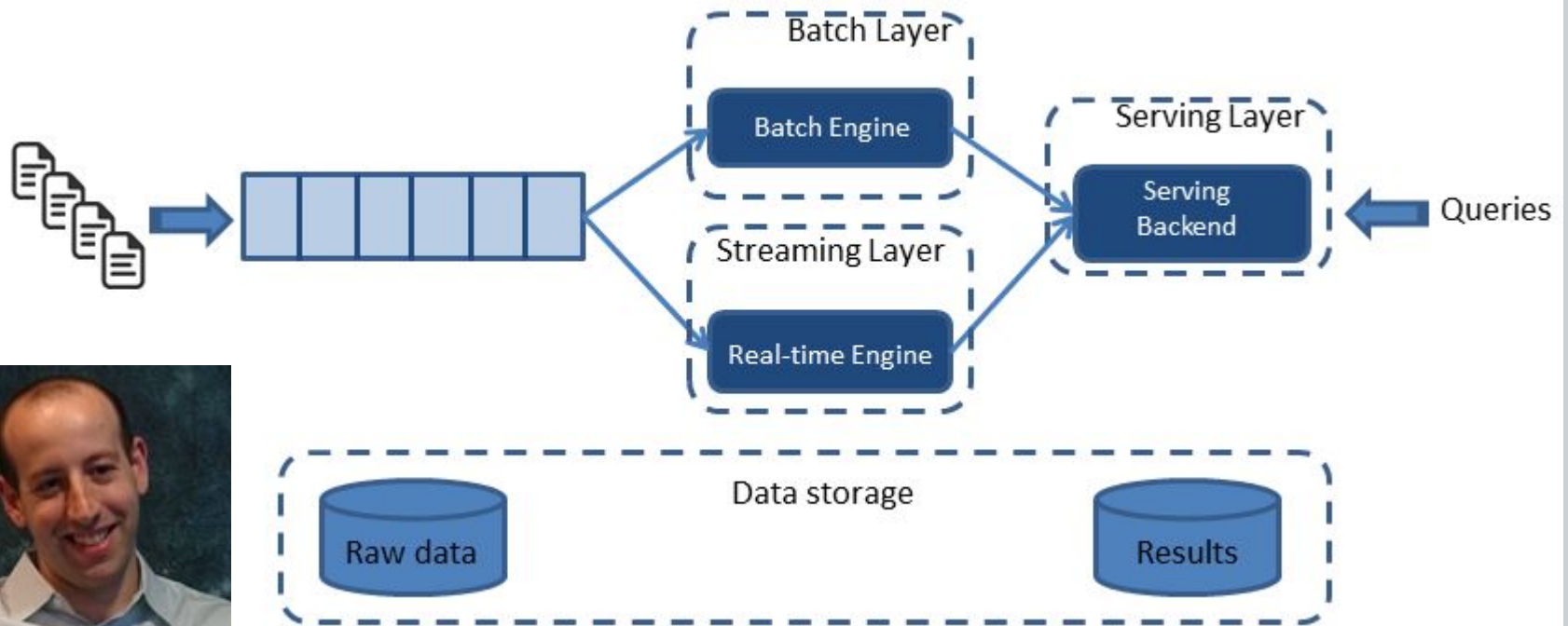
Figure 1: An Open IoT Ecosystem for Smart Irrigation



Streaming

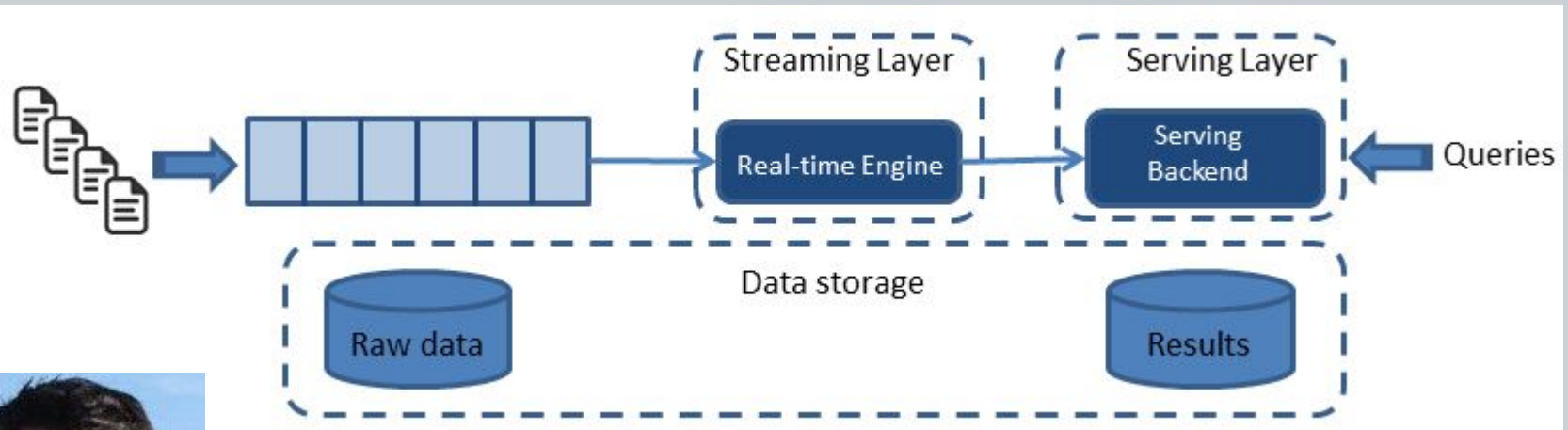
Streaming é tempo real?

