

BIG DATA

Olá!

Felipe

Dr. Finanças - EAESP - FGV
Me. Estatística - UFSCar
Me. Eng Prod. - USP São Carlos
Adm - USP Ribeirão Preto

tumenas@ufba.br



Trainee - 2006



McKinsey Fellow Analyst -
2008



Consultor de Riscos - 2011



Consultor Sênior - 2015



BNP PARIBAS
CARDIF

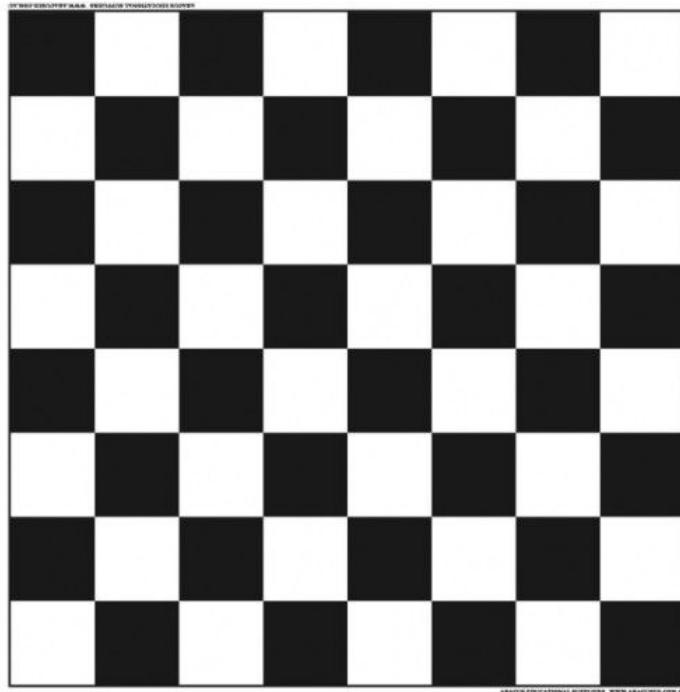
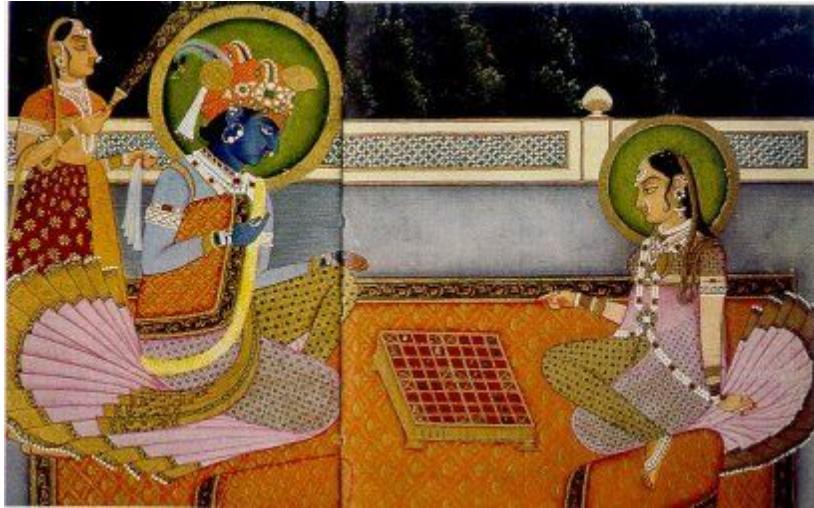
Gerente de Big Data e
Monitoramento de Riscos - 2016
(2018)

AGENDA

Dia 1 - Visão Geral e Hadoop

Dia 2 - Spark, Parquet, MongoDB



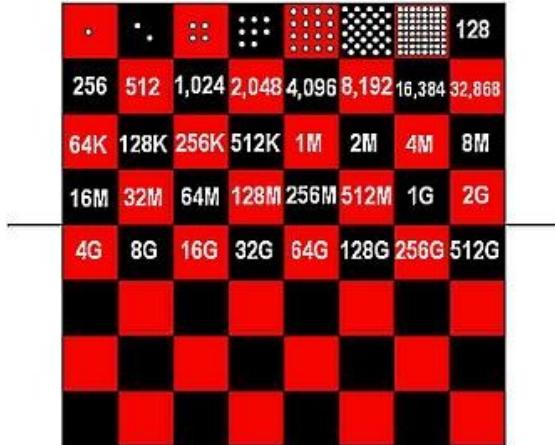


QUAL O VOLUME DE ARROZ?



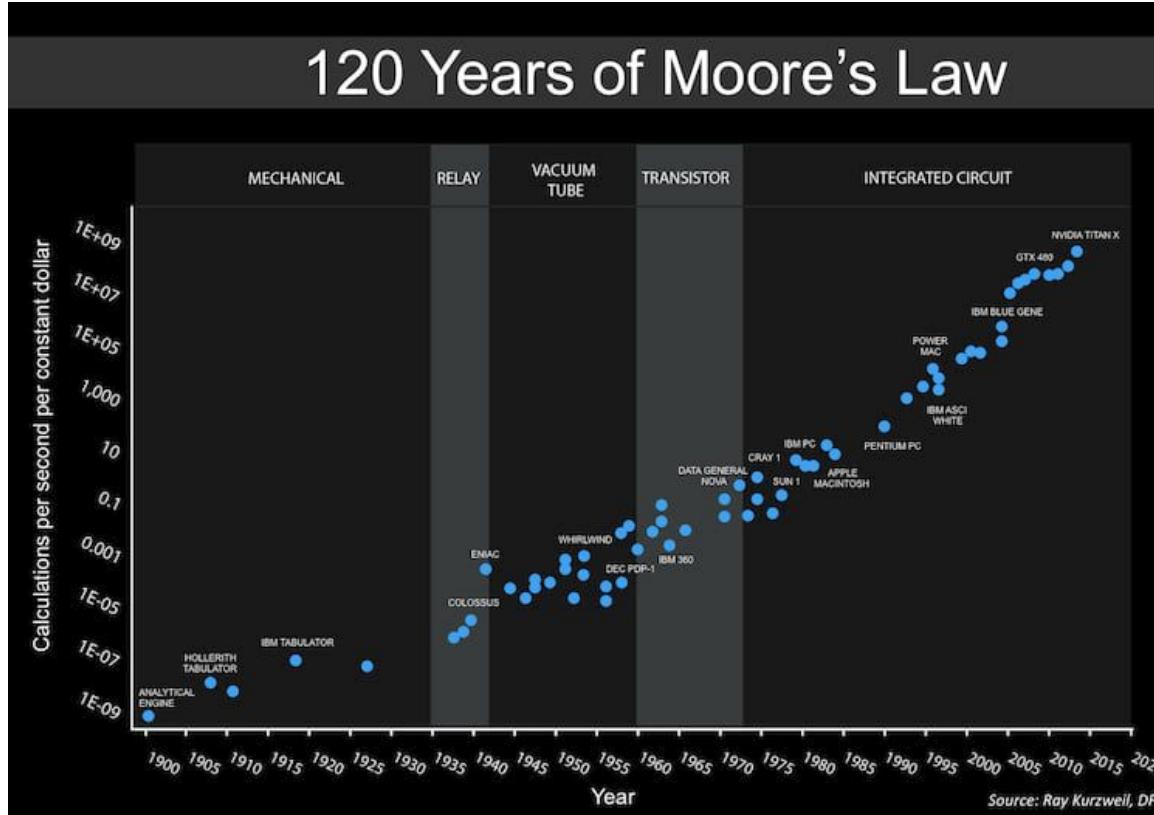


QUAL O VOLUME DE ARROZ?



210 bilhões de toneladas
de arroz

LEI DE MOORE



Gordon Moore (Intel): A cada 2 anos o número de transistores em um processador é dobrado

LEI DE MOORE

nature > nature electronics > perspectives > article

**nature
electronics**

Perspective | Published: 08 January 2018

Science and research policy at the end of Moore's law

Hassan N. Khan, David A. Hounshell & Erica R. H. Fuchs ✎

Nature Electronics 1, 14–21 (2018) | Download Citation ↴

ⓘ A Publisher Correction to this article was published on 05 February 2018

ⓘ This article has been updated

Abstract

The integrated circuit is today synonymous with the concept of technological progress. In the seven decades since the invention of the transistor at Bell Labs, relentless progress in the development of semiconductor devices — Moore's law — has been achieved despite regular warnings from industry observers about impending limits. Here, drawing on technical and organizational archival work and oral histories, we argue that the current technological and structural challenges facing the industry are unprecedented and undermine the incentives for continued collective action in research and development.

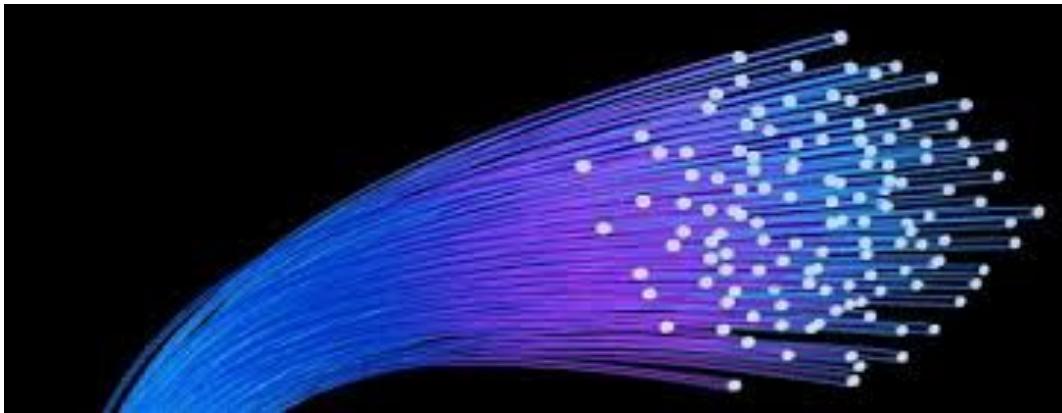
<https://www.nature.com/articles/s41928-017-0005-9>

LEI DE KRYDER



Mark Kryder (Intel): A 10.5 anos a densidade de informações armazenada em um disco rígido é aumentada por um fator de 1.000

LEI DE BUTTER

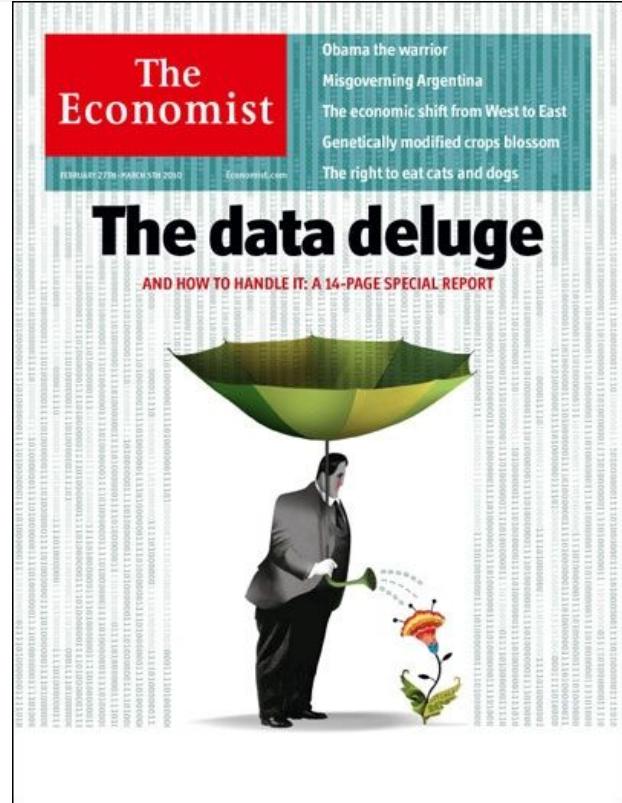


Gerry Butter (Bell Labs): O volume de informações que pode passar por uma fibra ótica dobra a cada 9 meses

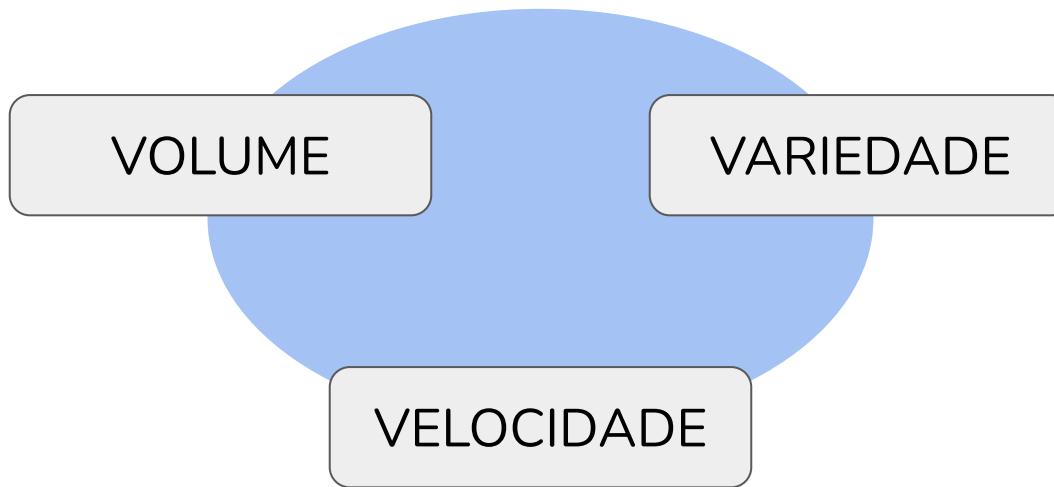
BIG DATA

“The term “*big data*” refers to *data* that is so large, fast or complex that it’s difficult or impossible to process using traditional methods” SAS

https://www.sas.com/pt_br/insights/big-data/what-is-big-data.html



3Vs



V's



CHANGEMAKER

Os 10 Vs do big data

O termo big data começou a aparecer com moderação no início dos anos 90, e sua prevalência e importância aumentaram exponencialmente com o passar dos últimos anos

4 min de leitura

Ouça

Home > Infra > Big Data

Big Data: os cinco Vs que todo mundo deveria saber

Por Redação

© shutterstock

E em um mundo cada vez mais conectado, o Big Data é um dos temas mais relevantes do mercado de TI. Para te ajudar a entender melhor sobre este assunto, o LinkedIn fez uma lista que ensina exatamente o que é essa tecnologia.

