Universidade Presbiteriana Mackenzie

Processamento em Big Data

O processing em uma visão prática

Flávio Clésio, Me.

Faculdade de Computação e Informática

Ementa Principal

- Escalabilidade
- Ingestão de Dados (Big Data)
- Integração de Dados Estruturados (Extração, Enriquecimento, Transformação e Carga)
- Construção de Aplicações ETL;
- Data Compression;
- Garantia da Qualidade de Dados;
- Gerenciamento de Dados Mestres;
- Ferramentas Disponíveis

Bibliografia Básica

- KRISHNAN, Krish. *Data Warehousing in the Age of Big Data*. EUA:MK,2013.

- SAWANT, Nitin; SAHAH, Himanshu. *Big Data Application Architecture Q&A: A Problem - Solution Approach*. EUA: Apress, 2013.

- MAYER-SCHONBERGER, Viktor;CUKIER, Kenneth. *BIG DATA: Como extrair volume,variedade,velocidade e valor da avalanche de informação cotidiana*. Rio de Janeiro: Campus,2013.

Bibliografia Complementar

- AMAZON, WebService. *Getting Started Guide: Analyzing Big Data with AWS*. EUA: AWS,2013.

- PROVOST,Foster; FAWCETT, Tom. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. EUA: O'Reilly Media,2013.

- O'REILLY,Media. *Big Data Now: 2012 Edition*. EUA: O'Reilly Media,2012.

Quem é você?

- Nome
- Formação
- Empresa que trabalha
- Objetivos com a Pós-Graduação
- Stack de tecnologia
- Qual é a expectativa com a disciplina?

Sobre mim

- Core Team Machine Learning na Movile
- Me. em Engenharia de Produção (Aplicação de Inteligência Computacional em derivativos de crédito) pela UNINOVE
- Esp. em Engenharia de Banco de Dados pela UNICAMP
- Bel. em Sistemas de Informação pelas Faculdades Integradas Rio Branco (FIRB)
- Especialista em Machine Learning, Engenharia de Banco de Dados e Business Intelligence
- Blogger no Mineração de Dados - http://mineracaodedados.wordpress.com
- @flavioclesio (Twitter)

PRESENTED BY O'REILLY AND CLOUDERA SAN JOSE • LONDON • BEIJING • NEW YORK • SINGAPORE

Strata+Hadoop WORLD

DECEMBER 5-6, 2016: TRAINING
DECEMBER 6-8, 2016: TUTORIALS & CONFERENCE
SINGAPORE

SCHEDULE SPEAKERS SPONSOR PAVILION EVENTS VENUE ABOUT RESOURCES ACCOUNT

Flavio Clesio

Machine Learning Engineer, Movile

Website | @flavioclesio

Flavio Clesio is specialist in machine learning and revenue assurance at Movile, where he helps build core intelligent applications to exploit revenue opportunities and automation in decision making. Prior to Movile, Flavio was a business intelligence consultant in financial markets, specifically in nonperforming loans. He holds a master's degree in computational intelligence applied in financial markets.

Sessions

2:35pm–3:15pm Thursday, December 8, 2016
Machine learning in practice with Spark MLlib: An intelligent data analyzer
Chat, machine learning, and AI
Location: Summit 1
Level: Intermediate
Tags: telecom

Flavio Clesio (Movile), Eiti Kimura (Movile)
Average rating: ★★★★☆ (3.50, 4 ratings)

Strata Hadoop World Singapura 2016

Link:

https://goo.gl/IeTCOc

Plano de Aula (WIP)

- Encontro 1: Preparação de Ambiente / Spark / Scala Programming
- Encontro 2: Spark / ~~Amazon Web Services~~
- Encontro 3: H2O.ai / ~~Azure Marketplace~~
- Encontro 4: ETL Modeling / Feature Engineering
- Encontro 5: ETL Modeling / Feature Engineering
- Encontro 6: TBD ~~Amazon Redshift / Google Big Query~~
- Encontro 7: Arquitetura de Processamento de Dados em Big Data / Stack de Tecnologia
- Encontro 8: TBD

Avaliações

- 3 Projetos (60%) – $P_{rojeto}1$: 20% / $P_{rojeto}2$: 20% / $P_{rojeto}3$: 20%
- 2 Provas (20%) – $P_{rova}1$: 10% / $P_{rova}2$: 10%
- 2 Relatórios (20%)

