If you do not look at things on a large scale,
it will be difficult to master strategy.
—Miyamoto Musashi

Big Data e Computação em Nuvem

Aula 01

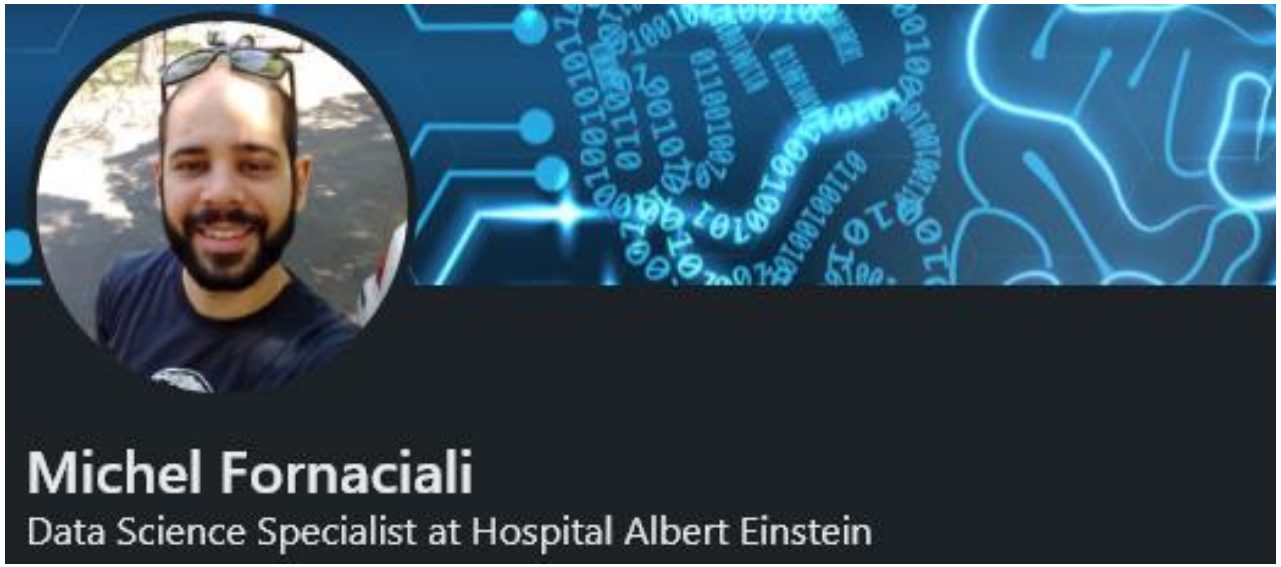
Introdução à Computação em Larga Escala

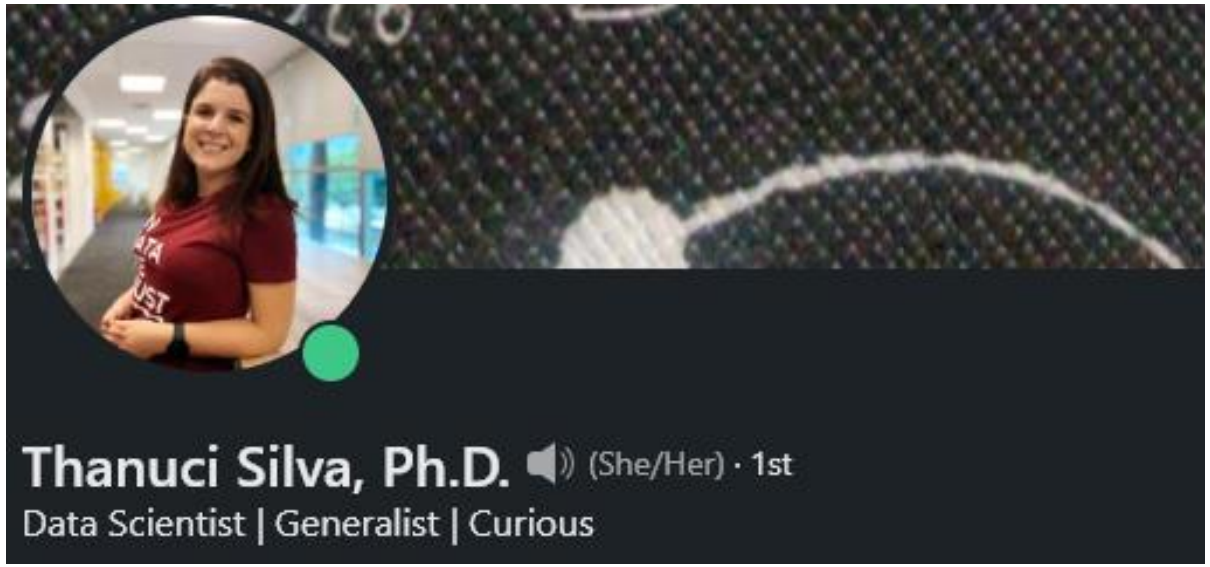
Prof. Michel Fornaciali, PhD.
Prof.ª Thanuci Silva, PhD.

Contatos:
MichelSF@insper.edu.br
thanucis@insper.edu.br

Insper

Apresentação





Insper

Nosso objetivo é conseguir tratar grandes quantidades de dados.

Objetivos de aprendizagem

Objetivos específicos:

- Desenhar arquiteturas para análise de dados em grande escala baseadas em serviços na nuvem;

- Desenvolver algoritmos para a análise de dados em grande escala utilizando Python, Dask, Spark e arquiteturas em nuvem.

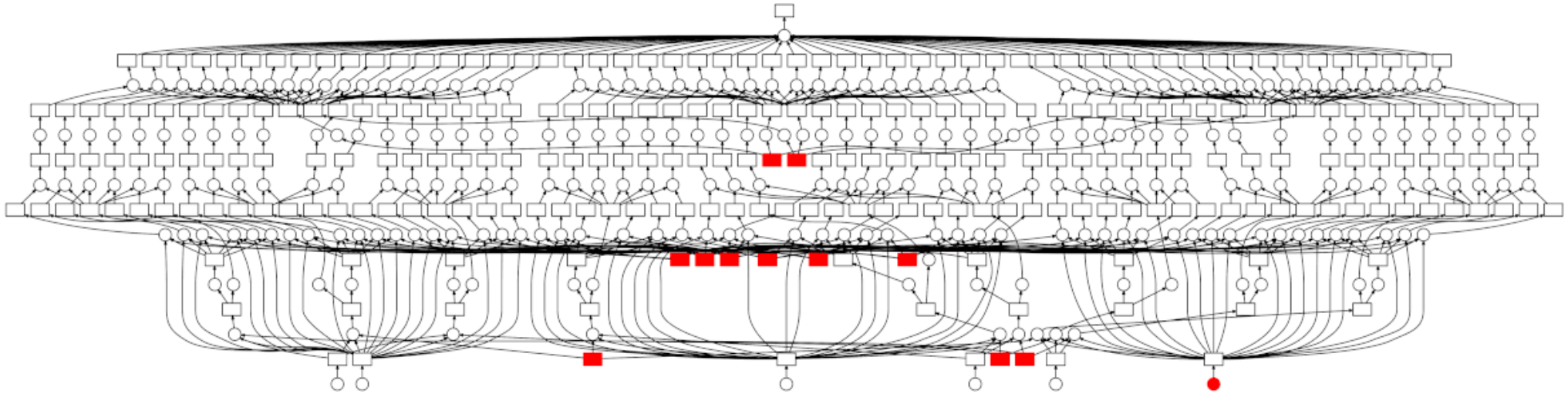


Imagem: dask.org

Nossa disciplina utilizará Python 3. Além disso, utilizaremos alguns conhecimentos básicos de Linux, RegEx e programação funcional.