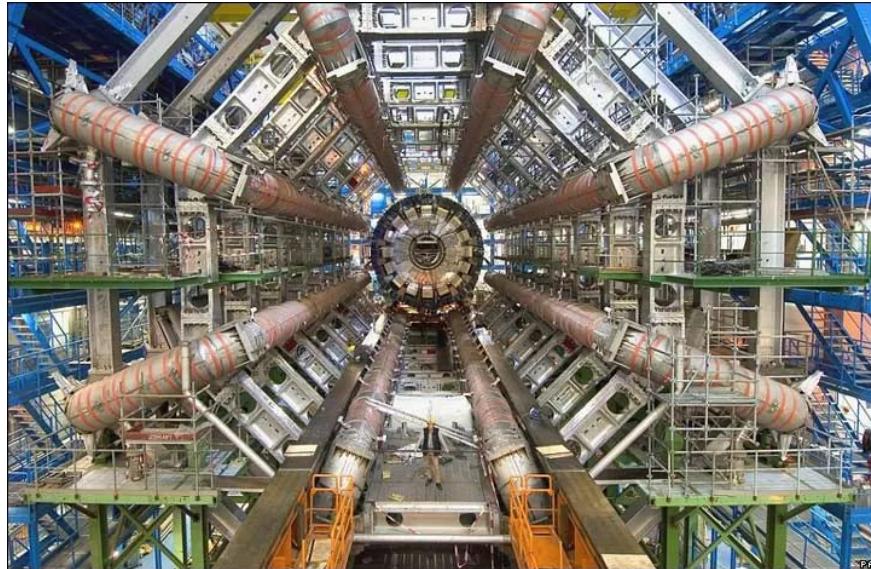
1. Volume

Big Data é uma grande quantidade de dados gerada a cada segundo. Pense em todos os e-mails, mensagens de Twitter, fotos e vídeos que circulam na rede a cada instante. Não são terabytes e sim zetabytes e brontobytes. Só no Facebook são 10 bilhões de mensagens, 4,5 bilhões de curtidas e 350 milhões de fotos compartilhadas todos os dias. A tecnologia do Big Data serve exatamente para lidar com esse volume de dados,

CERN (LHC)

“The raw data per event is around one million bytes (1 MB), produced at a rate of about 600 million events per second.”

$600.000.000 \times 1\text{MB} \sim \text{6 PB/s}$



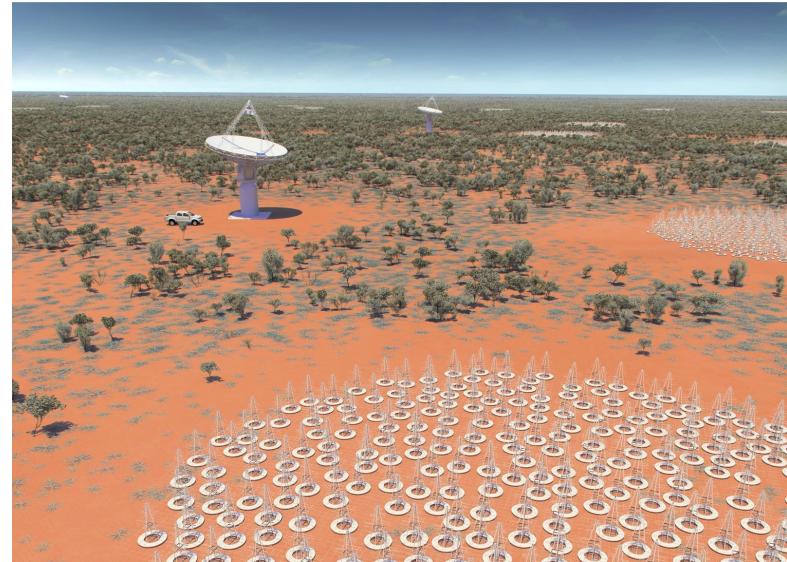
https://root.cern/

The screenshot shows the official website for the ROOT Data Analysis Framework. At the top, there is a navigation bar with links for About, Install, Get Started, Forum & Help, Manual, Blog Posts, Contribute, For Developers, and a search icon. Below the navigation bar is a large banner featuring a complex 3D visualization of particle collision tracks and energy deposits. On the left side of the banner, the text "ROOT: analyzing petabytes of data, scientifically." is displayed, followed by the subtitle "An open-source data analysis framework used by high energy physics and others." Below the banner are two buttons: "Learn more" and "Install v6.22/06". Underneath the banner, there are four main navigation links with icons: "Get Started" (document icon), "Reference" (book icon), "Forum & Help" (speech bubble icon), and "Gallery" (bar chart icon). At the bottom of the page, there are social media sharing icons for LinkedIn, GitHub, and Twitter.

https://root.cern/

SKA - (AUSTRÁLIA e ÁFRICA DO SUL)

“The SKA-Low array will generate 5 zettabytes of data every year—an unimaginable volume when considering global Internet traffic only passed 1 zettabyte for the first time in 2016. “



<https://spie.org/news/photonics-focus/mayjun-2020/square-kilometer-array-big-data?SSO=1>

[SUBSCRIBE](#) [RENEW](#) [GIVE A GIFT](#)

Smithsonian
MAGAZINE

SMARTNEWS HISTORY SCIENCE INGENUITY ARTS & CULTURE TRAVEL AT THE SMITHSONIAN PHOTOS VIDEO

AGE OF HUMANS FUTURE OF SPACE EXPLORATION HUMAN BEHAVIOR MIND & BODY OUR PLANET SPACE WILDLIFE NEWSLETTER EARTH

ADVERTISING

PHOTO C



Red Stag
[PHOTO](#)

MOST PO

1. Study
Bridge

“Earlier this year, astronomers stumbled upon a fascinating finding: Thousands of black holes likely exist near the center of our galaxy.

The X-ray images that enabled this discovery weren’t from some state-of-the-art new telescope. Nor were they even recently taken—**some of the data was collected nearly 20 years ago.**”

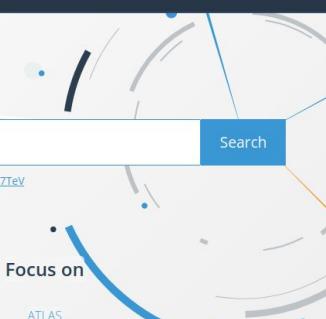
Big Data is Transforming How Astronomers Make Discoveries

The next game-changer is likely lurking in the data we already have—but it will take scientists years to uncover it



<https://www.smithsonianmag.com/science-nature/next-big-discovery-astronomy-scientists-probably-found-it-years-ago-they-dont-know-it-yet-180969073/>

Explore more than **two petabytes**
of open data from particle physics!



Start typing...

search examples: collision datasets, keywords:education, energy:7TeV

Explore

[datasets](#)
[software](#)
[environments](#)
[documentation](#)

Focus on

ATL
ALL
CN
LH
OPP

Registry of Open Data on AWS

About

This registry exists to help people discover and share datasets that are available via AWS resources. [Learn more about sharing data on AWS](#)

[See all usage examples for datasets listed in this request](#)

See datasets from Facebook Data for Good, NASA Space Act Agreement, NIH STRIDES, NOAA Big Data Program, Space Telescope Science Institute, and Amazon Sustainability Data Initiative.

Search datasets (currently 205 matching datasets)

Search database

Add to this register

If you want to add a dataset or example of how to use a dataset to the registry, please follow the instructions on the [Registry of Open Data](#) [AWS GitHub repository](#).

Unless specifically stated in the applicable dataset documentation, datasets available through the Registry of Open Data on AWS are not provided and maintained by AWS. Datasets are provided and maintained by a variety of third parties under a variety of licenses. Please check dataset licenses and related documentation to determine if a dataset can be used for your application.



 https://registry.opend



The Cancer Genome Atlas

cancer genomic life sciences STRIDES whole genome sequencing

The Cancer Genome Atlas (TCGA), a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), aims to generate comprehensive, multi-dimensional maps of the key genomic changes in major types and subtypes of cancer. TCGA has analyzed matched tumor and normal tissues from 11,000 patients, allowing for the comprehensive characterization of 33 cancer types and subtypes, including 10 rare cancers. The dataset contains open Clinical Supplement, Biopspecimen Supplement, RNA-Seq Gene Expression Quantification, miRNA-Seq Isoform Expression Quantification, and Expression Quantification.

Details

Usage example

- Broad Institute FireCloud by The Broad Institute of MIT & Harvard
 - An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics by Jianfang Liu, Tara Lichtenberg, et al.
 - The Immune Landscape of Cancer by Vésteinn Thorsson, David L. Gibbs, et al.
 - The chromatin accessibility landscape of primary human cancers by M. Ryan Corces, Jeffrey M. Granja, et al.
 - Molecular Characterization and Clinical Relevance of Metabolic Expression Subtypes in Human Cancers by Xinyia Peng, Zhongxuan Chen, et al.

[See 20 usage examples](#)

Therapeutically Applicable Research to Generate

SATÉLITES

https://www.geospatialworld.net/news/hedge-funds-use-satellite-imagery-to-predict-revenues

ABOUT US NEWS ARTICLES BLOGS VIDEOS WEBINARS EVENTS CONTACT US f G+ in t d

GEOSPATIAL WORLD™

GEOSPATIAL KNOWLEDGE INFRASTRUCTURE SUMMIT

Theme: Mapping our way to 4IR 24 - 25 FEBRUARY 2021 / 0700 - 1100 (EST)

REGISTER NOW

GIS & Maps Earth Observation GNSS & Positioning LiDAR Location Tech UAVs BIM & Modelling Trending Tech GW Weekly

Home > News > Business > Hedge funds use satellite imagery to predict revenues

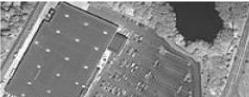
NEWS BUSINESS GENERAL

Hedge funds use satellite imagery to predict revenues

By Geospatial World - 09/17/2013 < 1 Minute Read

Share f t p in

US: Analysts at hedge funds are using analysis of satellite imagery to learn more about business and predict revenues. They are including their findings in the quarterly and annual reports also.



Recently, UBS Investment Research issued its earnings preview for Wal-Mart's second quarter, which publicly revealed that UBS used satellite imagery to gather data about the parking lots at Wal-

Upcoming Webinar:

Geospatial Knowledge Infrastructure Summit: Mapping our way to 4IR

Date: 24 - 25 February, 2021
Time: 7:00 - 11:00 (EST)

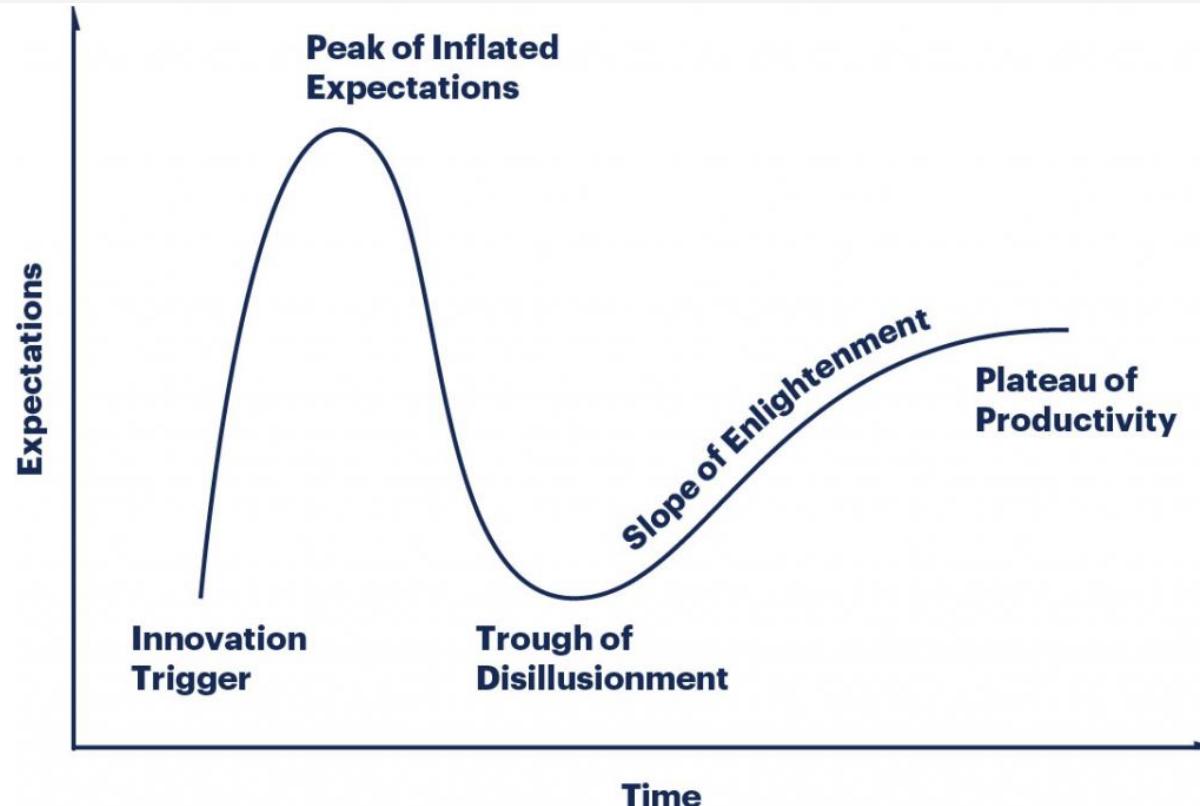
GW WEEKLY
Your Weekly Go-to Guide for Everything Geospatial

