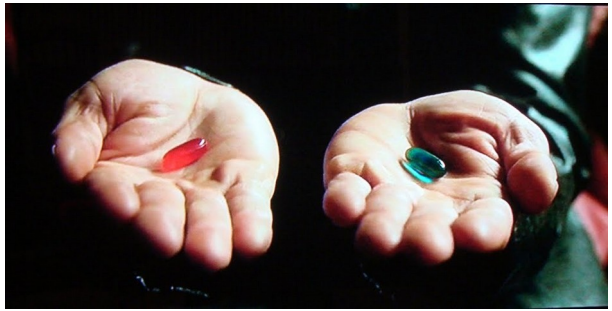
Big Data Analytics

MBI - Manufatura avançada
Indústria 4.0

Robespierre Pita

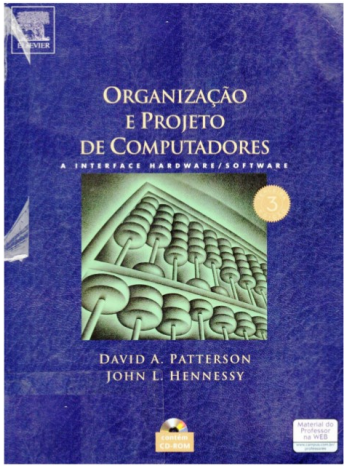
Roteiro

- Introdução: a arquitetura básica
- Processamento paralelo
- Sistemas distribuídos
- Big Data
- Ferramentas de Big Data
  - Hadoop MapReduce
  - Spark
- Big data analytics

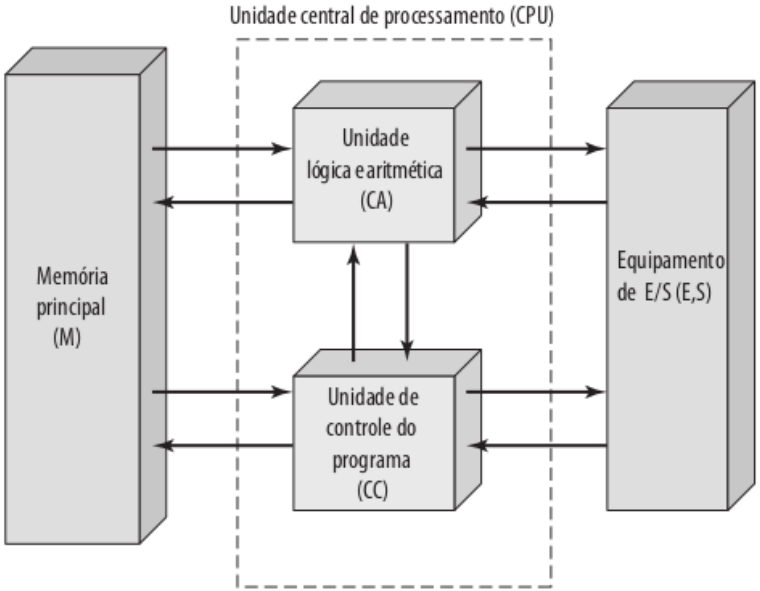


Roteiro

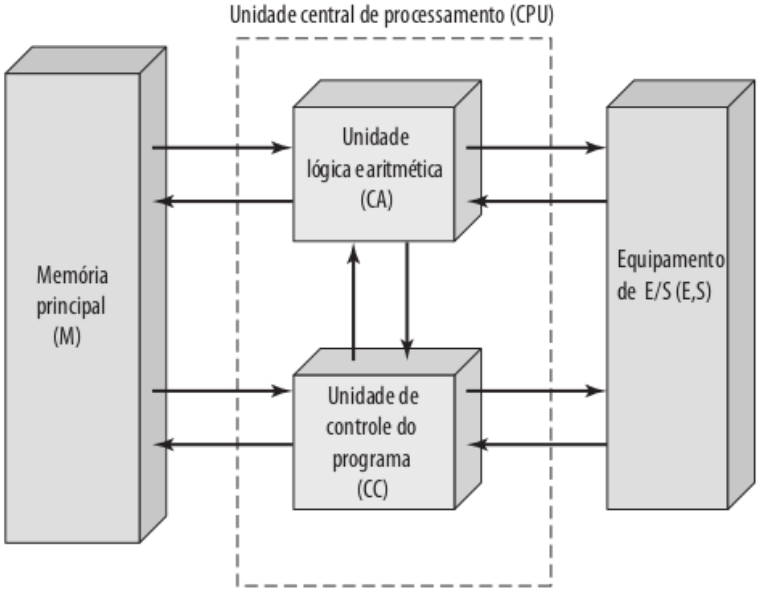
- **Introdução: a arquitetura básica**
- Processamento paralelo
- Sistemas distribuídos
- Big Data
- Ferramentas de Big Data
  - Hadoop MapReduce
  - Spark
- Big data analytics