Requisitos

Pré-requisitos:

- Python (pandas, numpy, etc.);
- Linux básico;
- Programação funcional básica;

Bibliotecas:

- Python & Anaconda
- Dask: https://dask.org
- PySpark: https://spark.apache.org/ (versão 3)
- AWS SDK for Python (boto3): https://aws.amazon.com/sdk-for-python/
- AWS Command Line Interface (AWSCLI): https://aws.amazon.com/cli/

Conjugaremos uma dinâmica de aula-laboratório em todos os encontros

Dinâmica

Aulas:

- Construção conjunta do conhecimento;
- Solução de problemas reais em conjunto;
- Exercícios e discussões;

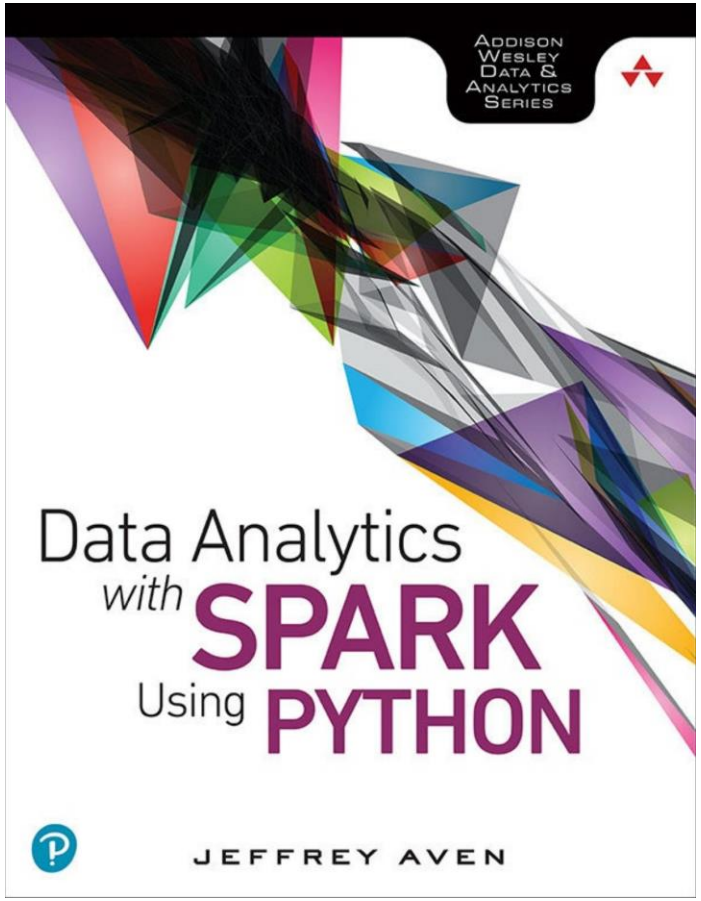
Laboratórios a cada aula:

- Dúvidas;
- Exercícios de fixação;
- Exercícios de extensão;
- Acompanhamento do projeto final;

Focaremos na nuvem Amazon Web Services (AWS) com a utilização do Apache Spark em Python.

Conteúdo Programático

- Big Data e computação em nuvem;
- Arquitetura, elementos e gestão de serviços em nuvem;
- Big Data e fundamentos de processamento e distribuído;
- Fundamentos do Apache Spark;
- Estruturas de dados do Apache Spark;
- Programação PySpark para Apache Spark e Spark SQL.



Bibliografia

Insper

Todas as nuvens possuem como base sistema operacional Linux. Caso você queira instalar, utilize as seguintes recomendações.

Material Complementar

- Instalação local, completa, no seu computador:
  - Tutorial de como instalar Ubuntu sobre VirtualBox:
  - Win: https://www.youtube.com/watch?v=zsqJhle7CXE&t=608s
  - Mac: https://www.youtube.com/watch?v=aJcc-xC6krE

- O Windows 10 tem um subsistema Linux (WSL).
  Permite você abrir um terminal simplificado, sem interface gráfica, suficiente para testar comandos:
  - https://www.youtube.com/watch?v=Eb6rEpXSTpY

- Cheat-sheet com comandos mais importantes Linux:
  - https://www.linuxtrainingacademy.com/linux-commands-cheat-sheet/

Nossa bibliografia básica é o livro do Aven. Utilizaremos ainda SQL e AWS.

Bibliografia

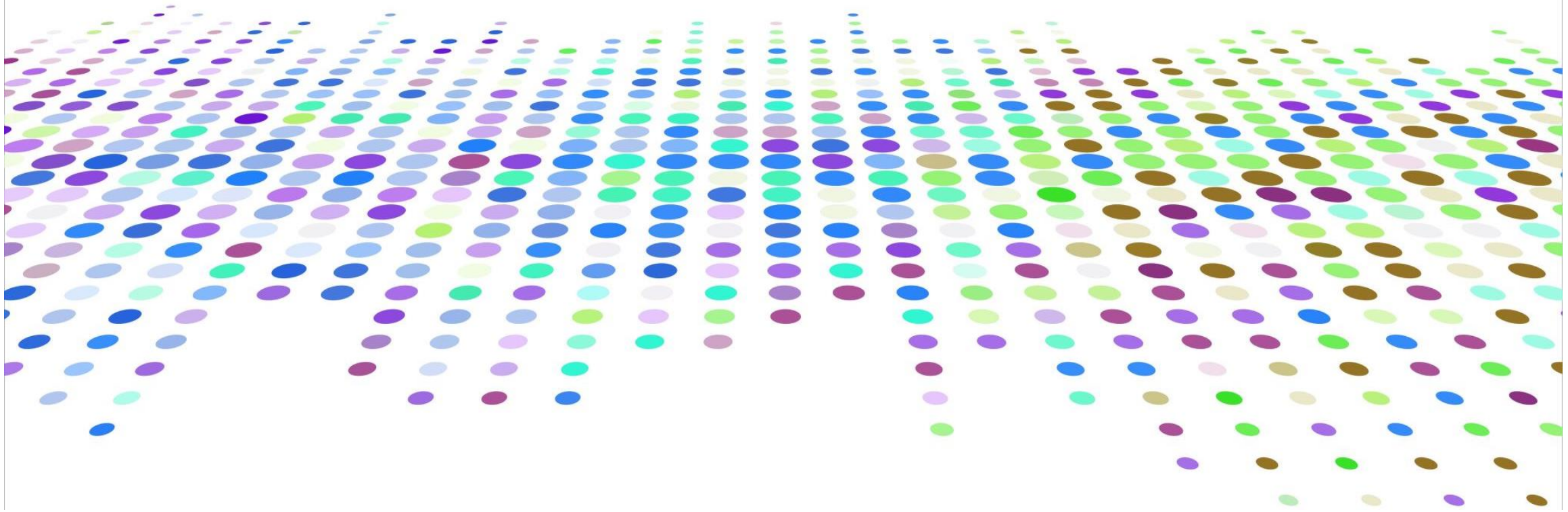
Bibliografia Básica:

- **Data Analytics with SPARK using Python**. Aven, J. 2018. Addison-Wesley.