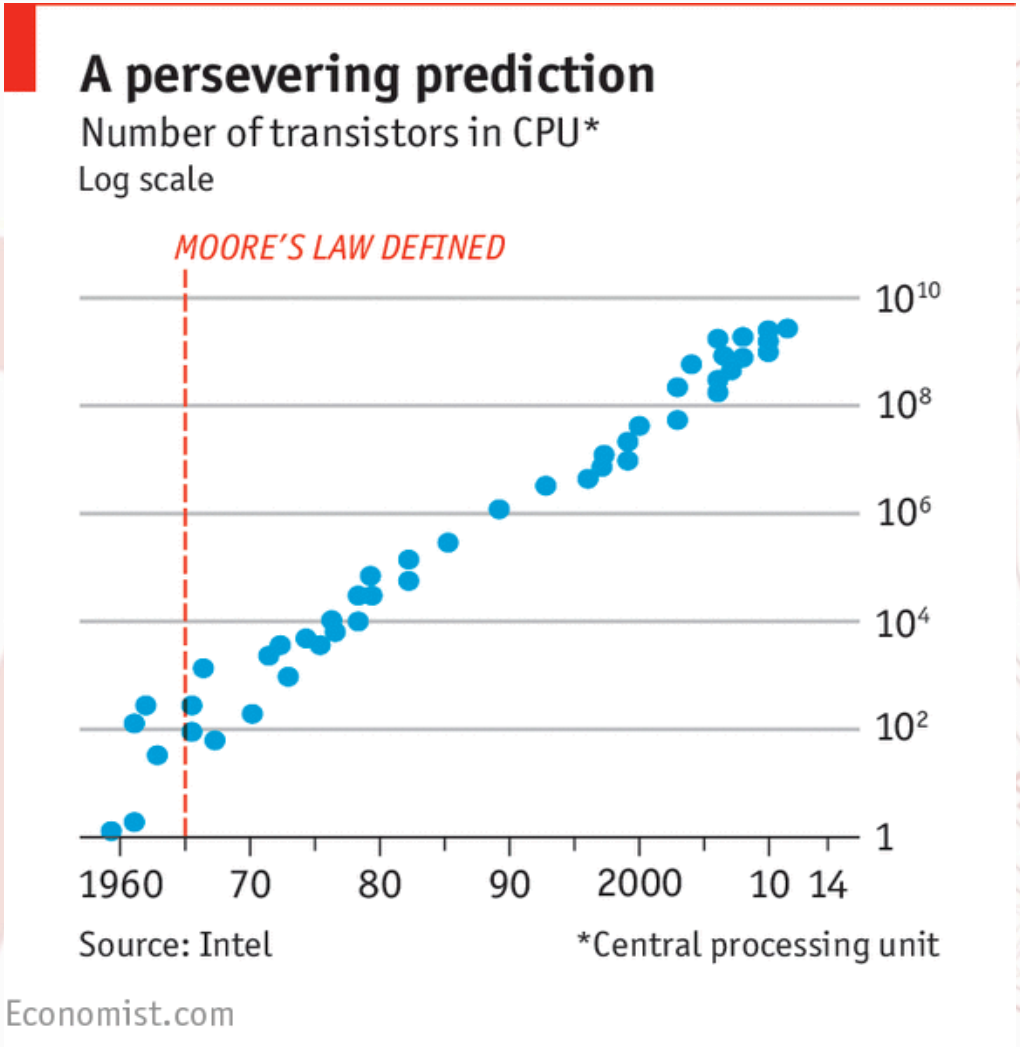
Processamento em Big Data nos dias atuais...

400.000 petabytes de Informação.