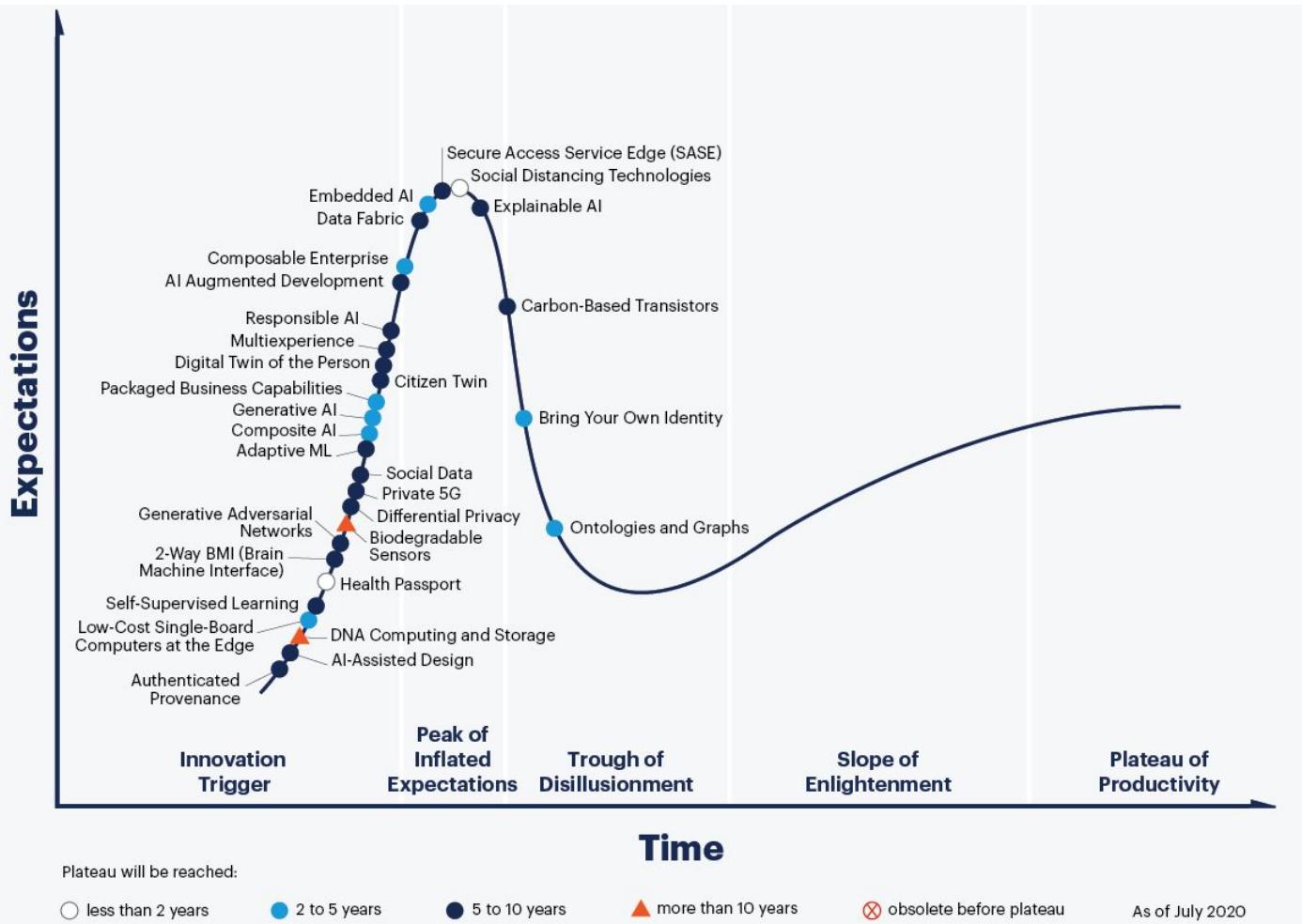
Get it delivered to your mailbox every Monday

Email *

Comments *

HYPE CYCLE (GARTNER)





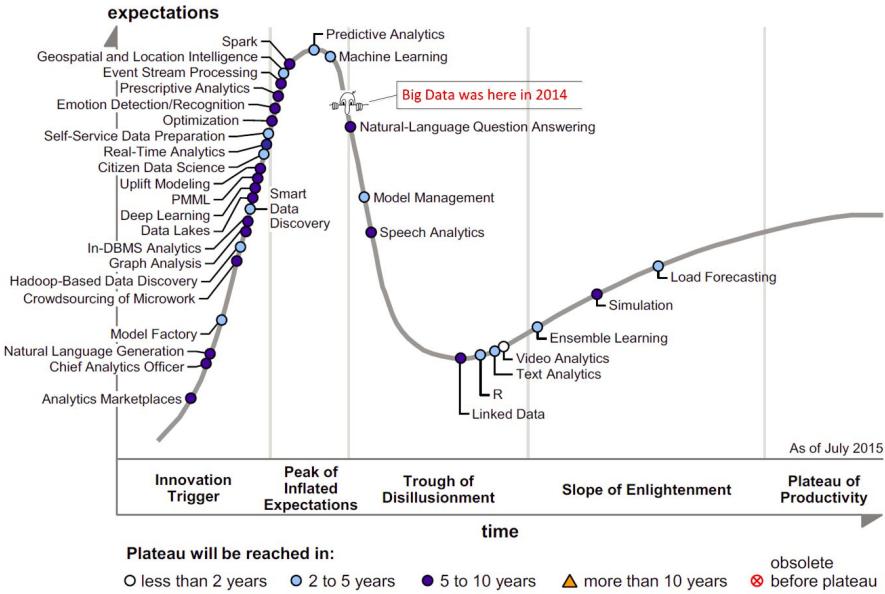
Big Data Falls Off the Hype Cycle

<https://www.datasciencecentral.com/profiles/blogs/big-data-falls-off-the-hype-cycle>

Why Gartner Dropped Big Data Off the Hype Curve

<https://www.datanami.com/2015/08/26/why-gartner-dropped-big-data-off-the-hype-curve/>

Figure 1. Hype Cycle for Advanced Analytics and Data Science, 2015



Source: Gartner (July 2015)

WATSON (DEEP BLUE)



WATSON (DEEP BLUE)



The screenshot shows the IBM Watson Health Oncology homepage. At the top, there's a navigation bar with the IBM logo, a Marketplace button, a search bar, and user icons. Below the header, the page title is "Watson for Oncology". A main heading "Watson for Oncology" is followed by a subtext: "Passe menos tempo procurando na literatura e no prontuário médico eletrônico e mais tempo tratando dos pacientes. Watson fornece aos médicos opções de tratamento baseadas em evidências e no treinamento dos médicos especialistas do Memorial Sloan Kettering (MSK)." Below this, there are two buttons: "Veja como funciona (05:07)" and "Obtenha os fatos (US)". To the right of the text, there's a photograph of a male doctor in a white coat and stethoscope, sitting at a desk and looking at a laptop screen. The background of the main content area is dark green.

Oncology and Genomics

Watson for Oncology

Passe menos tempo procurando na literatura e no prontuário médico eletrônico e mais tempo tratando dos pacientes. Watson fornece aos médicos opções de tratamento baseadas em evidências e no treinamento dos médicos especialistas do Memorial Sloan Kettering (MSK).

Veja como funciona (05:07) Obtenha os fatos (US)

Ajude a identificar opções de tratamento baseadas em evidência e centradas no paciente

A quantidade de pesquisas e dados disponíveis para ajudar na informação dos tratamentos contra o câncer está crescendo exponencialmente. No entanto, o tempo que as equipes de assistência possuem para consumir estas informações—buscando insights específicos para as necessidades de cada paciente para potencialmente melhorar os resultados do tratamento—é mais limitado do que

Watson for Oncology ajuda médicos a identificar rapidamente informações importantes no registro médico de um paciente, buscar artigos relevantes e explorar opções de tratamento para reduzir a variação indesejada de assistência e devolver tempo aos seus pacientes.

Vamos conversar

https://www.youtube.com/watch?v=HkEOJnn_zlg

GIZMODO
BRASIL

ESPECIAIS GALERIAS REVIEWS

INTELLIGÊNCIA ARTIFICIAL

IBM Watson teria recomendado tratamentos contra câncer “inseguros e incorretos”

Por: Jennings Brown
25 de julho de 2018 às 18:36



Documentos corporativos internos da IBM mostram que especialistas médicos trabalhando com o supercomputador Watson, da empresa, encontraram “vários exemplos de recomendações de tratamento inseguras e incorretas” ao usarem o software, segundo uma reportagem do *Stat*.

O site revisou documentos que foram incluídos em duas apresentações feitas em junho e julho de

<https://gizmodo.uol.com.br/ibm-watson-saude-recomendacao-tratamentos-cancer-inseguros-incorretos/>

<https://www.technologyreview.com/s/607965/a-reality-check-for-ibms-ai-ambitions/>

<https://gizmodo.com/why-everyone-is-hating-on-watson-including-the-people-w-1797510888>

GADGETS

Why Everyone Is Hating on IBM Watson —Including the People Who Helped Make It

MIT
Technology
Review

Log in / Create an account Search

Topics+ The Download Magazine Events More+

Subscribe

MIT Technology Review Global Panel [JOIN NOW](#)

Business Impact

A Reality Check for IBM's AI Ambitions



IBM, number 39 on our list of the 50 Smartest Companies, overhyped its Watson machine-learning

GOOGLE FLU (2013)



google.org Flu Trends

[Google.org home](#)

[Dengue Trends](#)

Flu Trends

[Home](#)

[Select country/region](#)

[How does this work?](#)

[FAQ](#)

Flu activity

Intense

High

Moderate

Low

Minimal

Explore flu trends around the world

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. [Learn more](#)



GOOGLE FLU

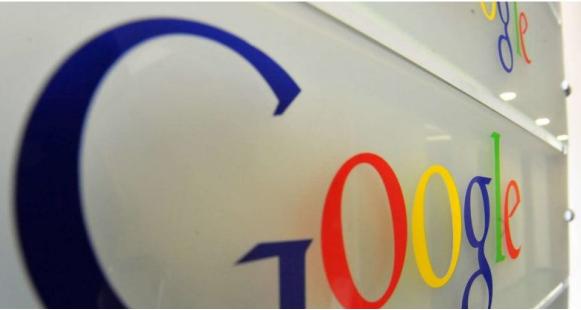
≡ TIME

U.S. POLITICS WORLD TECH TIME HEALTH ENTERTAINMENT SUBSCRIBE  

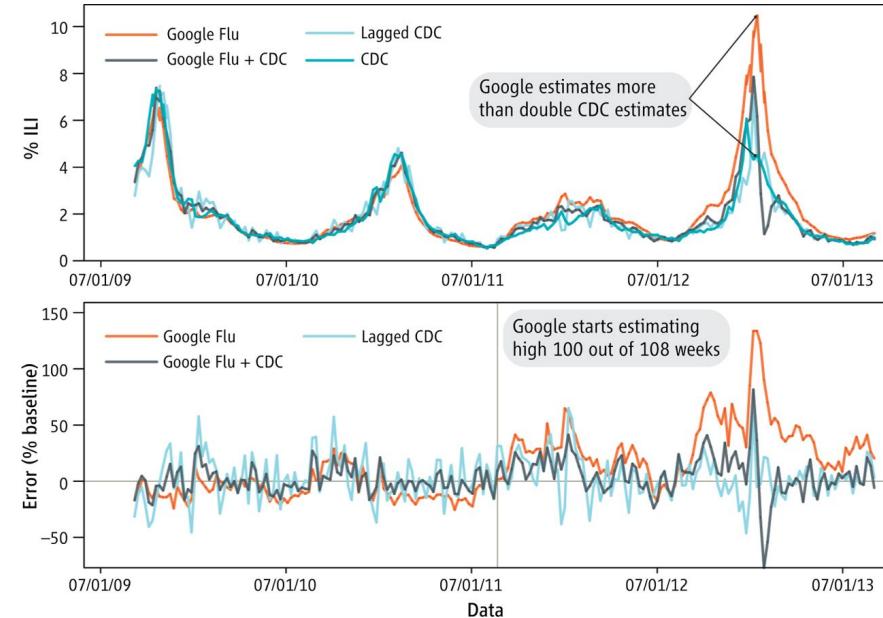
SCIENCE + BIG DATA

Google's Flu Project Shows the Failings of Big Data



Stories From  Ideal Media



<http://science.sciencemag.org/content/343/6176/1203>

GOOGLE FLU

DAVID LAZER AND RYAN KENNEDY OPINION 10.01.15 07:00 AM

WHAT WE CAN LEARN FROM THE EPIC FAILURE OF GOOGLE FLU TRENDS



RAFE SWAN/GETTY IMAGES

EVERY DAY, MILLIONS of people use Google to dig up information that drives their daily lives, from how long their commute will be to how to treat their child's illness. This search data reveals a lot about the searchers: their wants,

<https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/>

A screenshot of the Science magazine website. At the top, there is a banner for a Sony MA900 Multi-Application Cell Sorter. Below the banner, the word "Science" is prominently displayed, followed by links to Home, News, Journals, Topics, and Careers. The main article title is "The Parable of Google Flu: Traps in Big Data Analysis". The authors listed are David Lazer^{1,2*}, Ryan Kennedy^{1,3,4}, Gary King³, Alessandro Vespignani^{5,6,7}. The article was published on 14 Mar 2014, Vol. 343, Issue 6176, pp. 1203-1205, DOI: 10.1126/science.1248506. There are also links for Article, Figures & Data, Info & Metrics, eLetters, and PDF.