Bibliografia Complementar:

- Amazon Web Services in Action. Wittig, M. Wittig, A. 2nd. Ed. 2019. Manning.
- Cloud Computing: Theory and Practice. Marinescu, D. 2nd. Ed. 2018. Morgan Kaufmann.
- Learning Spark. Lightning Fast Data Analysis. Karau, H. et. Al. 2015. O'Reilly
- Spark, The Definitive Guide. Big Data Processing Made Simple. Bill Chambers & Matei Zaharia. 2018. O'Reilly.

Computação em Larga Escala



Contexto: aplicações atuais de Big Data Analytics

Olhar Digital

Cientistas estudam uso de inteligência artificial para detectar câncer

... por meio de machine learning, a classificar os tecidos entre sau doente. Durante a endoscopia, o sistema analisa a imagem da ...

há 2 semanas

TechCrunch

Aidoc raises over $66M for AI radiology analysis technology

Aidoc develops "decision support" software based on artificial intelligence. This

EVENTOS

02/10/2019

Fórum CIO Club

LEROY MERLIN APOSTA EM BIG DATA PARA UMA NOVA JORNADA DO CONSUMIDOR

O foco é capturar o cliente 4.0 que é hiperconectado e está cada dia mais exigente

14/09/2019 - 07H05 - POR ÉPOCA NEGÓCIOS ONLINE

Nestlé usa inteligência artificial para monitorar a felicidade das vacas

A iniciativa faz parte do Projeto Cowsense, uma ação desenvolvida para o acompanhamento das fazendas produtoras de leite orgânico

GlobeNewswire

Artificial Intelligence in Diagnostic Market wort

Growth in this market is primarily driven by government initi the adoption of AI-based technologies, rising demand for AI

há 1 dia

Inteligência Artificial revoluciona o varejo

A tecnologia da Microsoft contribui para que todos os elos da cadeia – da manufatura ao ponto de venda incluindo o cliente final – saiam ganhando mais

DINO

Como as empresas estão utilizando o Big Data para melhorar os resultados

ntender o comportamento e as p

08/10/2019 - 06H18 - ATUALIZADA ÀS 08H37 - POR ESTADÃO CONTEÚDO

Inteligência artificial é testada para aliviar demanda da Justiça

STJ está trabalhando em dois projetos ligados à inteligência artificial

Analytics Insight

Impact of Data Science in Healthcare

Data Science helps in advancing healthcare facilities and processes. It helps boost productivity in diagnosis and treatment and enhances the ...

há 2 dias

Insper

O que caracteriza o Big Data?

Os 5 V's do Big Data

- Volume
- Variedade
- Velocidade
- Veracidade
- Valor

Imagem reproduzida de: https://bit.ly/3meRz5z

Insper

Reflexões

- Quais são as dificuldades de se trabalhar com computação em larga escala?
  - Infra
  - Tempo
  - Dev / debug
  - Orquestração de operações

- Como poderemos contorná-las?
  - Particionar os dados
  - Escalabilidade de infra (elasticidade)
  - Tolerância à falha

Escalabilidade

- É a habilidade de um sistema lidar com o aumento da carga de processamento sem apresentar uma degradação significante em seu desempenho

- Há duas formas de se obter a escalabilidade:
  - **Escalabilidade vertical – *scale-up***
    - Fazer upgrade na infra existente (+ memória, por exemplo)
  - **Escalabilidade horizontal – *scale-out***
    - Adicionar novas máquinas ao parque computacional
    - Distribuir os dados e o trabalho de processamento em diversas máquinas

Desafios de processar Big Data

- É fato que a capacidade dos discos aumentou muito nos últimos anos:
  - Um disco típico de 1990 poderia armazenar **1.370 MB** de dados
    - Exemplo: Seagate ST-41600n (foto)

  - Um SSD atual pode armazenar **2 TB** de dados
    - Exemplo: Seagate Barracuda 120 (foto)

- Todavia, a velocidade de transferência de dados não seguiu na mesma proporção
  - Em 1990, um disco tinha uma velocidade de transferência de **4,4 MB/s**
  - Um SSD atual possui uma velocidade de transferência de **500 MB/s**

Desafios de processar Big Data

- É fato que a capacidade dos discos aumentou muito nos últimos anos:
  - Um disco típico de 1990 poderia armazenar **1.370 MB** de dados
    - Exemplo: Seagate ST-41600n (foto)

  - Um SSD atual pode armazenar **2 TB** de dados
    - Exemplo: Seagate Barracuda 120 (foto)

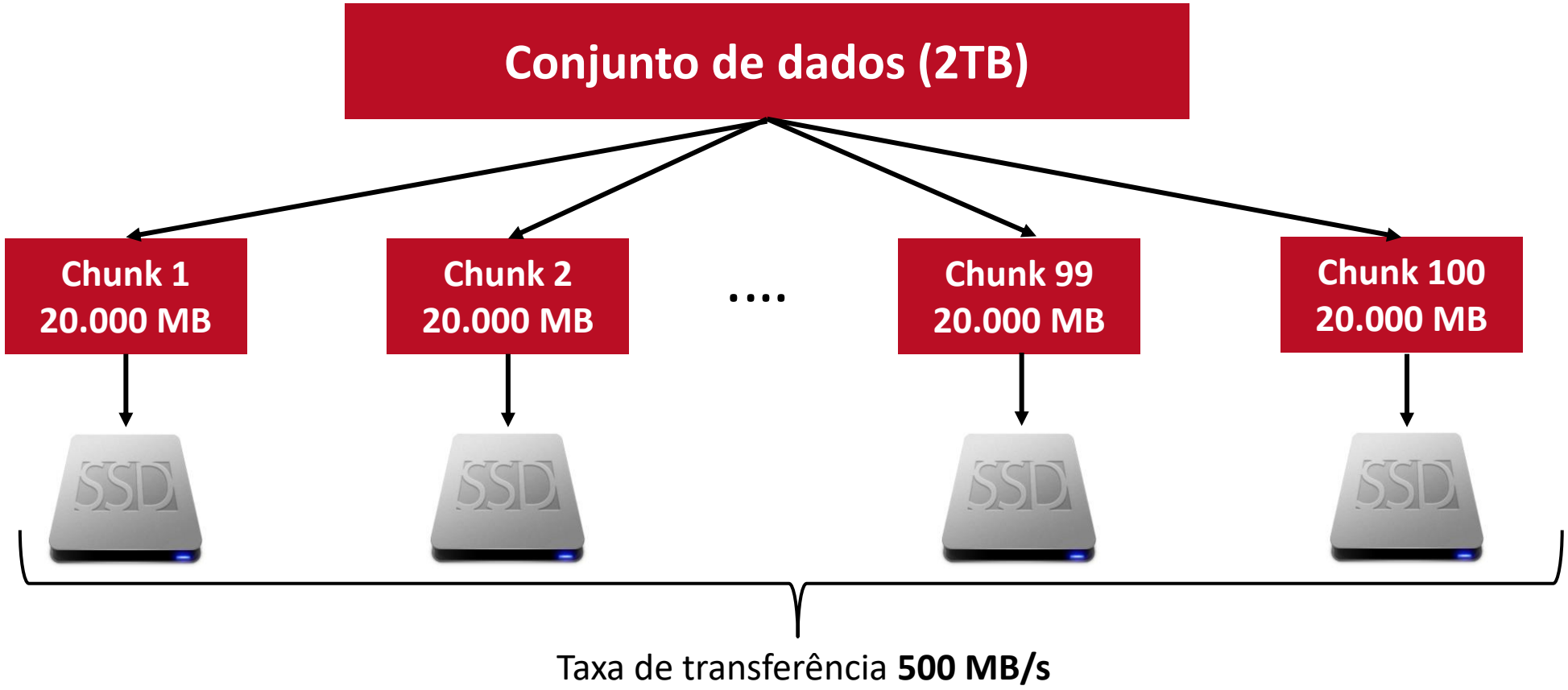
- Todavia, a velocidade de transferência de dados não seguiu na mesma proporção
  - Em 1990, um disco tinha uma velocidade de transferência de **4,4 MB/s**
  - Um SSD atual possui uma velocidade de transferência de **500 MB/s**