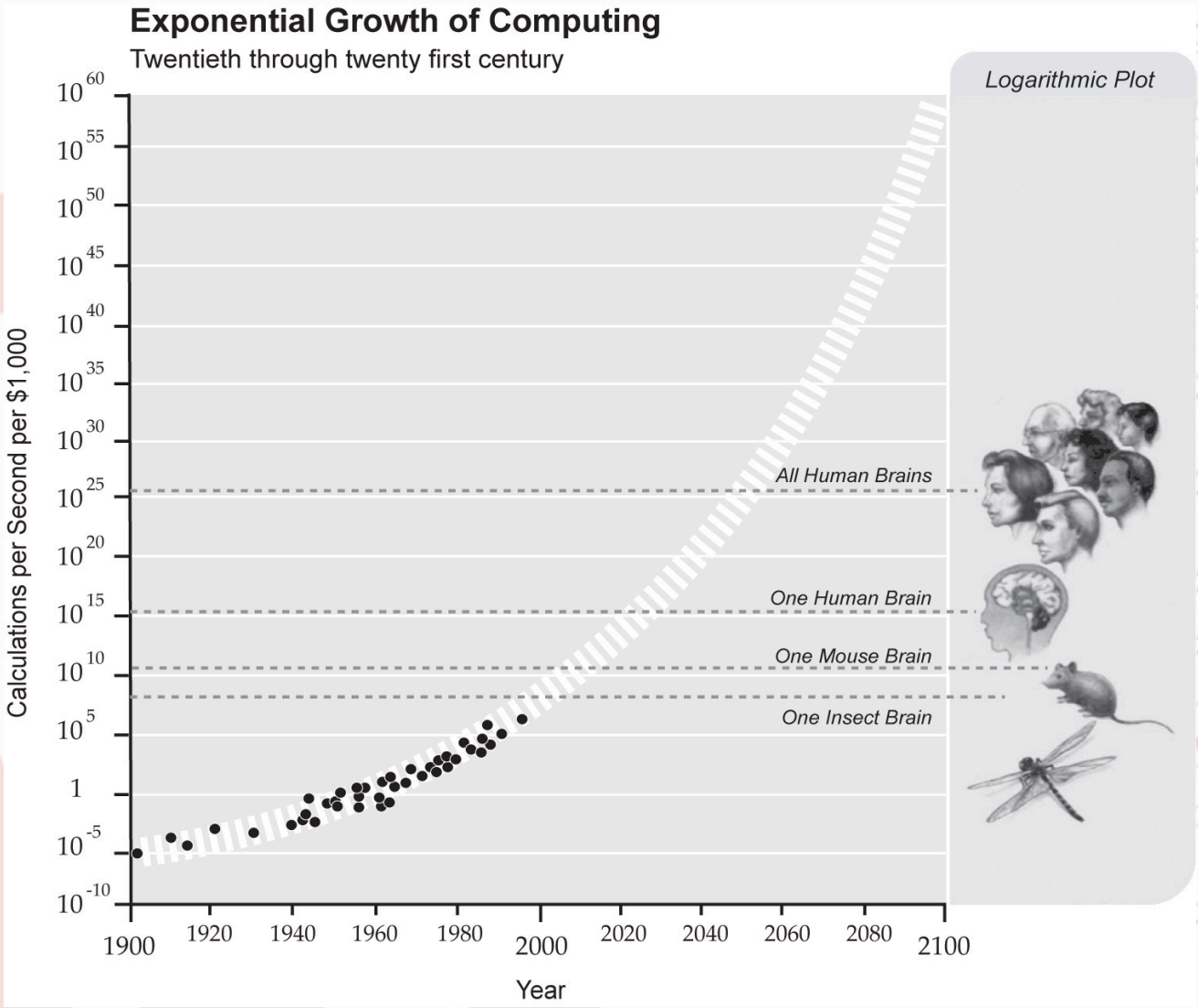
Fonte: Quora

Nos horários de pico chega elevar em 37% o tráfego de internet. Fonte: [Variety]