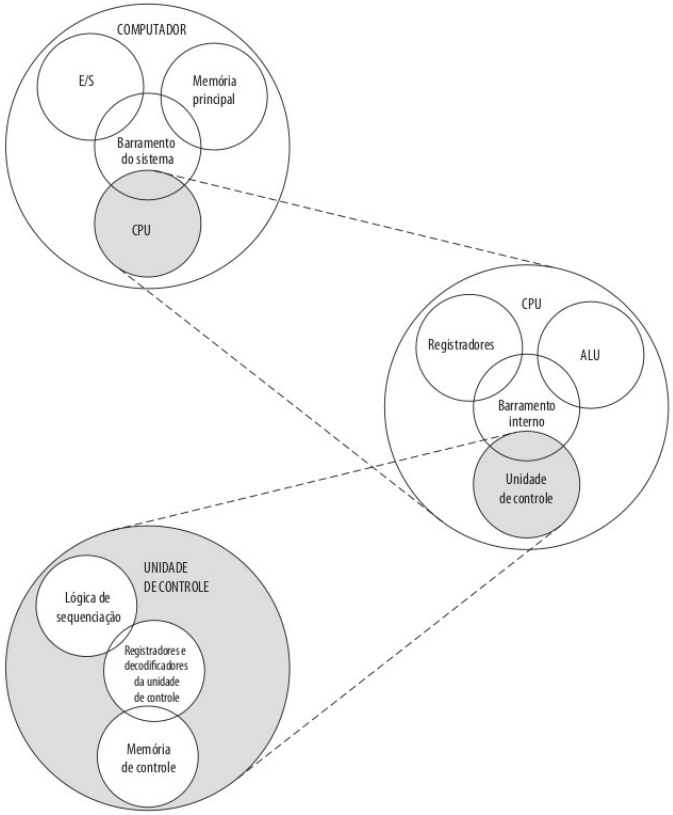


Introdução: a arquitetura básica



Introdução: a arquitetura básica

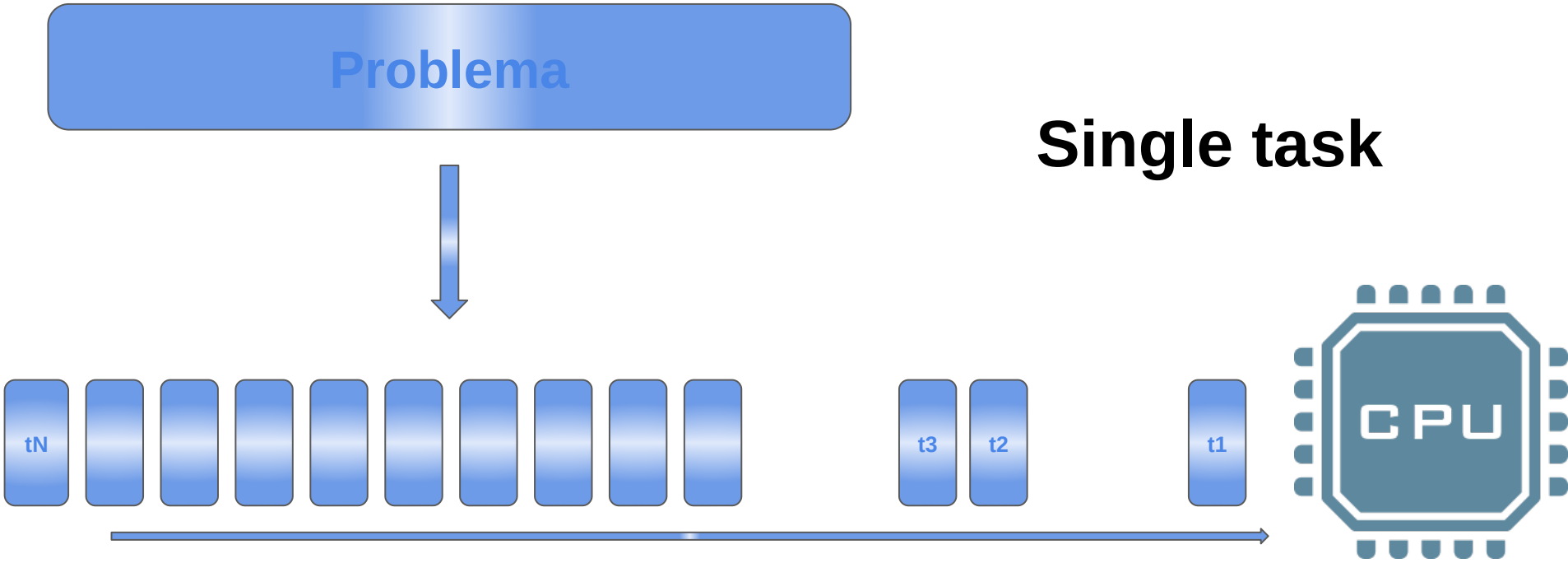