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (1, 2). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (3, 4), what lessons can we draw from this error?



CREDIT: ADAPTED FROM AXEL KOPES/DESIGN & ART
DIRECTION: ISTOCKPHOTO.COM

<http://science.sciencemag.org/content/343/6176/1203.full>



A má interpretação da posse de bola e os enganos trazidos pelos números frios no Brasil

mais de 3 anos

Variação de sistema, mais jogo curto e presença ofensiva: como funciona o "novo" Atlético de Diego Simeone?

1 mês

Por que o gol da vitória do São Paulo sobre o Flamengo vai muito além do erro de Hugo Souza?

2 meses

Sai Coupet, entra Abel: quais as consequências de tais mudanças no Internacional?

3 meses

A má interpretação da posse de bola e os enganos trazidos pelos números frios no Brasil



Renato Rodrigues

19 Sep, 2017



Mourinho e Guardiola sempre travaram grandes duelos de times com e sem bola Getty Images

O Campeonato Brasileiro de 2017 tem levantado um grande debate sobre a posse de

C ⌂ ⌂ https://www.digitalnewsasia.com/insights/why-85-big-data-projects-fail



Digital Economy
powered by 'MDEC' Insights 'Software Testing is'
Business Personal Tech Nat
Infra

DNA
DIGITAL NEWS ASIA
Your Eye on the Tech Ecosystem

Why 85% of Big Data projects fail

By Sharala Axyrd April 16, 2019

- People from the top must define clear problem statements
- Deciding to become data driven can be a long, difficult process



<https://www.digitalnewsasia.com/insights/why-85-big-data-projects-fail>

https://www.datanami.com/2020/10/01/most-data-science-projects-fail-but-yours-doesnt-have-to/



DATANAMI
DATA SCIENCE • AI • ADVANCED ANALYTICS

About Resources Events **Subscribe**

HOME COVID-19 FEATURES SECTORS APPLICATIONS TECHNOLOGIES V

October 1, 2020

Most Data Science Projects Fail, But Yours Doesn't Have To

Ryohei Fujimaki



In an effort to remain competitive in today's increasingly challenging economic times, companies are moving forward with digital transformations — powered by data science and machine learning — at an unprecedented rate. According to PwC's global study, AI will provide up to 26% boost in GDP for local economies by 2030. Yet, for many companies, implementing data science into various aspects of their businesses can prove difficult if not daunting.

According to Gartner analyst Nick Heudecker, over 85% of data science projects fail. A report from Dimensional Research indicated that only 4% of companies have succeeded in deploying ML models to production environment.

<https://www.datanami.com/2020/10/01/most-data-science-projects-fail-but-yours-doesnt-have-to/>

Oct 20, 2020, 01:14pm EDT | 7,786 views

The ‘Failure’ Of Big Data

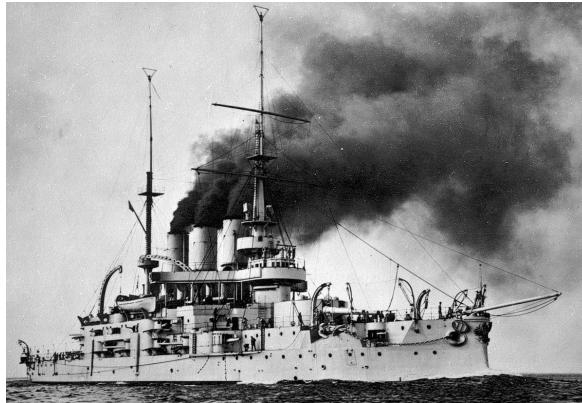


Randy Bean Contributor

CIO Network

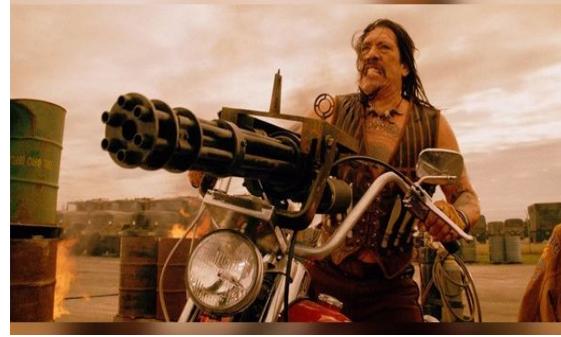
“Yet, the connection of data investments to business insights and successful business outcomes remains an elusive ambition for most.”











Estatística

Matemática formal

Relações entre as variáveis

Insights

Machine Learning

Algoritmos

Black box

Foco nos resultados

“There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems.”

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

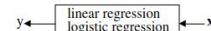
Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables \mathbf{x} (independent variables) go in one side, and on the other side the response variables \mathbf{y} come out. Inside the black box, nature functions to

The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



SMALL DATA

BIG
DATA
tells us
what's
happening.



small data tells us
why it's
happening.
A close-up photograph of a single green leaf, symbolizing small data.

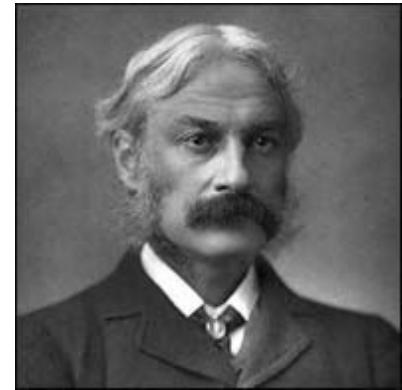
THICK DATA

TO FORM A COMPLETE PICTURE, BOTH BIG AND THICK DATA ARE CRITICAL BECAUSE THEY PRODUCE DIFFERENT TYPES OF INSIGHTS AT VARYING SCALES AND DEPTHS



“Alguns usam a estatística como os bêbados usam postes: mais para apoio do que para iluminação.”

Andrew Lang



STELLA CUNLIFFE – GUINNESS

“Como parte do processo de controle de qualidade, os estatísticos acompanhavam as medidas da capacidade dos barris e quais eram descartados. Ao examinar o gráfico de medidas de capacidade, Stella Cunliffe se deu conta de que havia um número incomumente alto de barris que passavam no teste por muito pouco, e um número incomumente baixo de barris que eram rejeitados por muito pouco. Examinaram as condições de trabalho da mulher que media os barris. Ela devia jogar um barril descartado no alto de uma pilha e colocar um barril aprovado em uma esteira transportadora. Por sugestão de Stella Cunliffe, a posição da balança foi deslocada para acima do depósito de barris descartados. Então, tudo que ela tinha de fazer era chutar o barril rejeitado para o depósito. O excesso de barris que eram aprovados por pouco desapareceu. “



(Uma senhora toma chá... Como a estatística revolucionou a ciência no século XX. Capítulo 25)

GENCHI GENBUTSU – GEMBA

Criado na década de 80 por Masaaki Imai, Genchi Genbutsu significa “vá e veja por você mesmo”. Princípio que estabelece a primazia do conhecimento tácito e pessoal, e que o primeiro passo para a solução de qualquer problema está em ir onde ele ocorre.

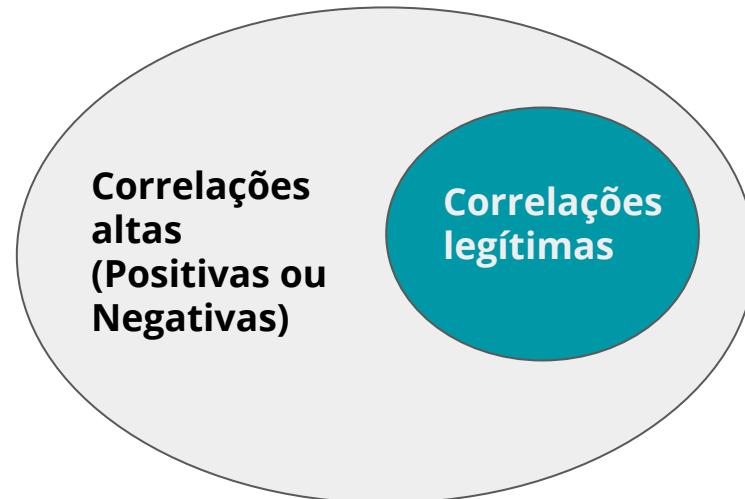


O local onde as coisas ocorrem, ou o “chão de fábrica”, é chamado de Gemba.

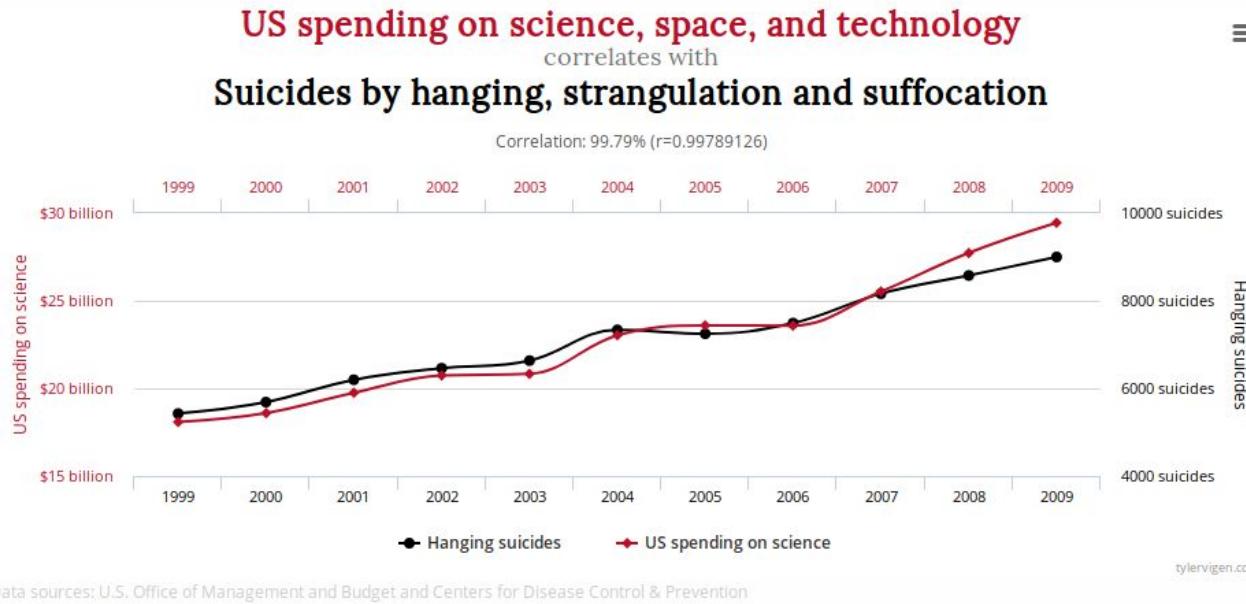
A técnica que os gestores sempre vão onde as coisas ocorrem é chamada Gemba Walk e é utilizada para observar e compreender como o trabalho é feito.