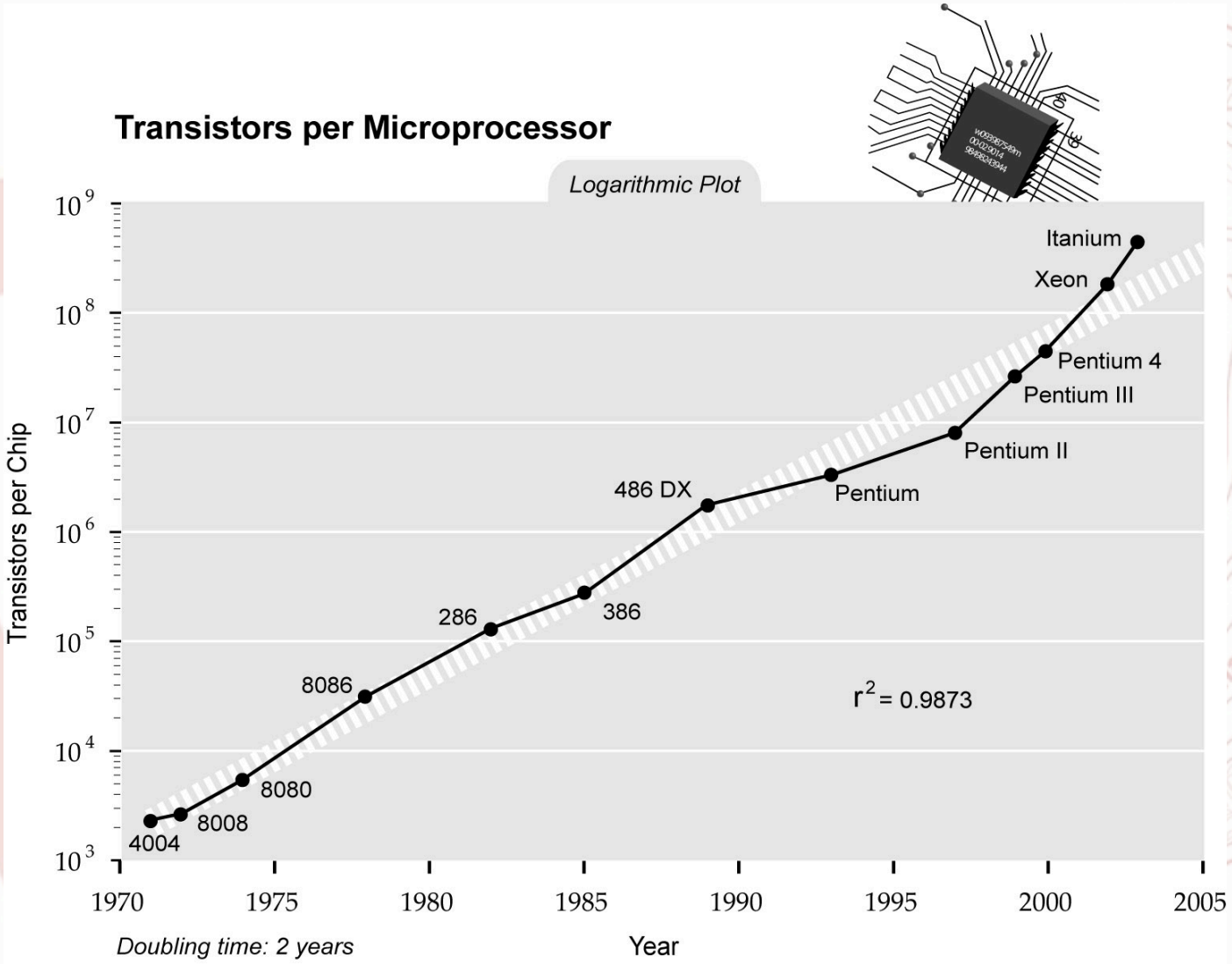

300 petabytes (3 milhões de Gb). Fonte: Facebook Engineering Blog
Facebook



Fonte: [The Economist]



Fonte: Singularity



Fonte: [Singularity]

Situação Atual

- Barateamento da memória e processamento (Lei de Moore)
- Aplicações mais complexas
- Hardware já virou Commodity (AWS, GCS, Azure Market Place)

Necessidades diante deste cenário

- Trafego de rede de todo esse volume de dados
- Computação distribuída
- Maximização de aproveitamento de hardware
- Paralelismo extremo de I/O (IOPS)
- Orquestramento de processos

Processamento de Big Data

- Haddop
- Spark
- SGBDs

Extract, Transform, Load

- E: CSV, SGBD, JSON, WebServices, Protocol Buffers
- T: Surrogating, Binning, Agregações
- L: Carga no DW, Preparação para ferramenta de Analytics, Preparação para ferramenta de Machine Learning

Caracterização de um problema de processing

- Multiplas leituras ao longo do tempo
- Alto volume de dados
- Heterogeneidade das fontes de dados (Estruturada x Não-Estruturada)

Data Processing nos dias atuais...

SECTIONS HOME SEARCH The New York Times SUBSCRIBE NOW LOG IN

TECHNOLOGY

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.