- Em 1990, levaríamos **5 minutos** para ler todos os dados do disco.
- Atualmente, levamos **mais do que uma hora** para ler todos os dados do disco!

Como lidar com esse problema?

Paralelizar?

- Dividir os dados em blocos (*chunks*) armazenados em diferentes discos, que são **lidos em paralelo**



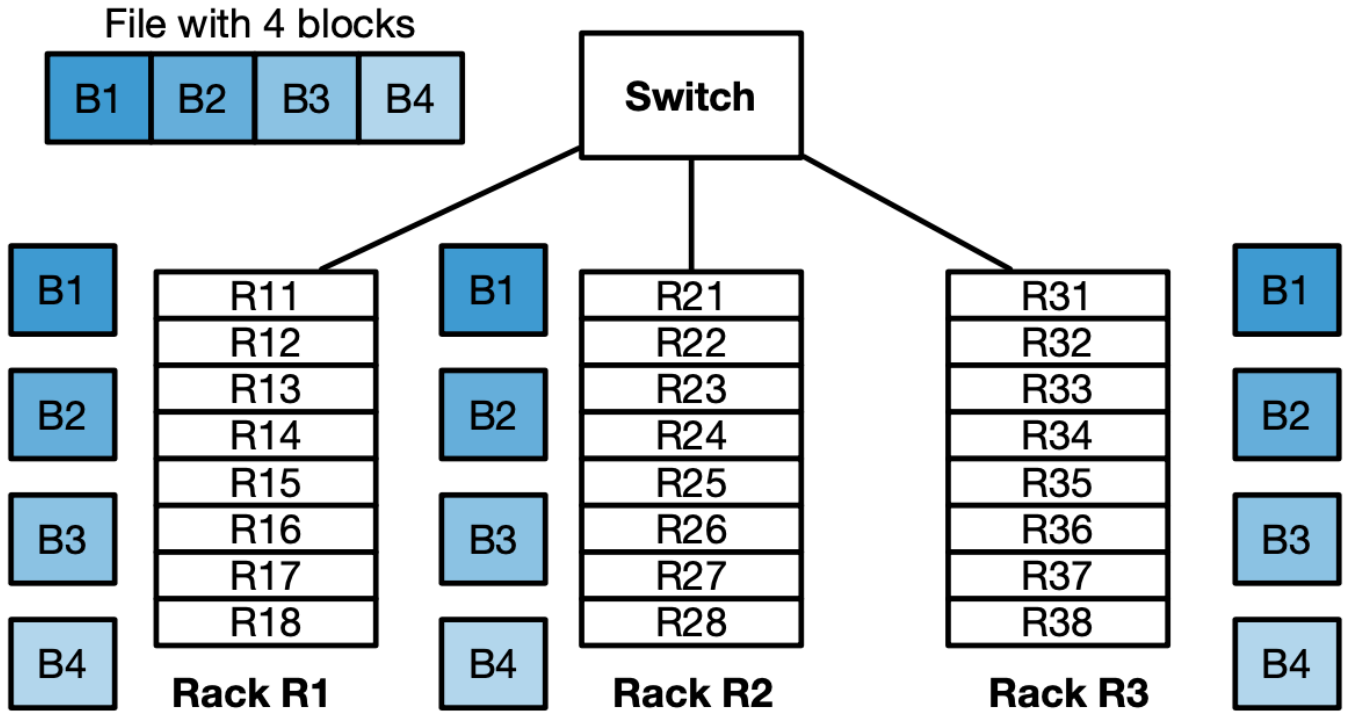
Desafios do Paralelismo

- Falha de hardware
  - Com mais discos, maior é a chance de falhas em hardware
  - Solução?
    - Fazer uso da **redundância**
    - Isto é, replicar os dados em vários discos
      - Esta é a proposta de sistemas de arquivos como o **Hadoop Distributed File System** (HDFS)
    - Processar os dados de uma maneira distribuída
      - Este é um dos papeis do **Apache Spark**