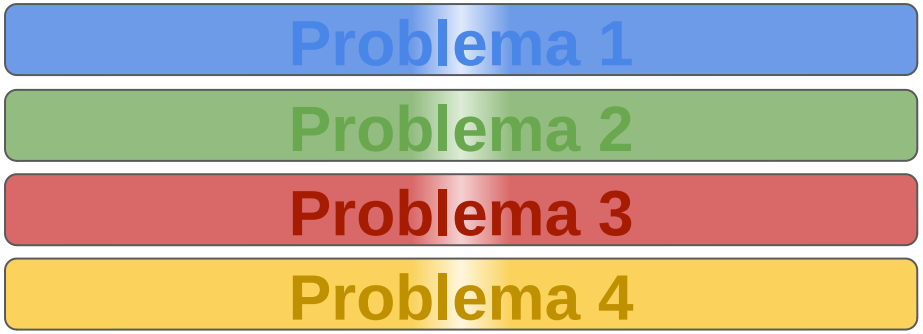
Introdução: a arquitetura básica

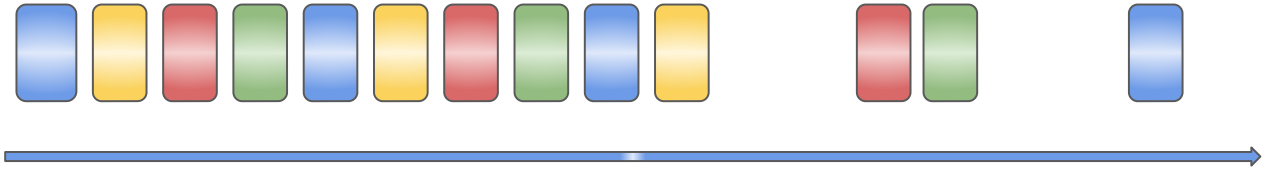
Qual o fluxo de execução quando uma operação chega à CPU?