Fonte: NYT

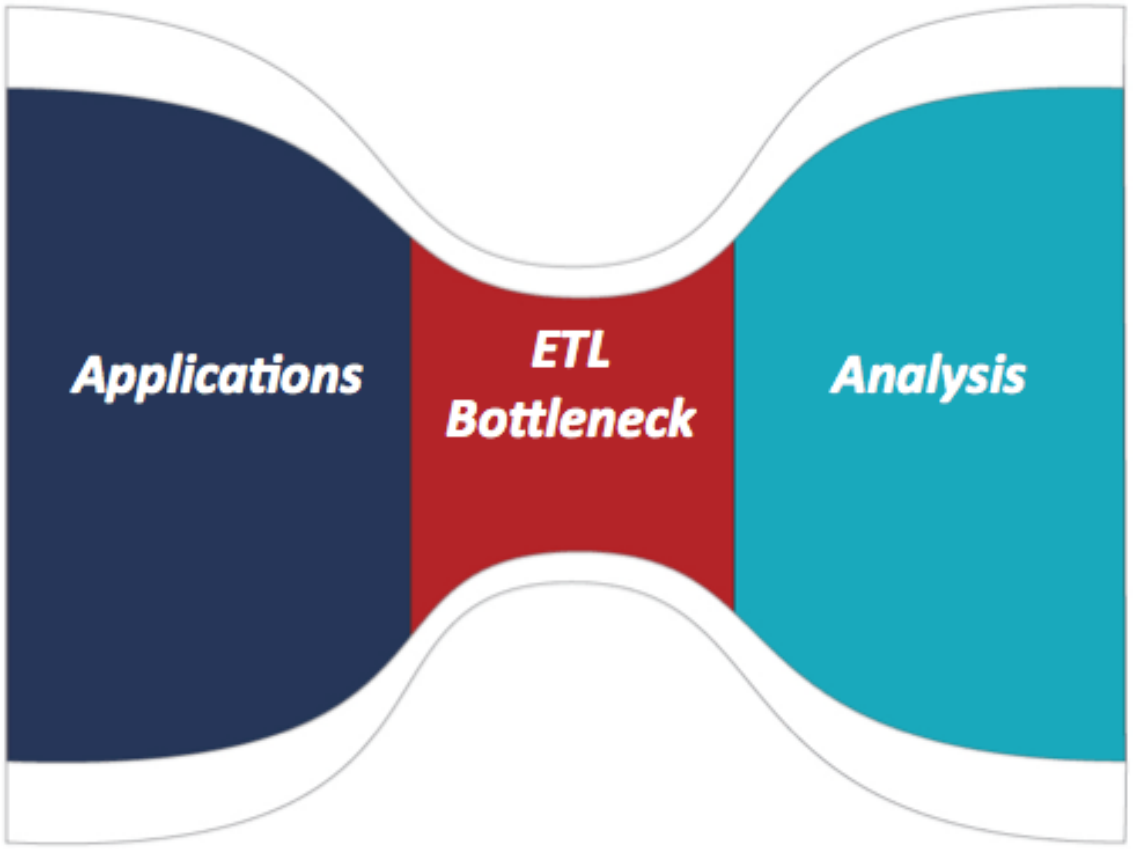
Data Processing nos dias atuais...

- *Yet far too much handcrafted work — what data scientists call "data wrangling," "data munging" and "data janitor work" — is still required. Data scientists, according to interviews and expert estimates,* ***spend from 50 percent to 80 percent of their time mired in this more mundane labor*** *of collecting and preparing unruly digital data, before it can be explored for useful nuggets.*