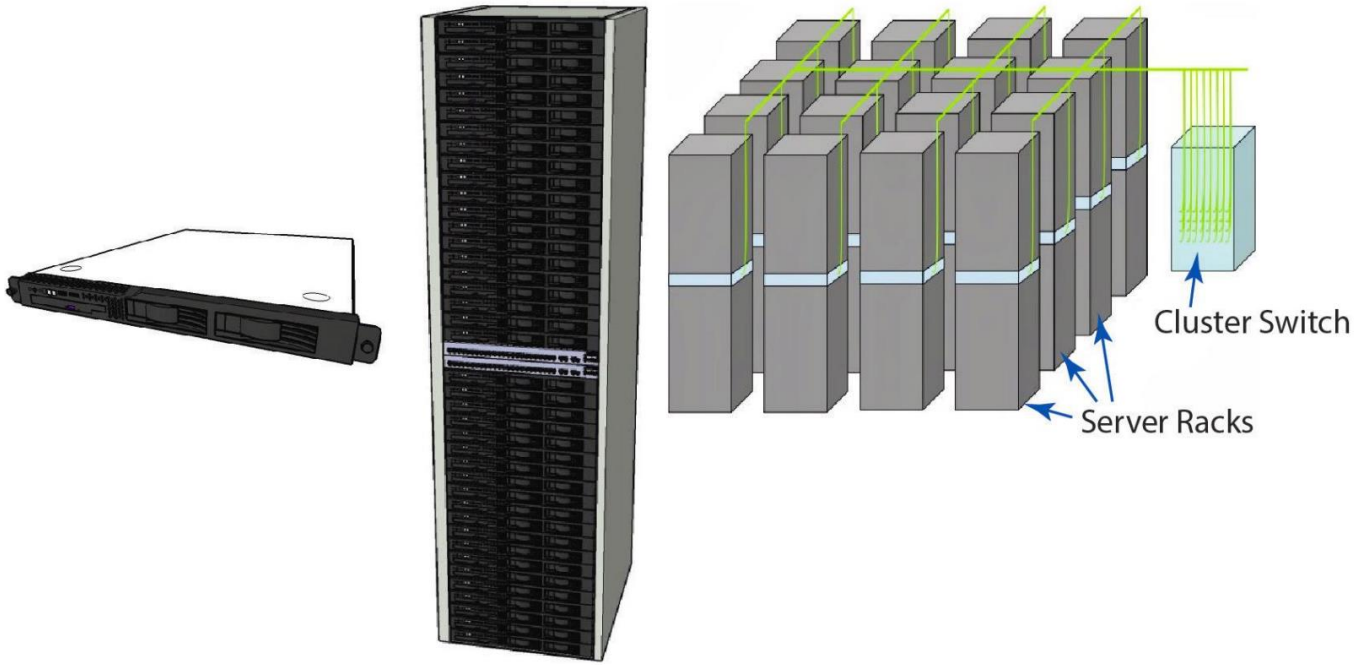
Desafios do Paralelismo

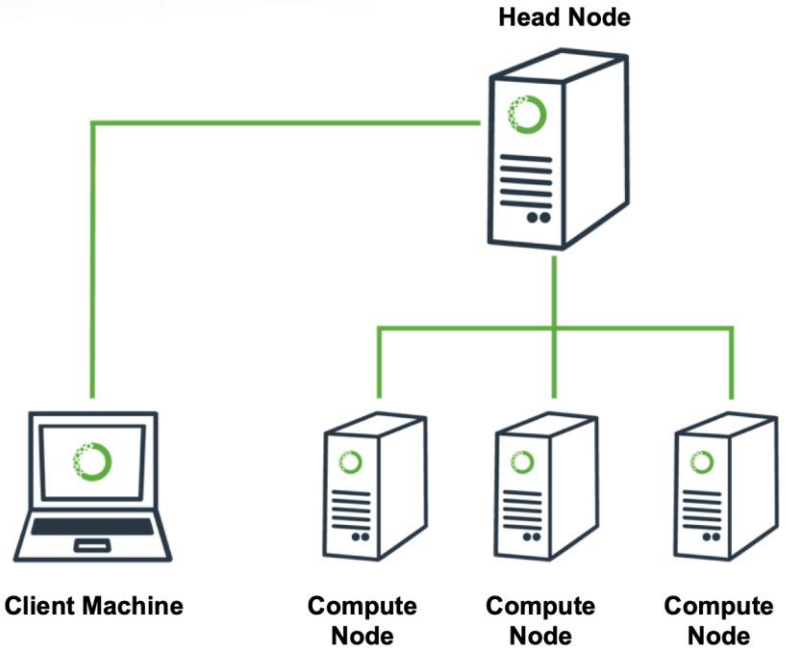
Scale out – escalabilidade horizontal

- Podemos fazer uso de um **cluster** de máquinas
- Os dados são armazenados em um sistema de arquivos distribuído (e.g., HDFS)
- Cada arquivo é divido em blocos de tamanho fixos
- Cada bloco é **replicado** em diversos nós do cluster



Cluster





Barroso, Clidaras, Holzle (2013)

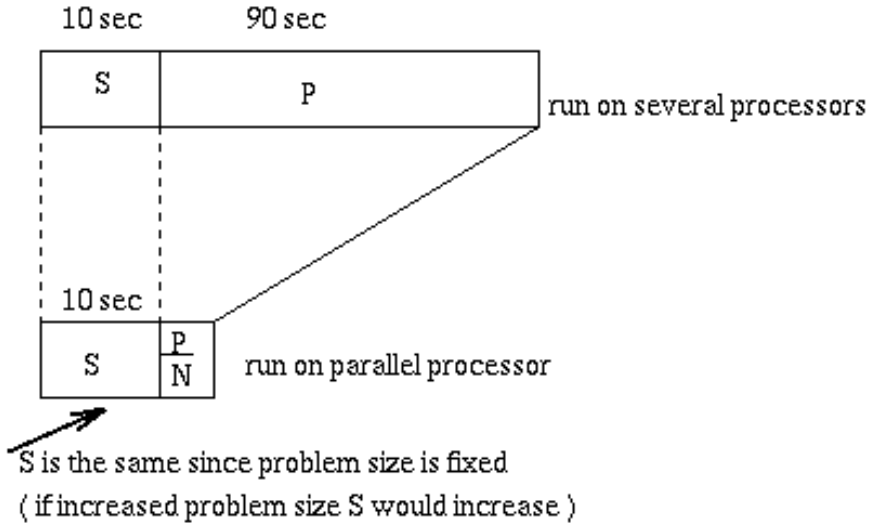
Insper

Paralelismo …

Speedup …. Lei de Amdahl

- Numa aplicação existe sempre uma parte que não pode ser paralelizada
- Seja **S** a parte do trabalho sequêncial, **1-S** é a parte susceptível de ser paralelizada
- Mesmo que a parte paralela seja perfeitamente escalável, o aumento do desempenho (speedup) está limitado pela parte sequêncial

n = número de processadores



$$\textbf{Speedup} = \frac{1}{S + \frac{(1-S)}{n}}$$

Insper

Paralelismo …

Speedup …. Lei de Amdahl

- Se 10% das operações de um código precisam ser feitas sequencialmente, então o speedup não pode ser maior do que 10, independente do número de processadores

$$S = \frac{1}{0.1 + \frac{0.9}{10}} \cong 5.3 \qquad p = 10 \text{ processors}$$

$$S = \frac{1}{0.1 + \frac{0.9}{\infty}} = 10 \qquad p = \infty \text{ processors}$$