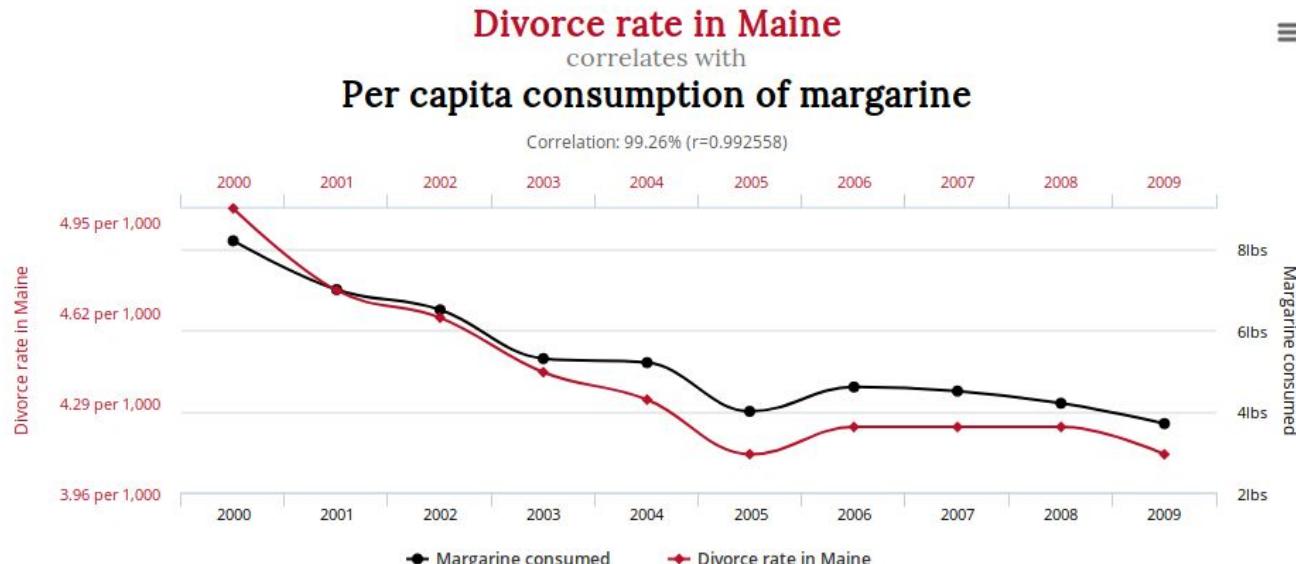
CORRELAÇÃO



CORRELAÇÃO ESPÚRIA



CORRELAÇÃO ESPÚRIA

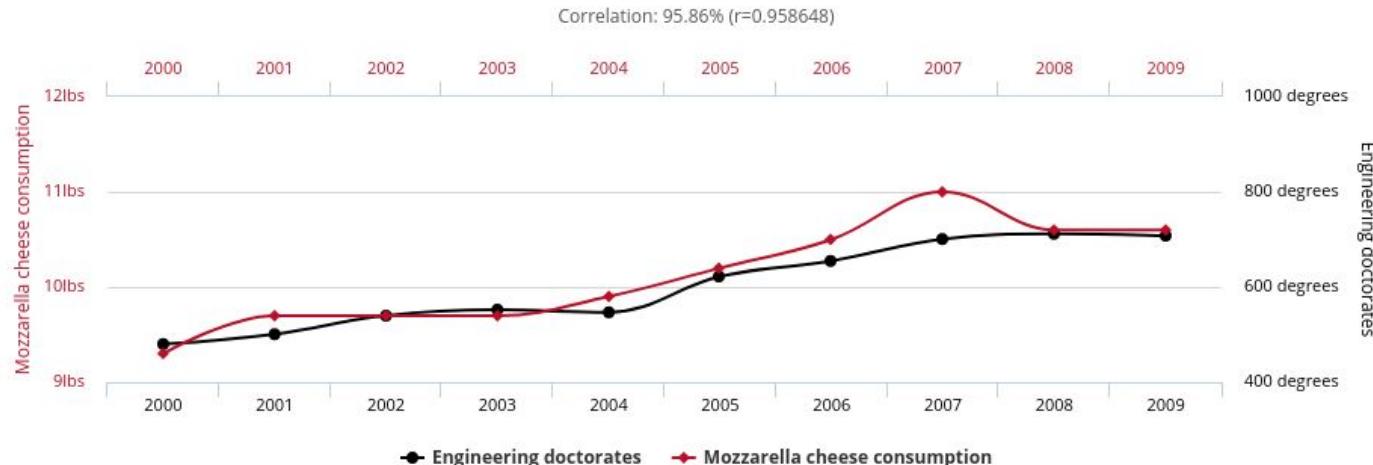


Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

tylervigen.com

CORRELAÇÃO ESPÚRIA

Per capita consumption of mozzarella cheese
correlates with
Civil engineering doctorates awarded



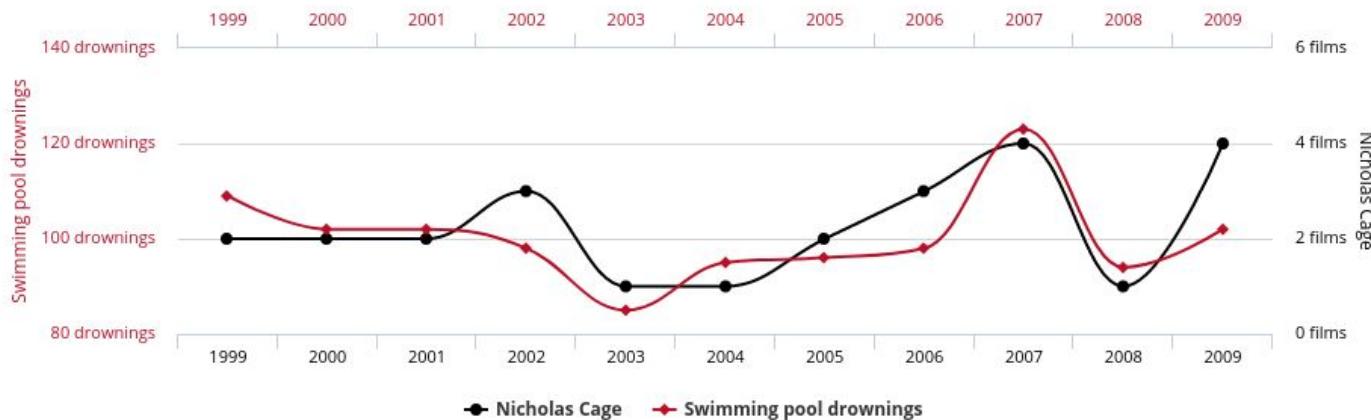
Data sources: U.S. Department of Agriculture and National Science Foundation

tylervigen.com

CORRELAÇÃO ESPÚRIA

Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in

Correlation: 66.6% ($r=0.666004$)



Data sources: Centers for Disease Control & Prevention and Internet Movie Database

tylervigen.com

LIMITE RAM



Bigmemory, biganalytics,

Opções de trabalhar com base de dados sem carregá-las na RAM (< 10GB):

R: ff, bigmemory, biganalytics, bigtabulate

Python: Dask

https://rpubs.com/msundar/large_data_analysis

ORACLE R DISTRIBUTION

Oracle R Distribution



Ability to dynamically load

Intel Math Kernel Library (MKL)
AMD Core Math Library (ACML)
Solaris Sun Performance Library

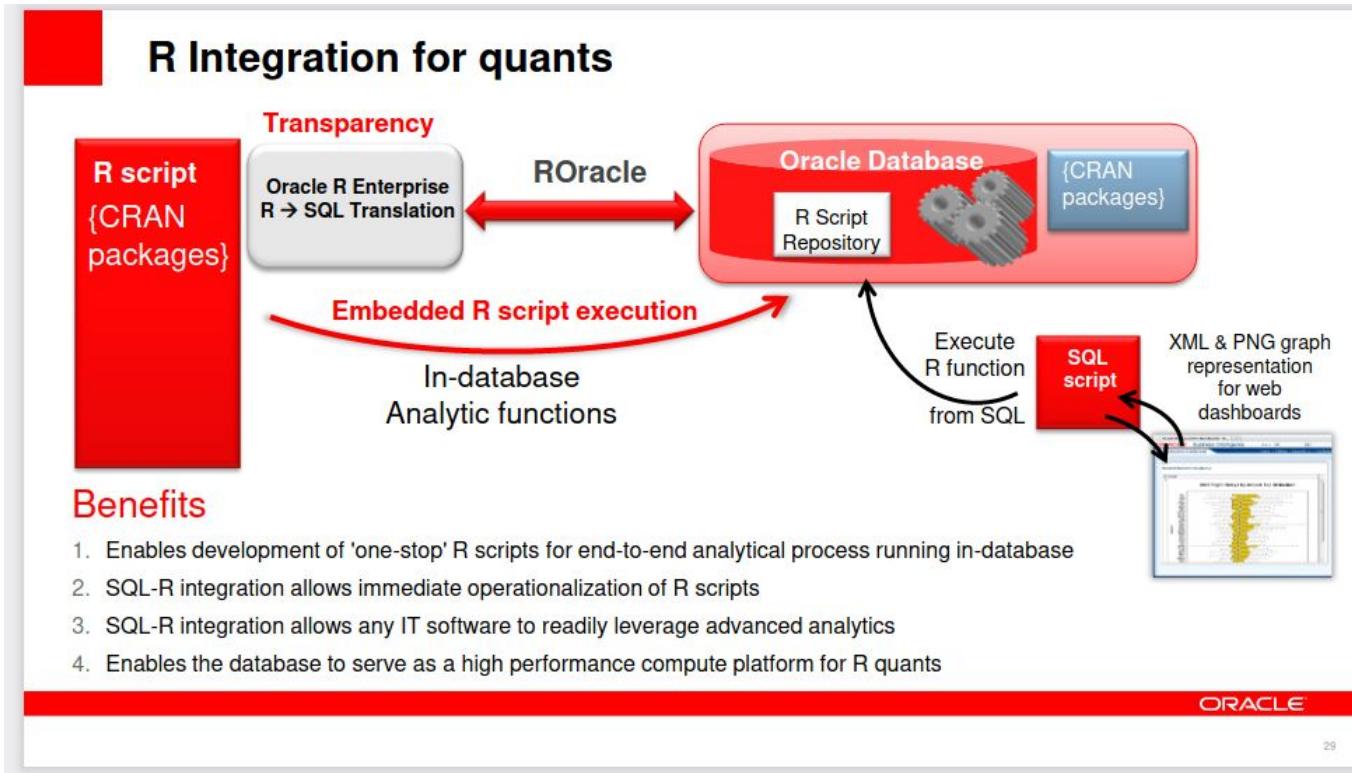


Oracle
Support

- Oracle's redistribution of open source R
- Enhanced linear algebra performance using Intel's MKL, AMD's ACML, and Sun Performance Library for Solaris
- Improve R scalability at client and at database server for embedded R execution
- Enterprise support for customers of Oracle Advanced Analytics option, Big Data Appliance, and Oracle Linux
- **Free** download
- Oracle makes bug fixes and enhancements available for open source R

ORACLE

ORACLE R DISTRIBUTION



ORACLE R DISTRIBUTION

<https://www.oracle.com/br/database/technologies/datawarehouse-bigdata/oml4r.html>

<https://oss.oracle.com/ORD/>

Documentos do SQL

Visão geral ▾ Instalar ▾ Seguro ▾ Desenvolver ▾ Administrar ▾ Analisar ▾ Referência ▾

Baixar o SQL Server

Docs / SQL / ML / Visão geral /

O que são os Serviços de Machine Learning (Python e R)?

 Indicador Comentários Editar Compartilhar

Ler em inglês



Versão

SQL Server 2019

 Filtrar por títuloDocumentação do aprendizado de
máquina do SQL

Documentação do Microsoft SQL >

Visão geral

O que são os Serviços de
Machine Learning (Python e R)?

Servidor autônomo

Novidades

> Instalar

> Inícios rápidos

> Tutoriais

Exemplos >>

O que são os Serviços de Machine Learning do SQL Server com Python e R?

10/11/2020 • 4 minutos para o fim da leitura •  Aplica-se a:  SQL Server 2017 (14.x) e  Instância Gerenciada do Azure SQL mais recente

Os Serviços de Machine Learning são um recurso no SQL Server que possibilita executar scripts do Python e do R usando dados relacionais. Você pode usar pacotes e estruturas de software livre, bem como os [pacotes do R e do Python da Microsoft](#) para análise preditiva e aprendizado de máquina. Os scripts são executados no banco de dados sem mover dados para fora do SQL Server ou pela rede. Este artigo explica os conceitos básicos dos Serviços de Machine Learning do SQL Server e como começar.

Para aprendizado de máquina em outras plataformas do SQL, confira a [documentação do aprendizado de máquina do SQL](#).

 Observação

Esta página é útil?

 Yes No

Neste artigo

 Executar scripts do
Python e do R no
SQL ServerIntrodução aos
Serviços de Machine
LearningVersões do Python e
do RPacotes do Python e
do R

Próximas etapas

MICROSOFT R OPEN

Microsoft R Application Network

Home About R Microsoft R Open R Packages R Community R Tools

Find an R Package 

Welcome to MRAN

Download Microsoft R Open 4.0.2 now.

R is the world's most powerful programming language for statistical computing, machine learning and graphics and has a thriving global community of users, developers and contributors.



Microsoft R Open
Microsoft R Open is the enhanced distribution of R from Microsoft Corporation. The current release, Microsoft R Open 4.0.2, is based the statistical language R-4.0.2 and includes additional capabilities for improved performance, reproducibility and platform support.

[!\[\]\(2b8ee4c0d46a8fa5c93fc1637df42851_img.jpg\) Download Now](#)

R Packages
Packages extend R with new function and data. Whether you're using R to optimize portfolio, analyze genomic sequences, or to predict component failure times, experts in every domain have made resources, applications and code available for free online.

[!\[\]\(e09c6b6a5121796ae9e08b55efcbc15f_img.jpg\) Explore Packages](#)

CRAN Time Machine
For the purpose of **reproducibility**, MRAN hosts **daily snapshots** of the CRAN R packages and R releases as far back as Sept. 17, 2014.
Use our **Time Machine** to browse CRAN contents from the past.