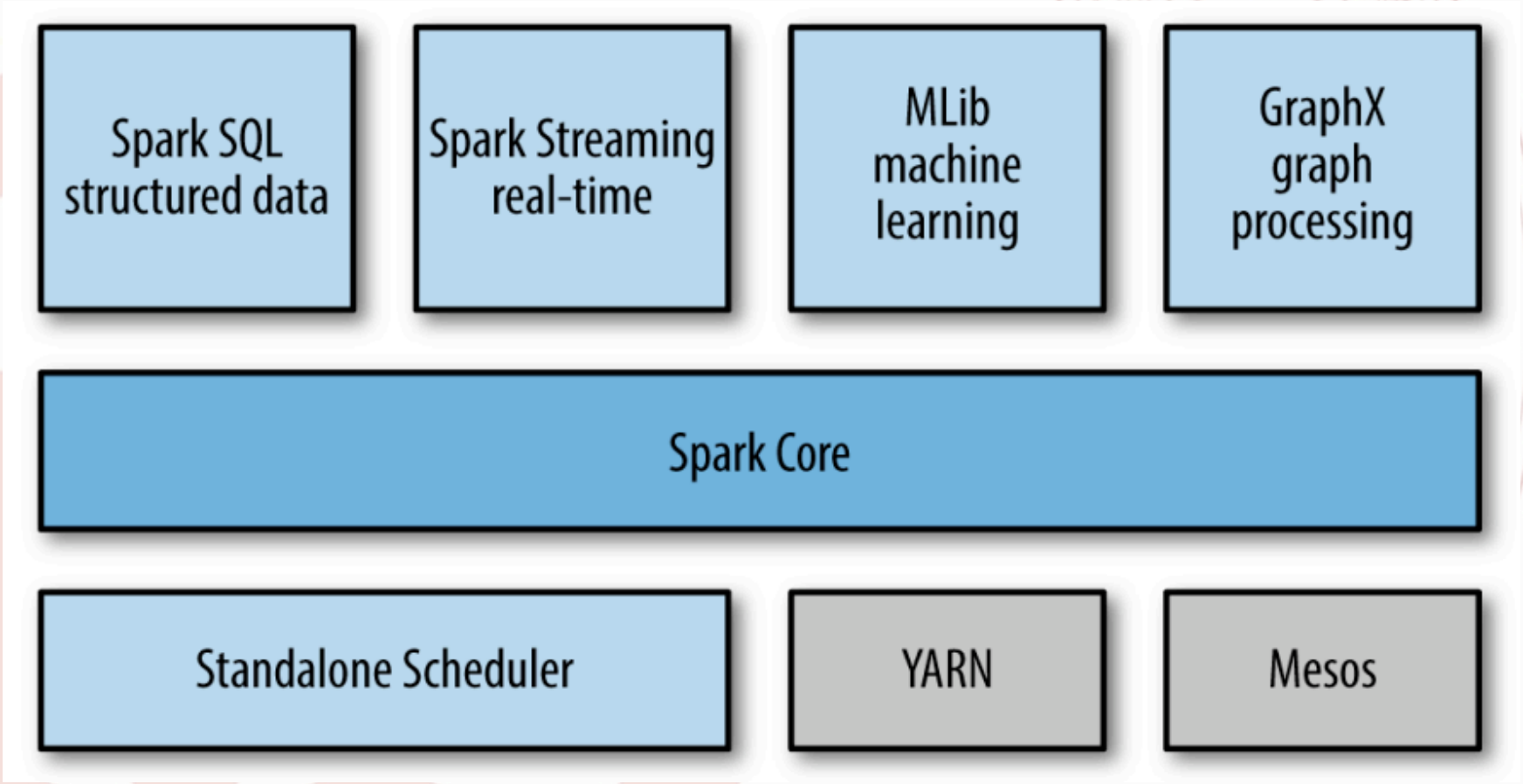
Fonte: NYT

ETL: O grande gargalo

30-40%
data growth
per year
Source: 2013 IBM Briefing Book



Apache Spark



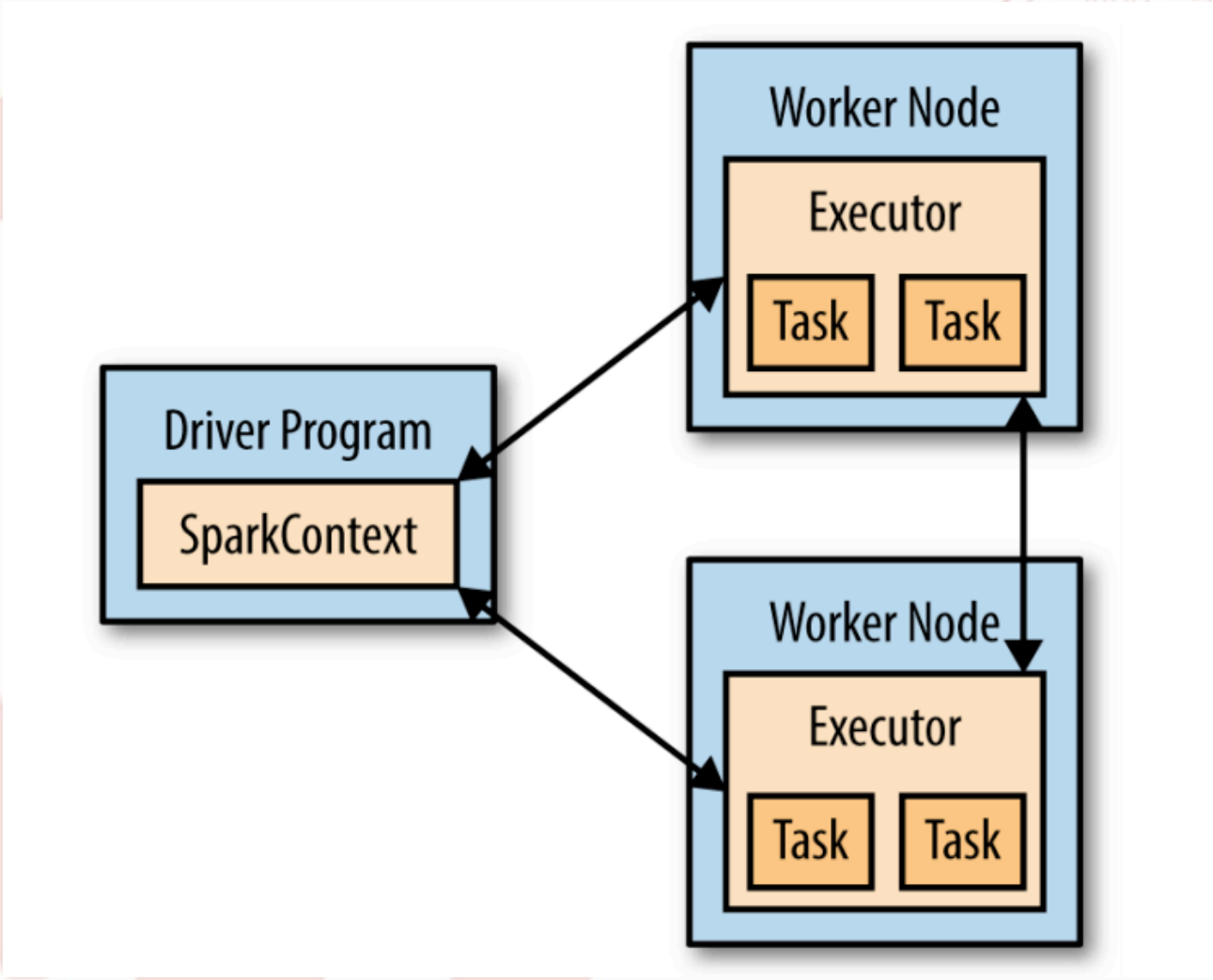
Características do Spark

- Tem integração com diversas linguagens (*e.g*. Python, Java, Scala e R)
- Tem o Scala como linguagem funcional
- Acesso nativo a múltiplos sistemas de arquivos (*e.g*. HDFS, Cassandra, Hbase, S3, CSV)
- Caching para performance (*i.e*. reduz acesso ao disco)

Características do Spark

- Projeto e comunidade muito ativos (JIRA)
- Até 40x mais rápido do que Haddop MapReduce
- Integração com EC2 (Elastic Cloud Computing), EMR (Elastic MapReduce), HDFS, S3, YARN, e Mesos