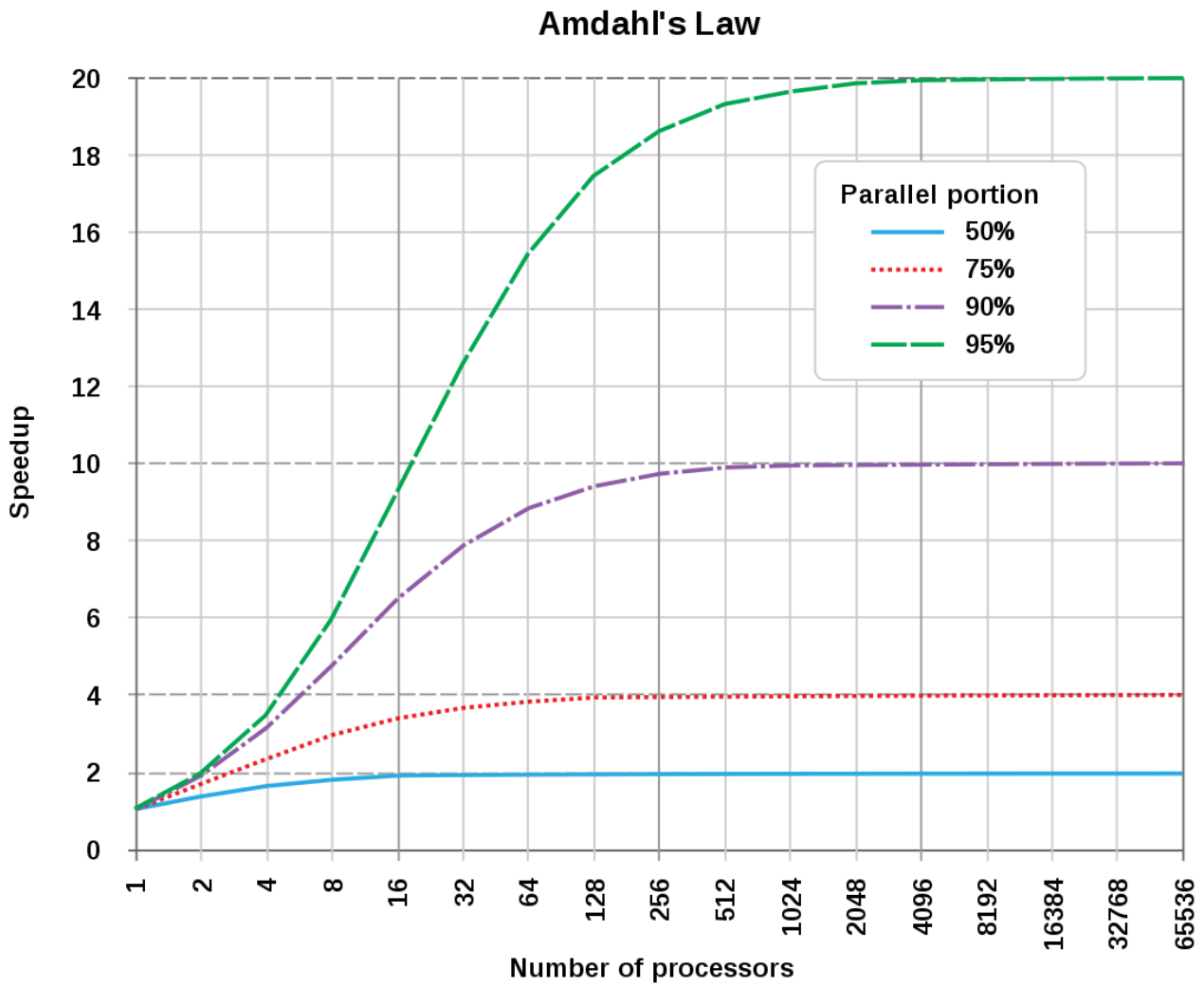


Imagem: Wikipedia

Insper

Big Data, Computação em Larga Escala, High Performance Computing

- Gartner:

Big Data faz referência ao grande volume, variedade e velocidade de dados que ***demandam formas inovadoras e rentáveis de processamento da informação****, para melhor percepção e tomada de decisão.*

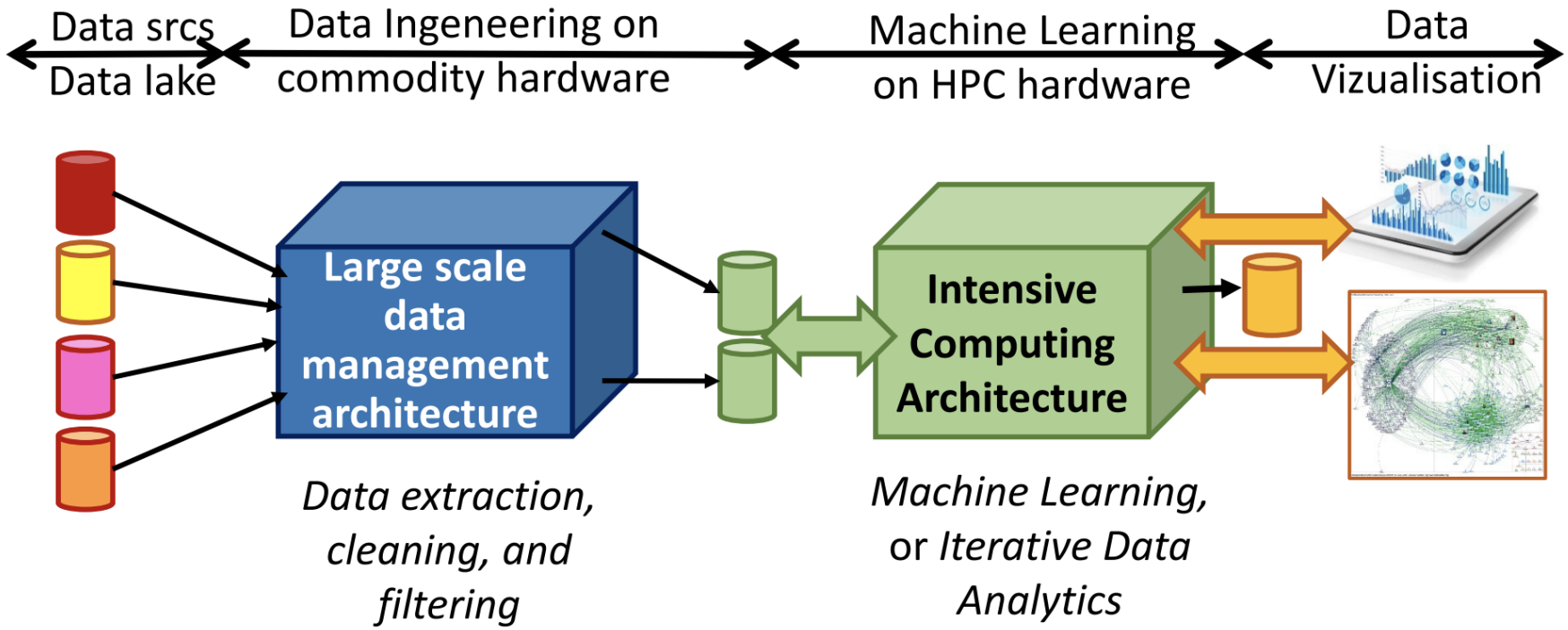


Imagem: Stéphane Vialle & Gianluca Quercini, CentraleSupélec – Universitè Paris-Saclay

Big Data, Computação em Larga Escala, High Performance Computing

- Gartner:

Big Data faz referência ao grande volume, variedade e velocidade de dados que ***demandam formas inovadoras e rentáveis de processamento da informação****, para melhor percepção e tomada de decisão.*