Introdução: a arquitetura básica

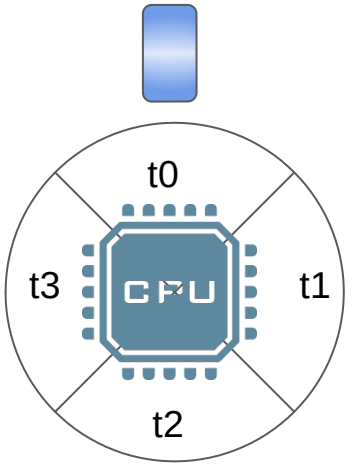


Introdução: a arquitetura básica

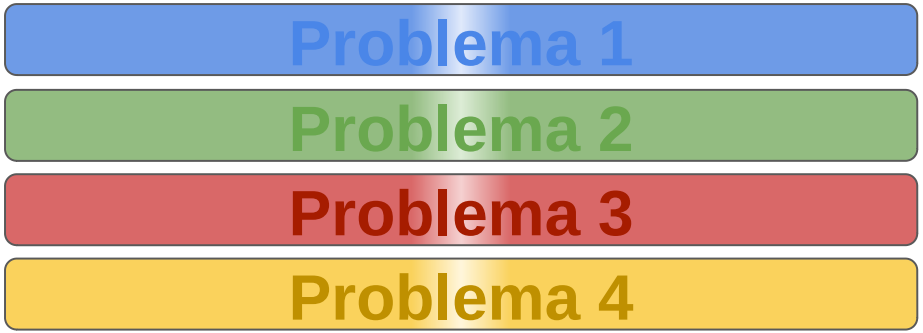




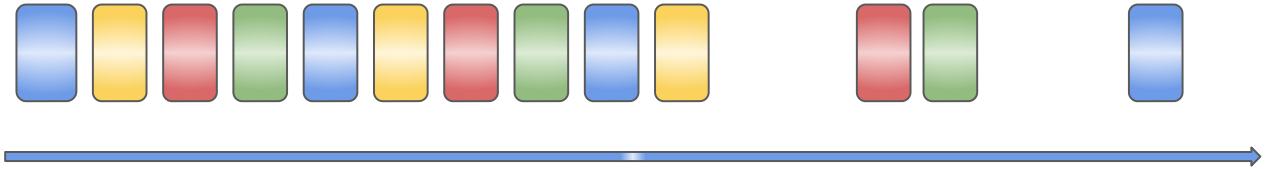
Single task

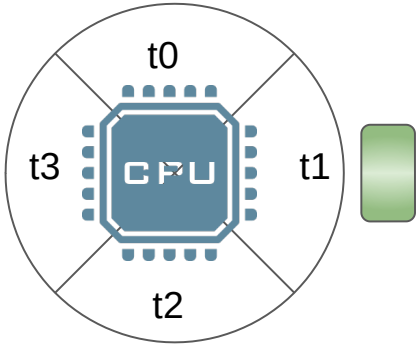


Introdução: a arquitetura básica

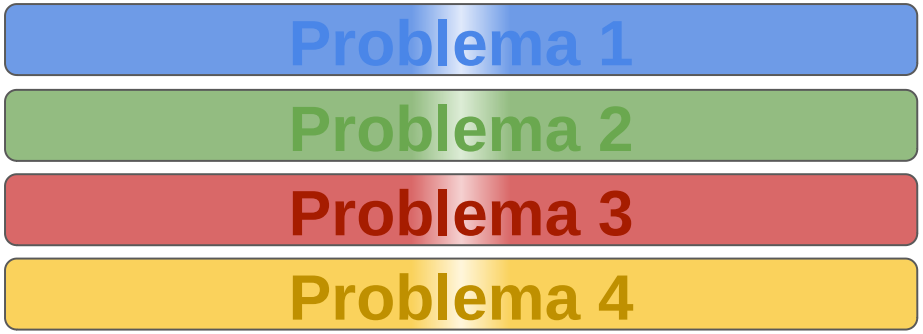


Single task

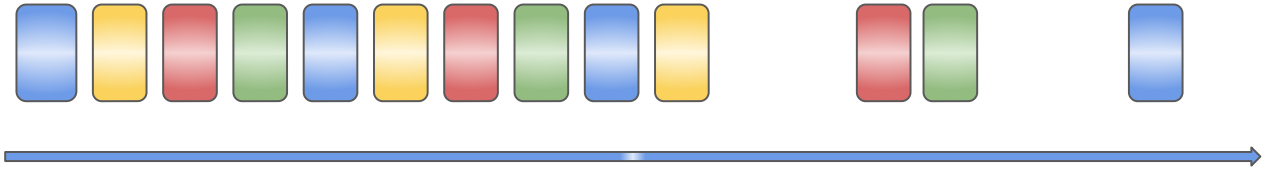


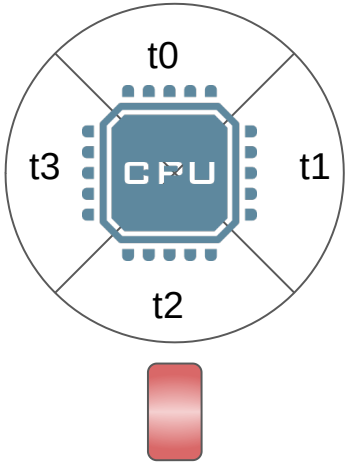