[!\[\]\(4e41e3d4ce57fdb2429f77aea497a461_img.jpg\) Browse Snapshots](#)



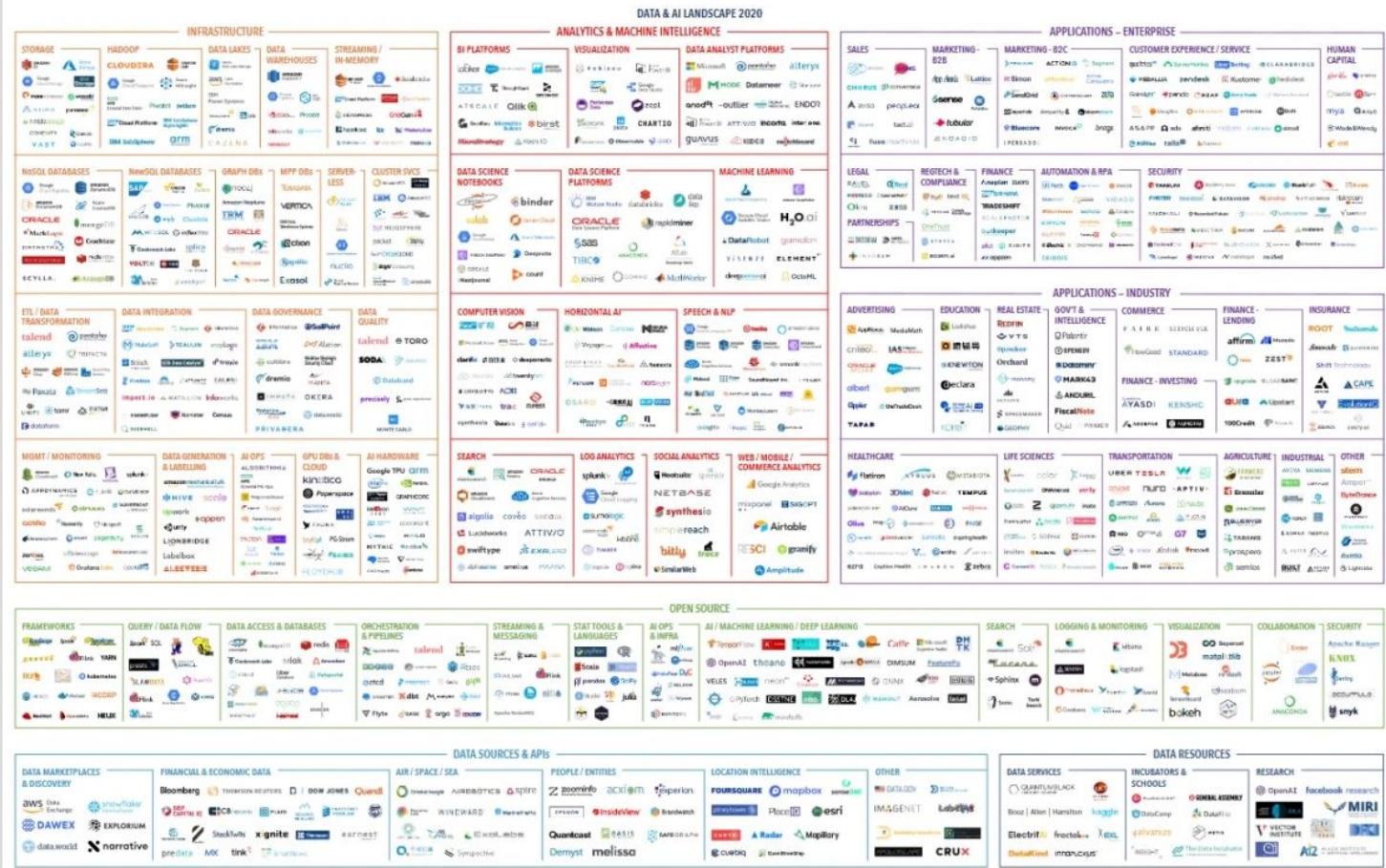
APACHE
HBASE



Flink



CLOUDERA



Version 1.0 - September 2020

© Matt Turck (@mattturck) & FirstMark (@firstmarkcap)

mattturck.com/data2020

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

<https://venturebeat.com/2020/10/21/the-2020-data-and-ai-landscape/>

HISTÓRICO

APACHE LUCENE (1997)

Open Source Search Engine



Doug Cutting

Apache 2.0 licensed

Apache Lucene and Solr are distributed under a commercially friendly Apache Software license

Welcome to Apache Lucene

The Apache Lucene™ project develops open-source search software. The project releases a core search library, named Lucene™ core, as well as the Solr™ search server.

LUCENE

Lucene Core is a Java library providing powerful indexing and search features, as well as spellchecking, hit highlighting and advanced analysis/tokenization capabilities. The PyLucene sub project provides Python bindings for Lucene Core.

SOLR

Solr™ is a high performance search server built using Lucene Core. Solr is highly scalable, providing fully fault tolerant distributed indexing, search and analytics. It exposes Lucene's features through easy to use RESTful interfaces or nothing clients for Java and

DOWNLOAD
Apache Lucene 8.7.0

DOWNLOAD
Apache Solr 8.7.0

Projects

- Lucene Core (Java)
- Solr
- PyLucene
- Open Relevance (Discontinued)

<https://lucene.apache.org/>

APACHE NUTCH (2001)

Web Crawler



Doug Cutting



Mike Cafarella

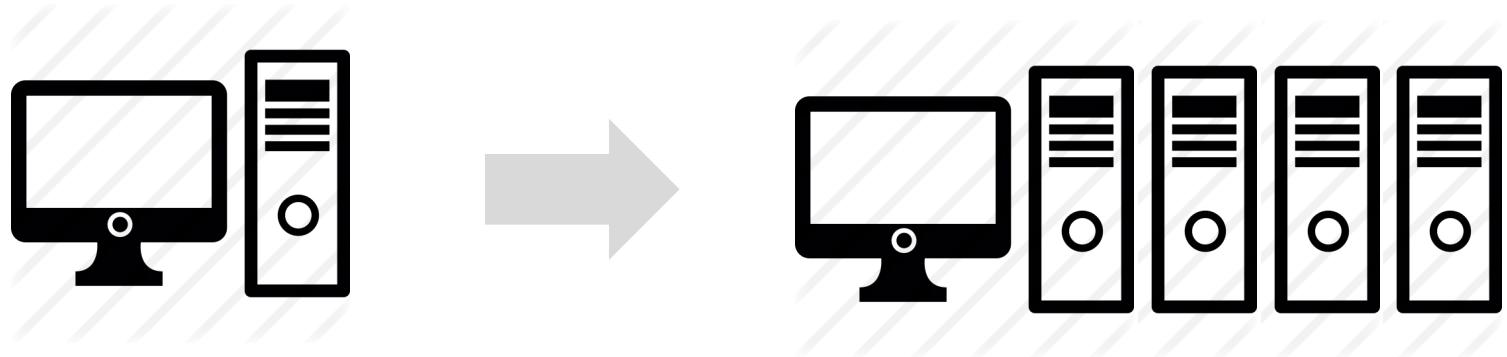
A screenshot of the Apache Nutch website homepage. The header features the Nutch logo and navigation links for Downloads, Community, Documentation, and Development. A search bar is also present. The main content area has a dark background with a blurred image of network cables and glowing green lights. The text "Pluggable parsing, protocols, indexing and more" is displayed prominently, followed by a description of Nutch's modular architecture and its integration with Apache Tika and Solr. A blue "Learn About" button is located at the bottom left of the main content area.

<http://nutch.apache.org/>

APACHE NUTCH

Começou em apenas uma máquina (single-core, 1GB de RAM, 1TB). O limite de indexação ficava em 100 milhões de páginas.

Aumento para 4 máquinas. Porém, a manutenção e gestão da utilização das máquinas era manual (além do problema de gestão crescer exponencialmente com o aumento no número de máquinas)



NECESSIDADES DO NUTCH

“

- ***schemaless*** with no predefined structure, i.e. no rigid schema with tables and columns (and column types and sizes)
- ***durable*** once data is written it should never be lost
- ***capable of handling component failure*** without human intervention (e.g. CPU, disk, memory, network, power supply, MB)
- ***automatically rebalanced*** to even out disk space consumption throughout cluster

“

<https://medium.com/@markobonaci/the-history-of-hadoop-68984a11704>

GOOGLE FILE SYSTEM (2003)

The screenshot shows a web browser displaying the Google Research Publications page. The URL in the address bar is <https://research.google/pubs/pub51/>. The page navigation menu includes links for Google Research, Philosophy, Research Areas, Publications (which is underlined), People, Tools & Downloads, Outreach, Careers, and Blog. A decorative graphic of a red geometric shape is visible on the right side of the page. The main content area displays the title 'The Google File System' and the authors 'Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung'. Below the title is the citation 'Proceedings of the 19th ACM Symposium on Operating Systems Principles, ACM, Bolton Landing, NY (2003), pp. 20-43'. At the bottom of this section are three buttons: 'Download', 'Google Scholar', and 'Copy Bibtex'.

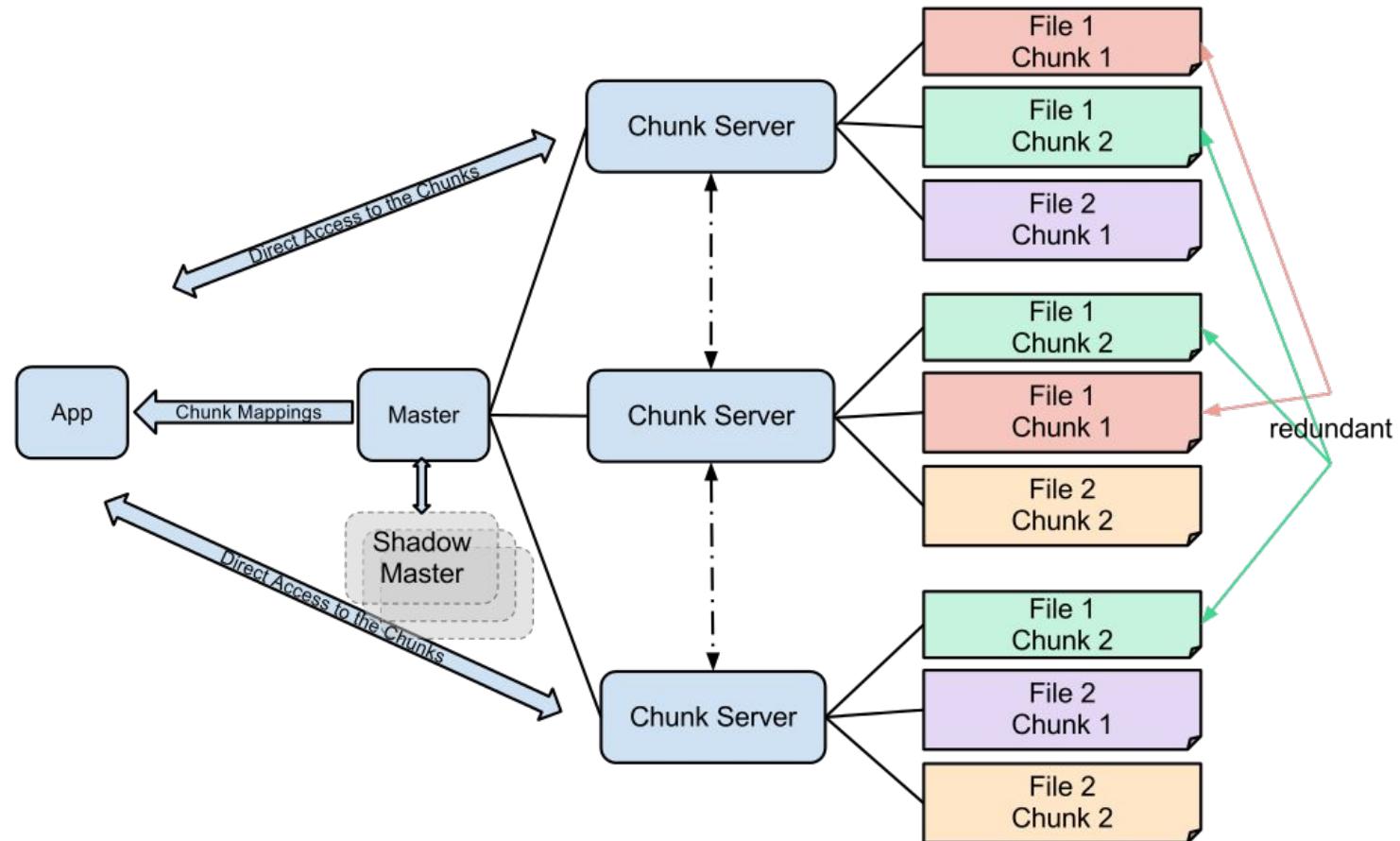
Abstract

We have designed and implemented the Google File System, a scalable distributed file system for large distributed data-intensive applications. It provides fault tolerance while running on inexpensive commodity hardware, and it delivers high aggregate performance to a large number of clients.

While sharing many of the same goals as previous distributed file systems, our design has been driven by observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system assumptions. This has led us to reexamine traditional choices and explore radically different design points.

The file system has successfully met our storage needs. It is widely deployed within Google as the storage platform for the generation and processing of data used by our service as well as research and development efforts that require large data sets. The largest cluster to date provides hundreds of terabytes of storage across thousands of disks on over a thousand machines, and it is concurrently accessed by hundreds of clients.