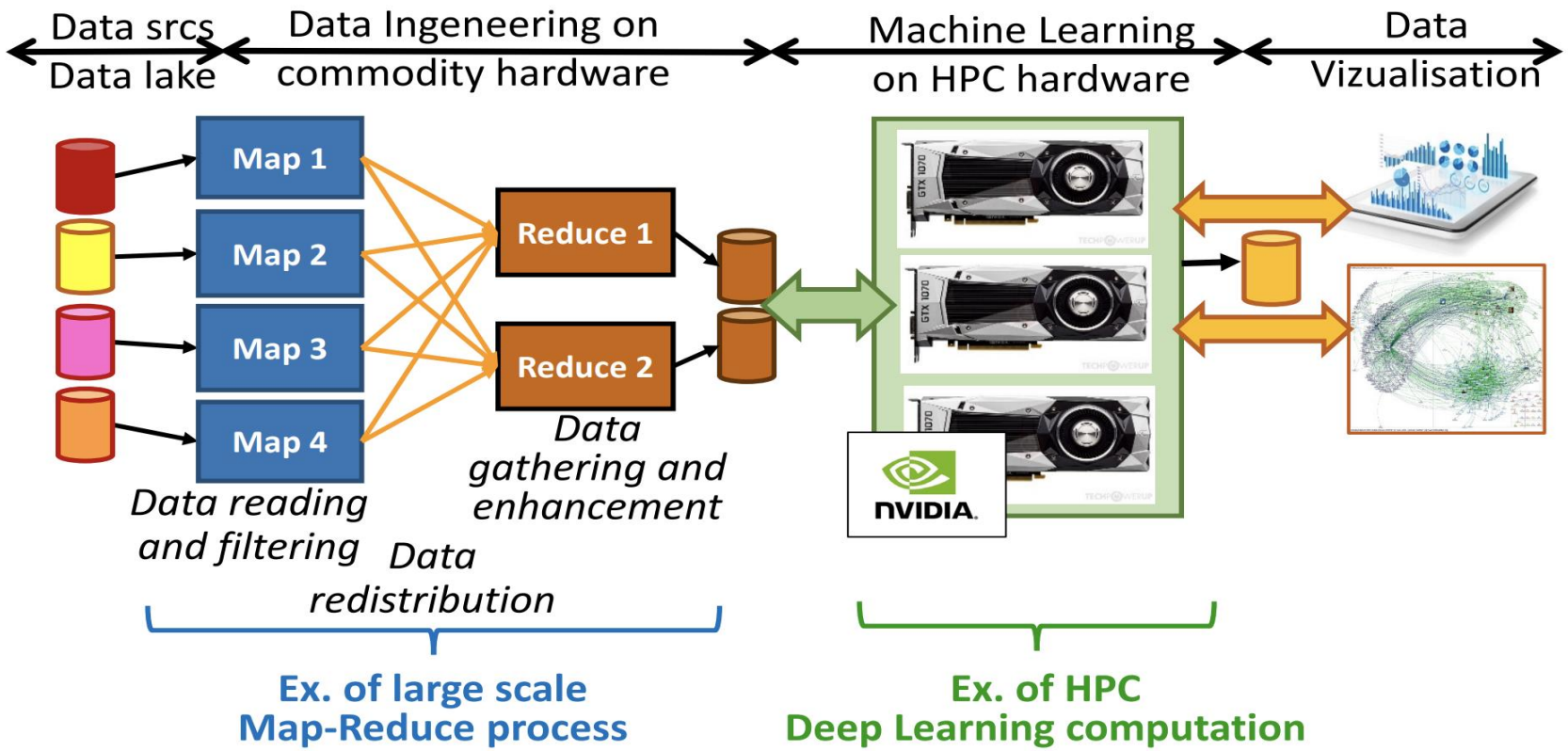
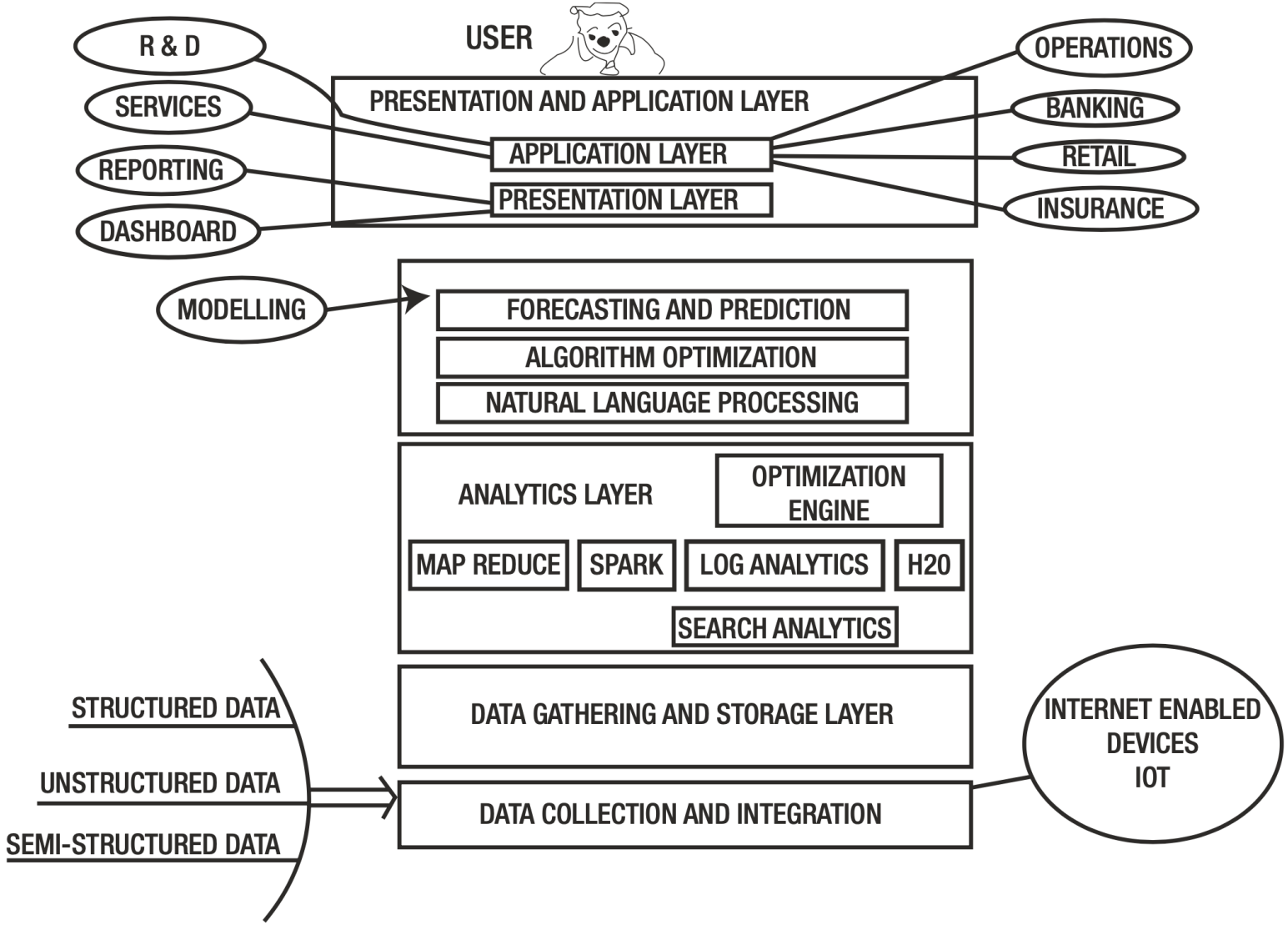


Imagem: Stéphane Vialle & Gianluca Quercini, CentraleSupélec – Universitè Paris-Saclay