Streaming



<https://www.linkedin.com/in/nathanmarz/>

Streaming

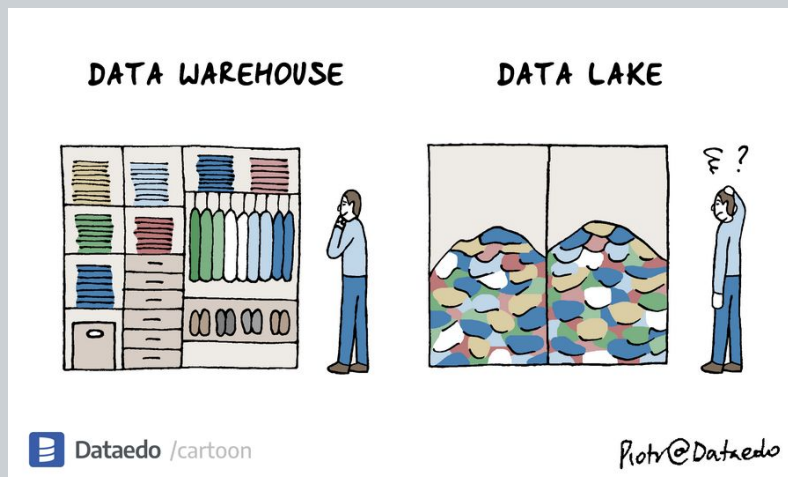


<https://www.linkedin.com/in/jaykreps/>

Data Lakes







A demanda e oferta de dados fez surgir a ideia de Data Lakes:

- Um repositório que armazena os dados de forma não estruturada;
- Pode ser usado para a descoberta de informações;
- Se mal usado, pode se tornar um Data Swamp.



Formatos de dados

BIG DATA FORMATS COMPARISON

	Avro	Parquet	ORC
Schema Evolution Support			
Compression			
Splitability			
Most Compatible Platforms	Kafka, Druid	Impala, Arrow Drill, Spark	Hive, Presto
Row or Column	Row	Column	Column
Read or Write	Write	Read	Read

Source: Nexla analysis, April 2018

Ao longo dos anos, novos formatos de dados foram sendo criados para com foco em:

- Custo de armazenamento
- Custo de transmissão
- Estruturação e facilidade de análise

NoSQL



Not only SQL

Sadalage e Fowler, 2012

<http://martinfowler.com/books/nosql.html>

“Banco de Dados Relacional
será nota de rodapé na história”

Nathan Marz, 2014

<http://goo.gl/WGXvPy>



NoSQL

SQL and NoSQL will merge “Not yet SQL”

Michael Stonebraker, 2015

<https://www.youtube.com/watch?v=KRcecxvGxvQ>



Como podemos usar essas ideias em domínios reais?

- Domínios
 - ESG
 - eHealth
 - Apostas em esporte
 - Reconhecimento facial para segurança pública
 - Gestão de força de trabalho em empresas terceirizadas
- Perguntas
 - Esse domínio precisa de Big Data?
 - Existe alguma oportunidade para uso de Streaming?
 - O domínio se beneficiaria de Machine Learning? Qual tipo?
 - Como os dados poderiam ser armazenados?

Luiz Henrique Zambom Santana

lhzsantana@gmail.com

<https://www.linkedin.com/in/luizsantana/>



UNIVERSIDADE FEDERAL
DE SANTA CATARINA