Introdução: a arquitetura básica

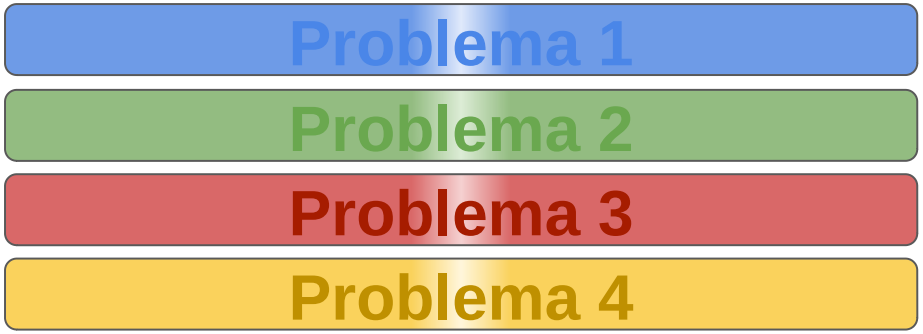


Single task

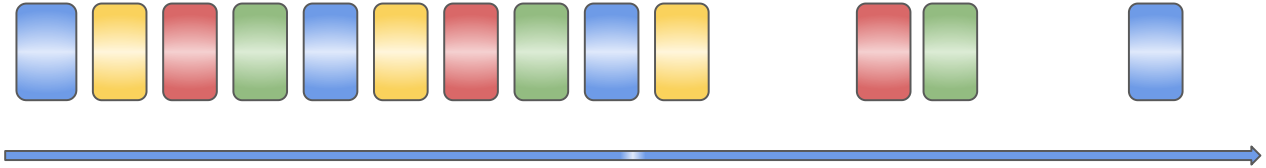


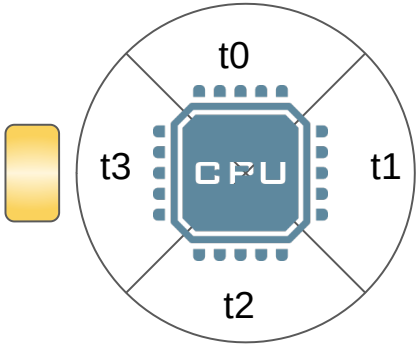


Introdução: a arquitetura básica

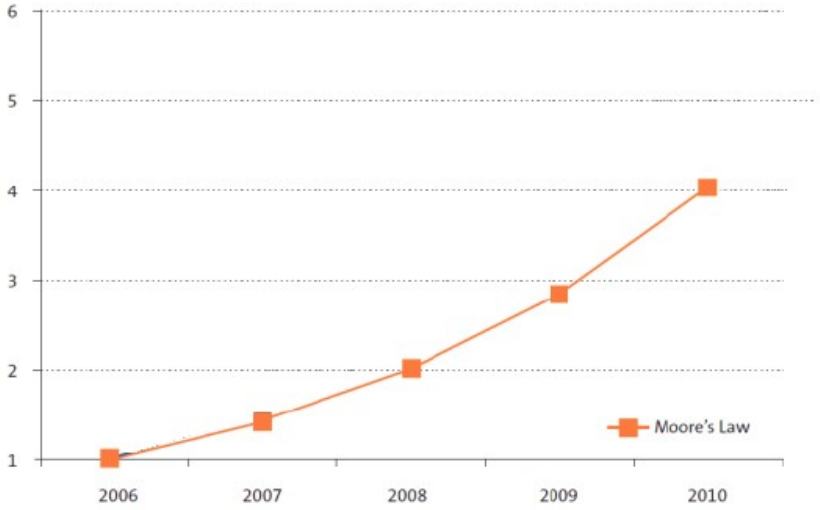


Multi-task

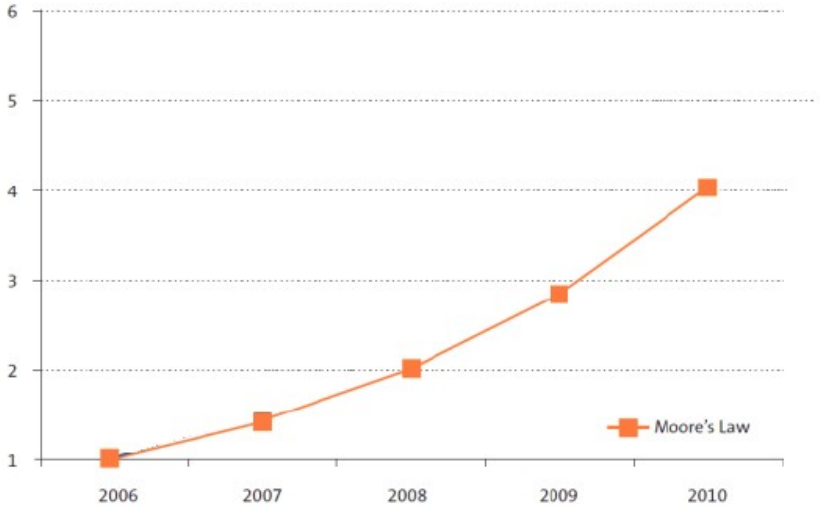




Introdução: a arquitetura básica

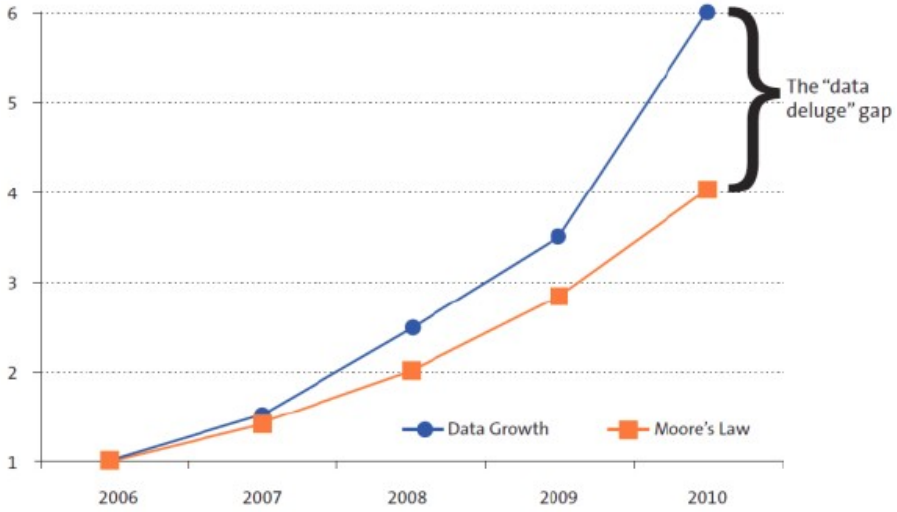


Introdução: a arquitetura básica