Spark: Internals

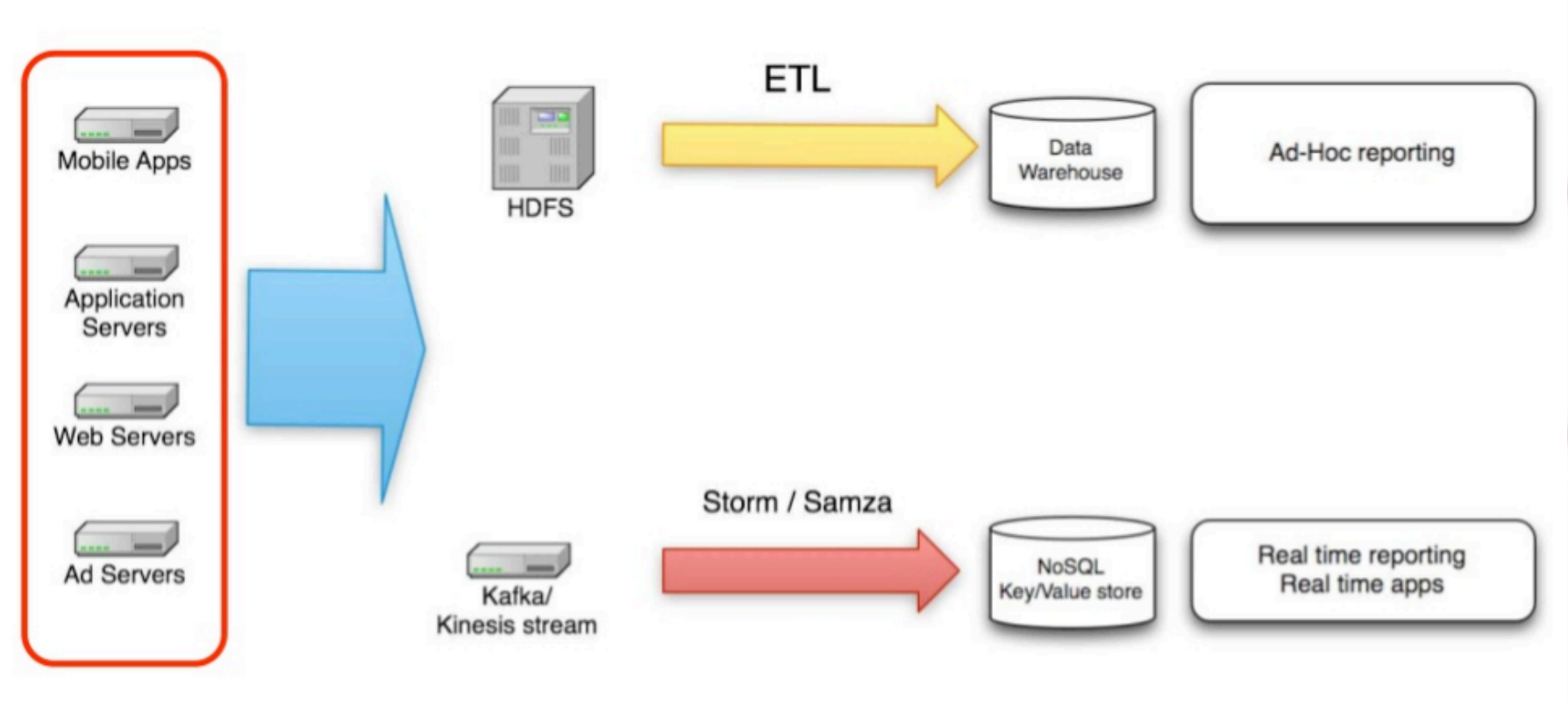


Map Reduce x Spark

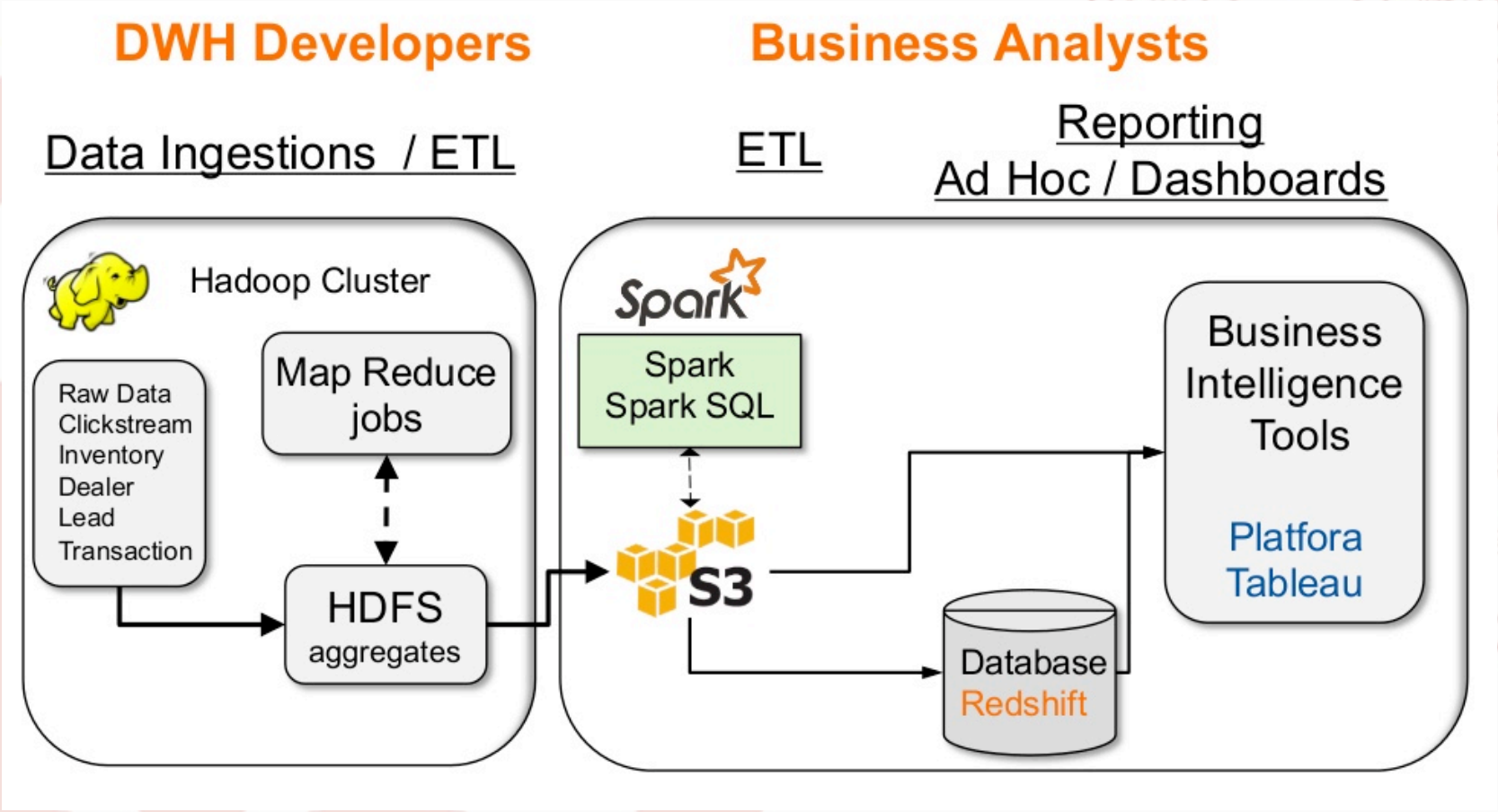
- MapReduce

- Spark

Arquitetura básica de Processing



Arquitetura básica de Processing



Leituras Complementares

- ETL with Spark
- Benchmark Hadoop x Spark
- Machine learning in practice with Spark MLlib: An intelligent data analyzer
- Data Integration is not ETL
- For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights

Obrigado

Flávio Clésio

flavioclesio@gmail.com

Visitem

http://mineracaodedados.wordpress.com