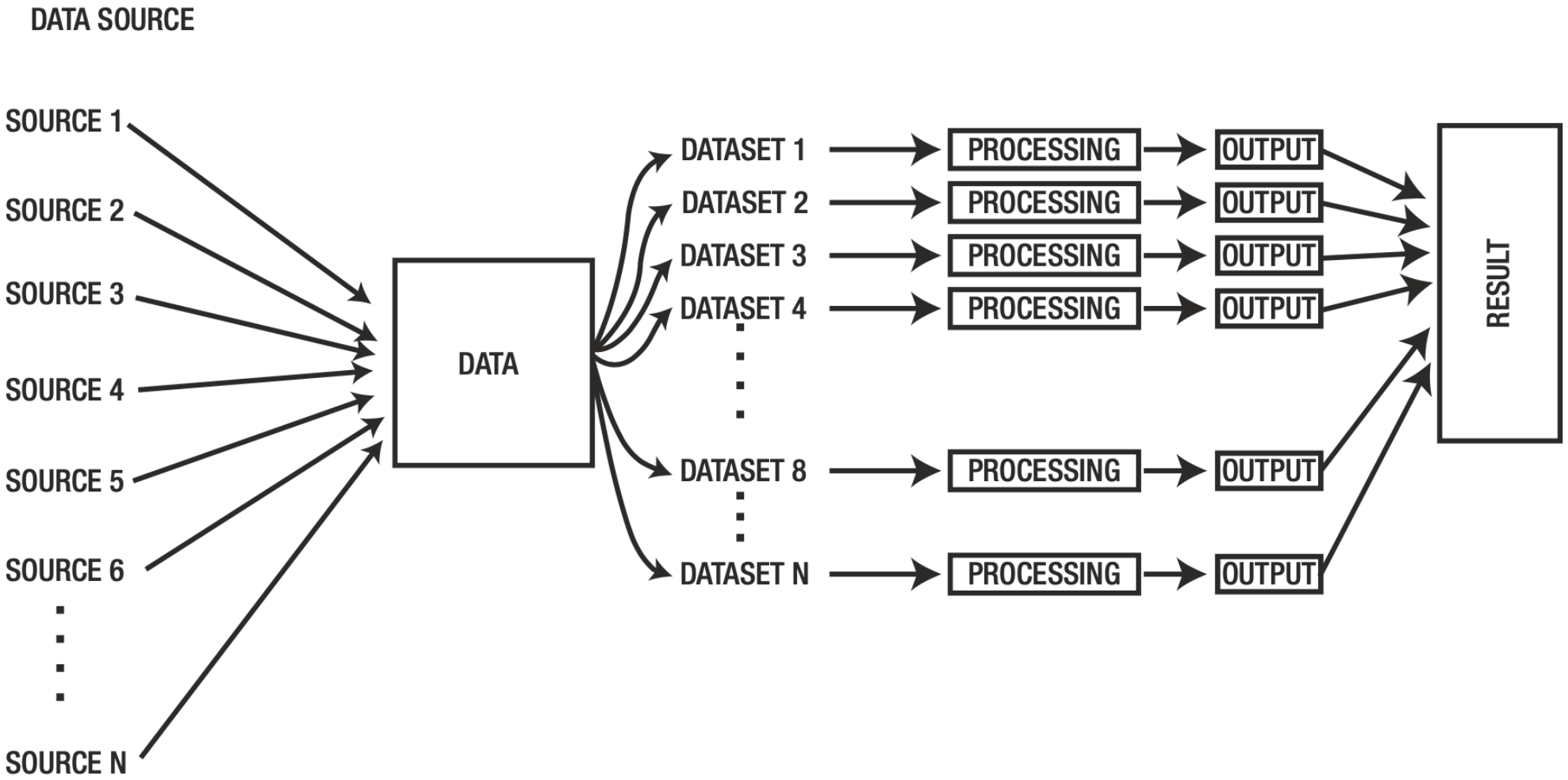
Big Data - Stack



Kashyap, Patanjali. *Machine Learning for Decision Makers: Cognitive Computing Fundamentals for Better Decision Making*. Apress, 2018.

Insper

Ideia básica

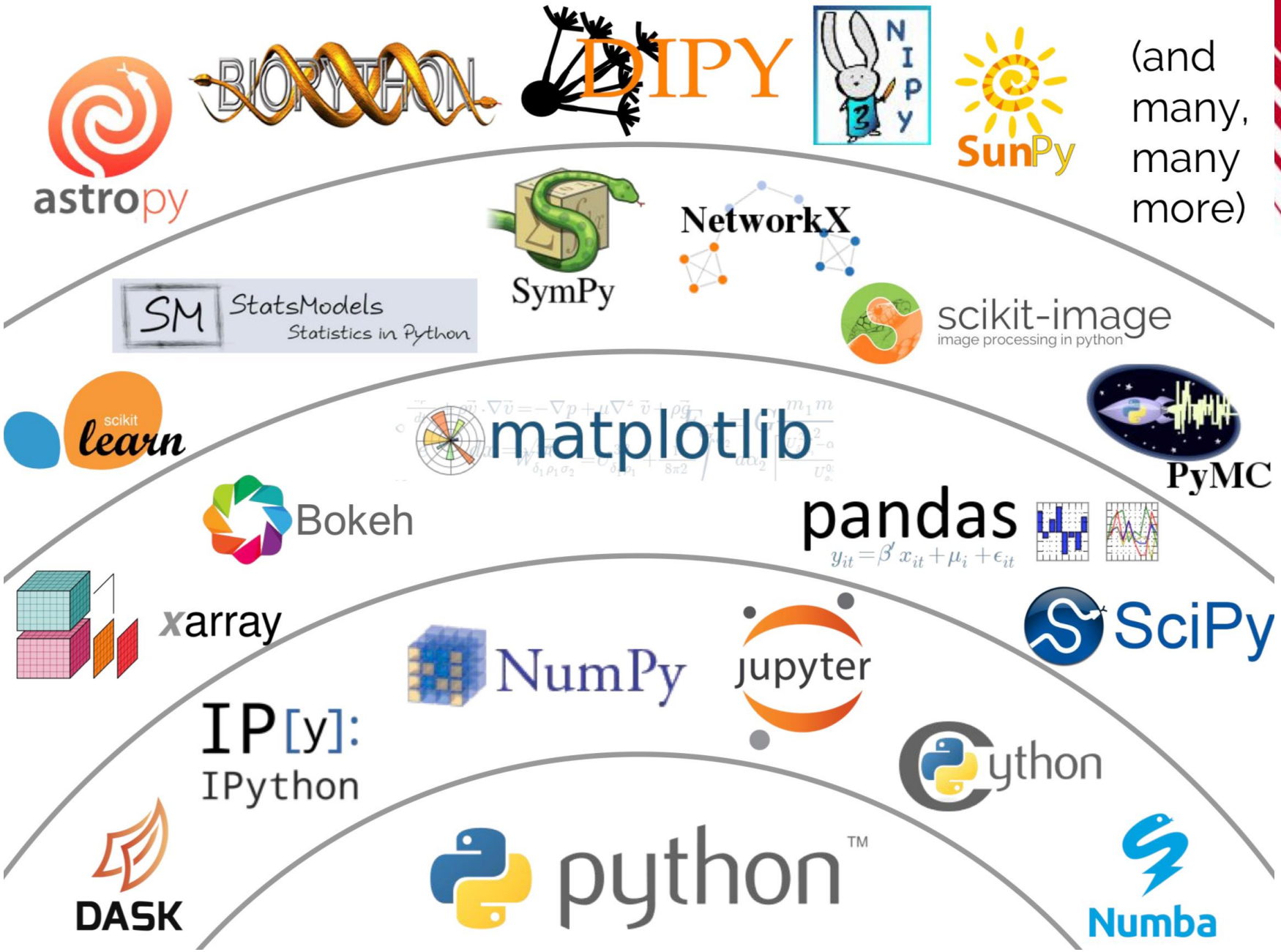


Kashyap, Patanjali. *Machine Learning for Decision Makers: Cognitive Computing Fundamentals for Better Decision Making*. Apress, 2018.

Insper

Prática de hoje

Dask



Leia o paper que descreve o Dask em detalhes http://conference.scipy.org/proceedings/scipy2015/pdfs/matthew_rocklin.pdf