In this paper, we present file system interface extensions designed to support distributed applications, discuss many aspects of our design, and



NUTCH DISTRIBUTED FILE SYSTEM (NDFS)

- Escrito em Java a partir das especificações do Google File System (GFS)
- Arquivos divididos em pedaços de 64MB em 3 diferentes nós (replicação automática)

MAP REDUCE (2004)

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

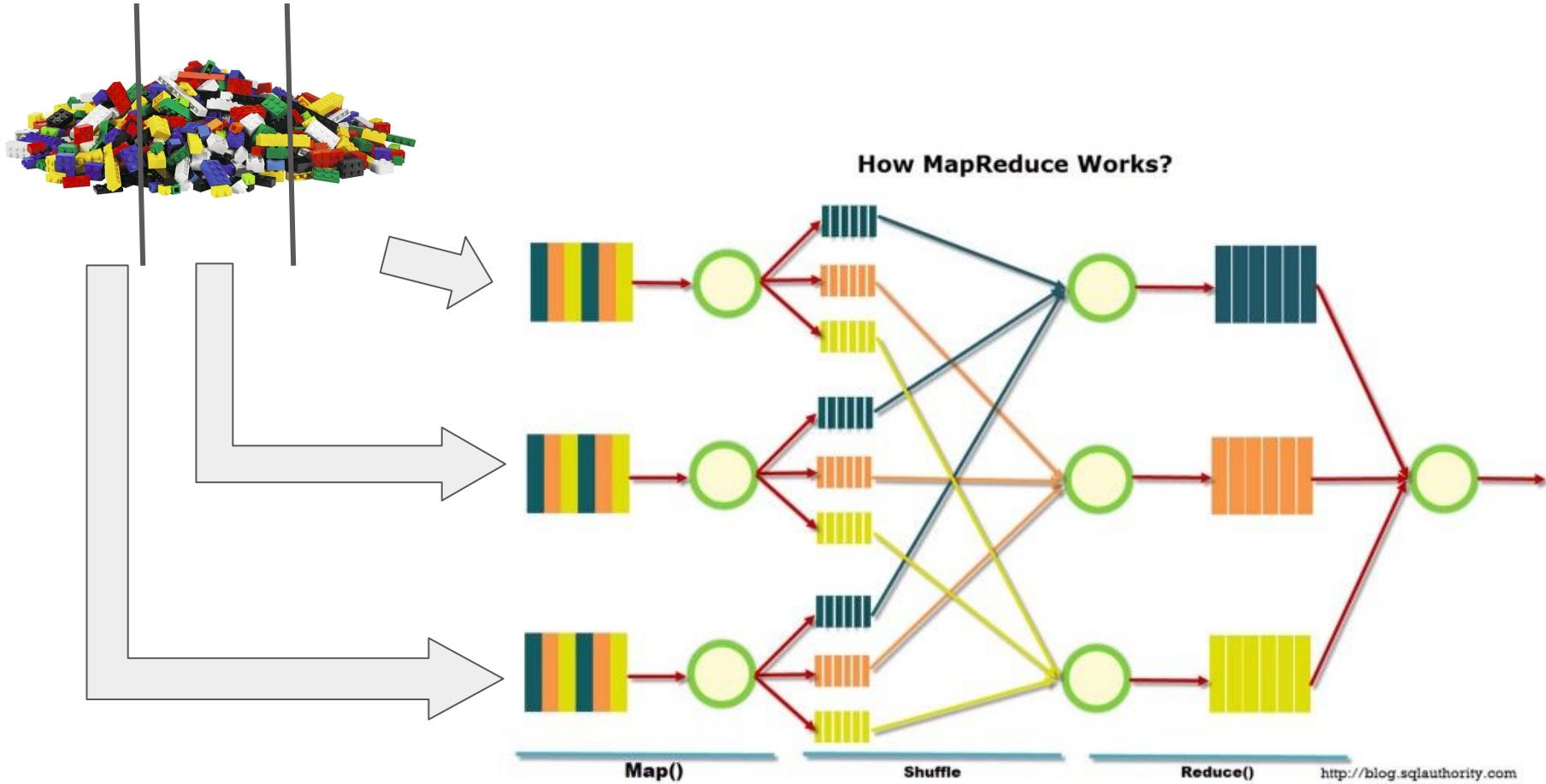
Abstract

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





HADOOP (2006)

2005 - Map Reduce incluído no Nutch

2006 - NDFS e MapReduce retirados do Nutch e é estabelecido o projeto Hadoop, dentro do Lucene (NDFS se torna HDFS)



2006 em diante ...

- 2006** - Yahoo contrata Cutting para implementar o sistema Hadoop
- 2007** - Yahoo reporta rodar Hadoop em um cluster de 1.000 máquinas
- 2008** - Hadoop se “gradua” (sai debaixo do Lucene)
- 2012** - Hadoop v2.0.0 (YARN, aka NextGen MapReduce)
- 2012** - Cluster Hadoop do Yahoo com 42.000 máquinas

...

DOUG CUTTING

Doug Cutting: The Origins of Hadoop

https://www.youtube.com/watch?v=ebgXN7ValZA&feature=emb_title

Doug Cutting: The Name of Hadoop

https://www.youtube.com/watch?v=irK7xHUmkUA&feature=emb_title

PYTHON

JAVA

C++

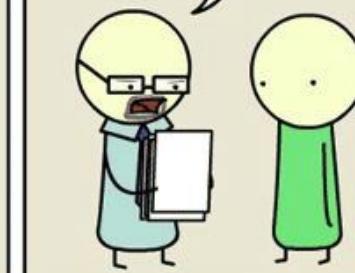
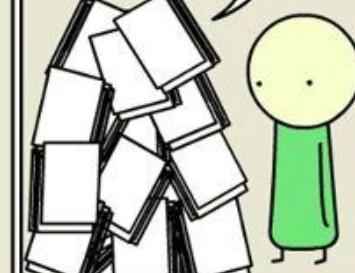
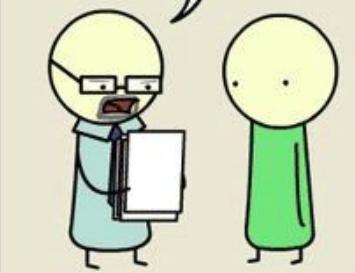
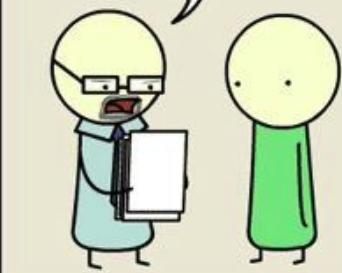
UNIX SHELL

THIS IS PLAGIARISM.
YOU CAN'T JUST "IMPORT ESSAY."

I'M TWO PAGES IN AND I STILL
HAVE NO IDEA WHAT YOU'RE SAYING.

I ASKED FOR ONE COPY,
NOT FOUR HUNDRED.

I DON'T HAVE PERMISSION TO
READ THIS.



ASSEMBLY

C

LATEX

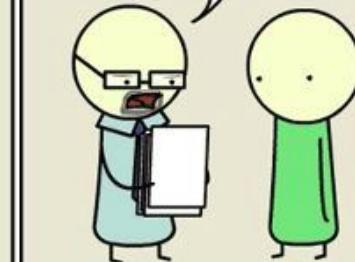
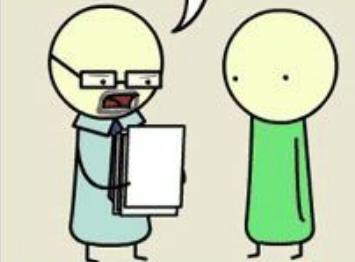
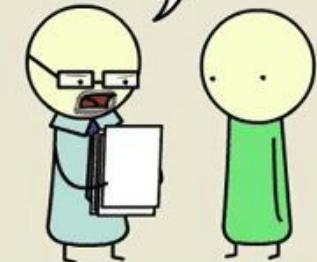
HTML

DID YOU REALLY HAVE TO REDEFINE EVERY
WORD IN THE ENGLISH LANGUAGE?

THIS IS GREAT, BUT YOU FORGOT TO ADD
A NULL TERMINATOR. NOW I'M JUST READING
GARBAGE.

YOUR PAPER MAKES NO GODDAMN SENSE,
BUT IT'S THE MOST BEAUTIFUL THING
I HAVE EVER LAID EYES ON.

THIS IS A FLOWER POT.



```
1. package com.javatpoint;
2.
3. import java.io.IOException;
4. import java.util.StringTokenizer;
5. import org.apache.hadoop.io.IntWritable;
6. import org.apache.hadoop.io.LongWritable;
7. import org.apache.hadoop.io.Text;
8. import org.apache.hadoop.mapred.MapReduceBase;
9. import org.apache.hadoop.mapred.Mapper;
10. import org.apache.hadoop.mapred.OutputCollector;
11. import org.apache.hadoop.mapred.Reporter;
12. public class WC_Mapper extends MapReduceBase implements Mapper<LongWritable,Text,Text,IntWritable>{
13.     private final static IntWritable one = new IntWritable(1);
14.     private Text word = new Text();
15.     public void map(LongWritable key, Text value,OutputCollector<Text,IntWritable> output,
16.         Reporter reporter) throws IOException{
17.         String line = value.toString();
18.         StringTokenizer tokenizer = new StringTokenizer(line);
19.         while (tokenizer.hasMoreTokens()){
20.             word.set(tokenizer.nextToken());
21.             output.collect(word, one);
22.         }
23.     }
24. }
```

```
1. package com.javatpoint;
2. import java.io.IOException;
3. import java.util.Iterator;
4. import org.apache.hadoop.io.IntWritable;
5. import org.apache.hadoop.io.Text;
6. import org.apache.hadoop.mapred.MapReduceBase;
7. import org.apache.hadoop.mapred.OutputCollector;
8. import org.apache.hadoop.mapred.Reducer;
9. import org.apache.hadoop.mapred.Reporter;
10.
11. public class WC_Reducer extends MapReduceBase implements Reducer<Text,IntWritable,Text,IntWritable> {
12.     public void reduce(Text key, Iterator<IntWritable> values,OutputCollector<Text,IntWritable> output,
13.     Reporter reporter) throws IOException {
14.         int sum=0;
15.         while (values.hasNext()) {
16.             sum+=values.next().get();
17.         }
18.         output.collect(key,new IntWritable(sum));
19.     }
20. }
```

```
1. package com.javatpoint;
2.
3. import java.io.IOException;
4. import org.apache.hadoop.fs.Path;
5. import org.apache.hadoop.io.IntWritable;
6. import org.apache.hadoop.io.Text;
7. import org.apache.hadoop.mapred.FileInputFormat;
8. import org.apache.hadoop.mapred.FileOutputFormat;
9. import org.apache.hadoop.mapred.JobClient;
10. import org.apache.hadoop.mapred.JobConf;
11. import org.apache.hadoop.mapred.TextInputFormat;
12. import org.apache.hadoop.mapred.TextOutputFormat;
13. public class WC_Runner {
14.     public static void main(String[] args) throws IOException{
15.         JobConf conf = new JobConf(WC_Runner.class);
16.         conf.setJobName("WordCount");
17.         conf.setOutputKeyClass(Text.class);
18.         conf.setOutputValueClass(IntWritable.class);
19.         conf.setMapperClass(WC_Mapper.class);
20.         conf.setCombinerClass(WC_Reducer.class);
21.         conf.setReducerClass(WC_Reducer.class);
22.         conf.setInputFormat(TextInputFormat.class);
23.         conf.setOutputFormat(TextOutputFormat.class);
24.         FileInputFormat.setInputPaths(conf,new Path(args[0]));
25.         FileOutputFormat.setOutputPath(conf,new Path(args[1]));
26.         JobClient.runJob(conf);
27.     }
28. }
```

```
#!/usr/bin/env python3
"""mapper.py"""

import sys

for line in sys.stdin:
    line = line.strip()
    words = line.split()
    for word in words:
        print('%s\t%s' % (word, 1))
```

```
#!/usr/bin/env python3
"""reducer.py"""

from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None

for line in sys.stdin:
    line = line.strip()
    word, count = line.split('\t', 1)
    try:
        count = int(count)
    except ValueError:
        continue

    if current_word == word:
        current_count += count
    else:
        if current_word:
            print('%s\t%s' % (current_word, current_count))
        current_count = count
        current_word = word

    if current_word == word:
        print('%s\t%s' % (current_word, current_count))
```

```
tumenas@tumenas-Lenovo-G40-80:~$ su hdoop
Password:
hdoop@tumenas-Lenovo-G40-80:/home/tumenas$ cd ../hadoop
hdoop@tumenas-Lenovo-G40-80:~$ ls
apache-hive-3.1.2-bin  dfsdata  hadoop-3.3.0  tmpdata
hdoop@tumenas-Lenovo-G40-80:~$ cd hadoop-3.3.0/
hdoop@tumenas-Lenovo-G40-80:~/hadoop-3.3.0$ 
hdoop@tumenas-Lenovo-G40-80:~/hadoop-3.3.0$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hdoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [tumenas-Lenovo-G40-80]
Starting resourcemanager
Starting nodemanagers
hdoop@tumenas-Lenovo-G40-80:~/hadoop-3.3.0$ jps
31153 NodeManager
31314 Jps
30613 DataNode
30473 NameNode
30794 SecondaryNameNode
31006 ResourceManager
hdoop@tumenas-Lenovo-G40-80:~/hadoop-3.3.0$ 
```

NAMENODE

localhost:9870/dfshealth.html#tab-datanode

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Datanode Information

In service Down Decommissioning Decommissioned Decommissioned & dead
Entering Maintenance In Maintenance In Maintenance & dead

Datanode usage histogram

Disk usage of each DataNode (%)

In operation

DataNode State All Show 25 entries Search:

Node	Http Address	Last contact	Last Block Report	Used	Non DFS Used	Capacity	Blocks	Block pool used	Version
✓ tumenass-Lenovo-G40-80:9866 (127.0.0.1:9866)	http://tumenass-Lenovo-G40-80:9864	0s	1m	3.01 GB	133.93 GB	915.4 GB	62	3.01 GB (0.33%)	3.3.0

DATANODE

The screenshot shows a web browser window with the URL `localhost:9864/datanode.html` in the address bar. The page has a green header bar with the text "Hadoop" and "Overview". Below the header, the main content area displays the following information:

DataNode on tumenas-Lenovo-G40-80:9866

Cluster ID:	CID-58f36133-4d3b-4a24-9389-7c7920666e3b
Version:	3.3.0, raa96f1871bfd858f9bac59cf2a81ec470da649af

Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)
localhost:9000	BP-217755931-127.0.1.1-1607461528932	RUNNING	1s	12 minutes	694 B (128 MB)

Volume Information

Directory	StorageType	Capacity Used	Capacity Left	Capacity Reserved	Reserved Space for Replicas	Blocks
/home/hadoop/dfsdata/datanode	DISK	3.01 GB	731.89 GB	0 B	0 B	62

Hadoop, 2020.

YARN

The screenshot shows the Hadoop YARN cluster web interface at `localhost:8088/cluster`. The title bar features the Hadoop logo and the text "All Applications".

The left sidebar has a "Cluster" section with the following navigation items:

- About
- Nodes
- Node Labels
- Applications
 - NEW
 - NEW SAVING
 - SUBMITTED
 - ACCEPTED
 - RUNNING
 - FINISHED
 - FAILED
 - KILLED
- Scheduler

Below the sidebar is a "Tools" button.

The main content area displays the following metrics:

Cluster Metrics	Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memor
	0	0	0	0	0	0 B	8 GB

Cluster Nodes Metrics	Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes
	1	0	0	0

Scheduler Metrics	Scheduler Type	Scheduling Resource Type	Minimum Allocation
	Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>

Below these tables is a table header:

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Contalners
----	------	------	------------------	------------------	-------	----------------------	-----------	------------	------------	-------	-------------	--------------------

A message below the table header states: "No data available in table".

At the bottom, it says "Showing 0 to 0 of 0 entries".

Project Gutenberg

About ▾ Search and Browse ▾ Help ▾

Quick search Go! Donation PayPal

Books: machado de assis (sorted by popularity)

Authors

One author name matches your search.

Subjects

One subject heading matches your search.

Sort Alphabetically

Sort by Release Date

Displaying results 1-14

DOM CASMURRO

[Dom Casmurro \(Portuguese\)](#)
Machado de Assis
259 downloads

Brazilian Tales

[Brazilian Tales](#)
Machado de Assis, Medeiros e Albuquerque, Henrique Coelho Netto, and Carmen Dolores



Quincas Borba (Portuguese)

Machado de Assis
62 downloads



Mémoires Posthumes de Braz Cubas (French)

Machado de Assis
45 downloads



Poesias Completas (Portuguese)

Machado de Assis
35 downloads



A Mao e A Luva (Portuguese)

Machado de Assis
32 downloads



Esau e Jacob (Portuguese)

Machado de Assis
27 downloads



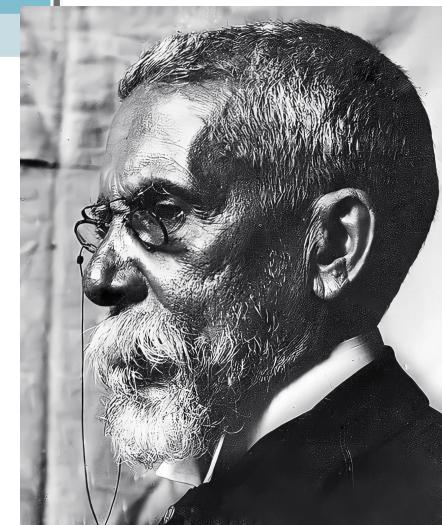
Memorial de Ayres (Portuguese)

Machado de Assis
25 downloads



De l'amour des femmes pour les sots. Portuguese (Portuguese)

Victor Hénaux
10 downloads





hadoop@tumenas-Lenovo-G40-80: ~/hadoop-3.3.0



```
hadoop@tumenas-Lenovo-G40-80:~/hadoop-3.3.0$ hadoop fs -put /home/tumenas/Aulas/BigData/dados/machado_a  
ssis/*.txt /user/input
```

```
hadoop@tumenas-Lenovo-G40-80:~/hadoop-3.3.0$ hadoop fs -ls /user/input
```

Found 9 items

-rw-r--r--	1	hadoop	supergroup	418721	2021-02-04	15:15	/user/input/dom_casmurro.txt
-rw-r--r--	1	hadoop	supergroup	453564	2021-02-04	15:15	/user/input/esau_jaco.txt
-rw-r--r--	1	hadoop	supergroup	331000	2021-02-04	15:15	/user/input/historias_sem_data.txt
-rw-r--r--	1	hadoop	supergroup	232652	2021-02-04	15:15	/user/input/mao_luva.txt
-rw-r--r--	1	hadoop	supergroup	322977	2021-02-04	15:15	/user/input/memorial_aires.txt
-rw-r--r--	1	hadoop	supergroup	403669	2021-02-04	15:15	/user/input/memorias_postumas.txt
-rw-r--r--	1	hadoop	supergroup	355240	2021-02-04	15:15	/user/input/papeis_avulsos.txt
-rw-r--r--	1	hadoop	supergroup	289770	2021-02-04	15:15	/user/input/poesias_completas.txt
-rw-r--r--	1	hadoop	supergroup	493323	2021-02-04	15:15	/user/input/quincas_borba.txt

```
hadoop@tumenas-Lenovo-G40-80:~/hadoop-3.3.0$ 
```

k to add speaker notes

```
hadoop@tumenas-Lenovo-G40-80:~/hadoop-3.3.0/bin$ hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.0.jar -input /user/input/*.txt -output /user/output/results -mapper $HADOOP_HOME/mapper.py -reducer $HADOOP_HOME/reducer.py
packageJobJar: [/tmp/hadoop-unjar3309211693858719328/] [] /tmp/streamjob7986355403806411076.jar tmpDir=null
2021-02-04 17:00:43,018 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8032
2021-02-04 17:00:43,603 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8032
2021-02-04 17:00:44,567 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1612468766211_0001
2021-02-04 17:00:45,771 INFO mapred.FileInputFormat: Total input files to process : 9
2021-02-04 17:00:46,649 INFO mapreduce.JobSubmitter: number of splits:9
2021-02-04 17:00:48,299 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1612468766211_0001
2021-02-04 17:00:48,299 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-02-04 17:00:49,321 INFO conf.Configuration: resource-types.xml not found
2021-02-04 17:00:49,321 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-02-04 17:00:49,990 INFO impl.YarnClientImpl: Submitted application application_1612468766211_0001
2021-02-04 17:00:50,046 INFO mapreduce.Job: The url to track the job: http://tumenas-Lenovo-G40-80:8088/proxy/application_1612468766211_0001/
2021-02-04 17:00:50,048 INFO mapreduce.Job: Running job: job_1612468766211_0001
2021-02-04 17:01:03,360 INFO mapreduce.Job: Job job_1612468766211_0001 running in uber mode : false
2021-02-04 17:01:03,361 INFO mapreduce.Job: map 0% reduce 0%
2021-02-04 17:01:15,585 INFO mapreduce.Job: map 11% reduce 0%
2021-02-04 17:01:16,593 INFO mapreduce.Job: map 22% reduce 0%
2021-02-04 17:01:17,610 INFO mapreduce.Job: map 67% reduce 0%
2021-02-04 17:01:24,674 INFO mapreduce.Job: map 89% reduce 0%
2021-02-04 17:01:25,679 INFO mapreduce.Job: map 100% reduce 0%
2021-02-04 17:01:28,694 INFO mapreduce.Job: map 100% reduce 100%
2021-02-04 17:01:30,710 INFO mapreduce.Job: Job job_1612468766211_0001 completed successfully
2021-02-04 17:01:30,815 INFO mapreduce.Job: Counters: 55
  File System Counters
    FILE: Number of bytes read=5341064
    FILE: Number of bytes written=13342111
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=3301839
    HDFS: Number of bytes written=803102
    HDFS: Number of read operations=32
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
```



Cluster	
About	
Nodes	
Node Labels	
Applications	
NEW	
NEW_SAVING	
SUBMITTED	
ACCEPTED	
RUNNING	
FINISHED	
FAILED	
KILLED	
Scheduler	
Tools	

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Queued
1	0	0	1	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
1	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Default Queue
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vcores:16

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority
application_1612468766211_0001	hadoop	streamjob7986355403806411076.jar	MAPREDUCE		default	0

Showing 1 to 1 of 1 entries



Application application_1612468766211_0001

Logged in as: dr.who

Cluster
About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools

Application Overview	
User:	hadoop
Name:	streamjob7986355403806411076.jar
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Qui fev 04 17:00:49 -0300 2021
Launched:	Qui fev 04 17:00:52 -0300 2021
Finished:	Qui fev 04 17:01:29 -0300 2021
Elapsed:	40sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Application Metrics	
Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>

hadoop@tumenas-Lenovo-G40-80: ~/hadoop-3.3.0/results

```
hadoop@tumenas-Lenovo-G40-80:~/hadoop-3.3.0$ $HADOOP_HOME/bin/hadoop fs -get /user/output/results $HADOOP_HOME
hadoop@tumenas-Lenovo-G40-80:~/hadoop-3.3.0$ cd results
hadoop@tumenas-Lenovo-G40-80:~/hadoop-3.3.0/results$ ls -l
total 788
-rw-r--r-- 1 hadoop hadoop 803102 fev 4 17:14 part-00000
-rw-r--r-- 1 hadoop hadoop      0 fev 4 17:14 _SUCCESS
hadoop@tumenas-Lenovo-G40-80:~/hadoop-3.3.0/results$ cat part-0000
```

```
é*      2
é,     80
é.    12
é...   2
é...?  1
é..»   2
é;    8
é?    52
échos  2
éco    11
écos   7
écos,  1
énigme 1
épico. 1
épicos, 1
éra    1
éra,   1
éramos 4
és     59
és,   4
és...  1
és?   2
été    1
être   1
ia    2
ihe   1
impia 2
impossíveis? 1
intima 1
in    1
ó     32
óbito 1
óccuper 1
últimos 4
única 1
---- 1
hadoop@tumenas-Lenovo-G40-80:~/hadoop-3.3.0/results$
```

APLICAÇÕES

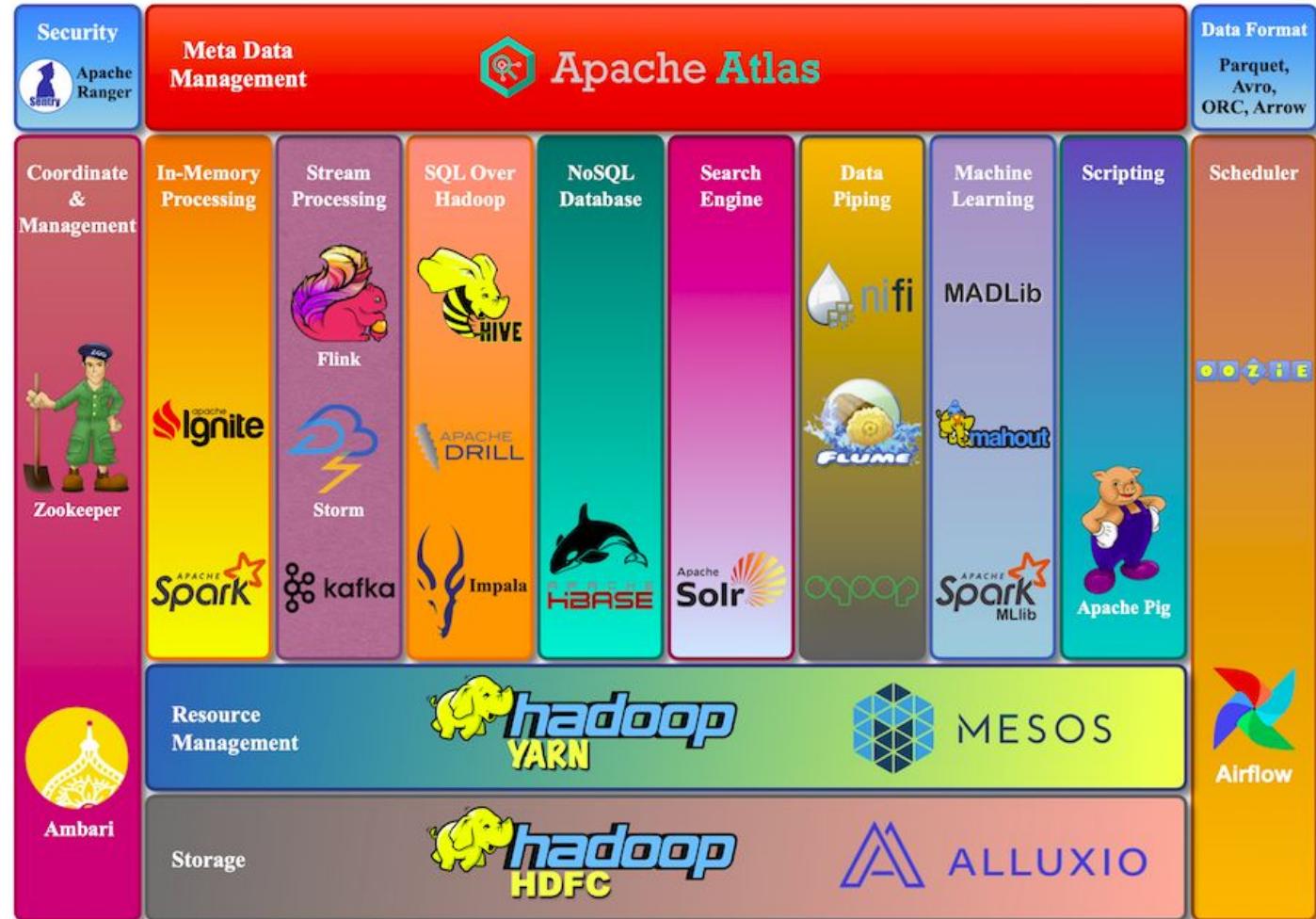
2008 - **PIG** (Yahoo), **Hive** (Facebook) : Linguagens “como SQL”

2009 - **Mahout**: Algoritmos de ML e Estatística

2010 - **Storm** (Twitter): Processamento Real Time

....







comic.browserling.com

DIFICULDADES x FACILIDADES



CLOUDERA



PRÓXIMA AULA

-SPARK, SPARKLYR

-PARQUET

-MONGODB