O que acontece quando o problema é grande/complexo demais para um único processador?

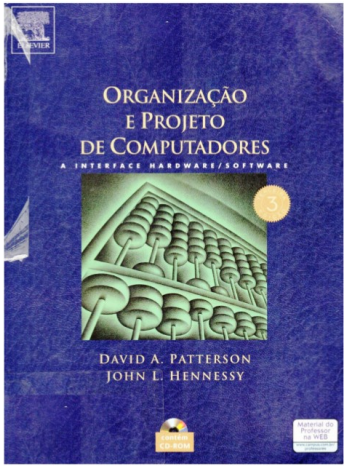
Introdução: a arquitetura básica



O que acontece quando o problema é grande/complexo demais para um único processador?

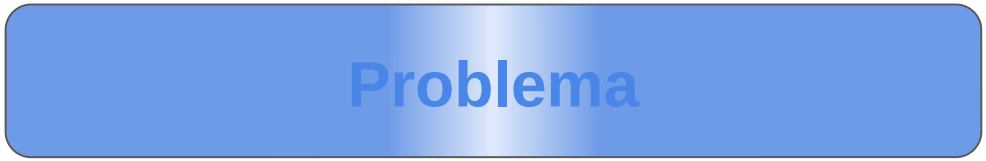
Roteiro

- Introdução: a arquitetura básica
- **Processamento paralelo**
- Sistemas distribuídos
- Big Data
- Ferramentas de Big Data
  - Hadoop MapReduce
  - Spark
- Big data analytics

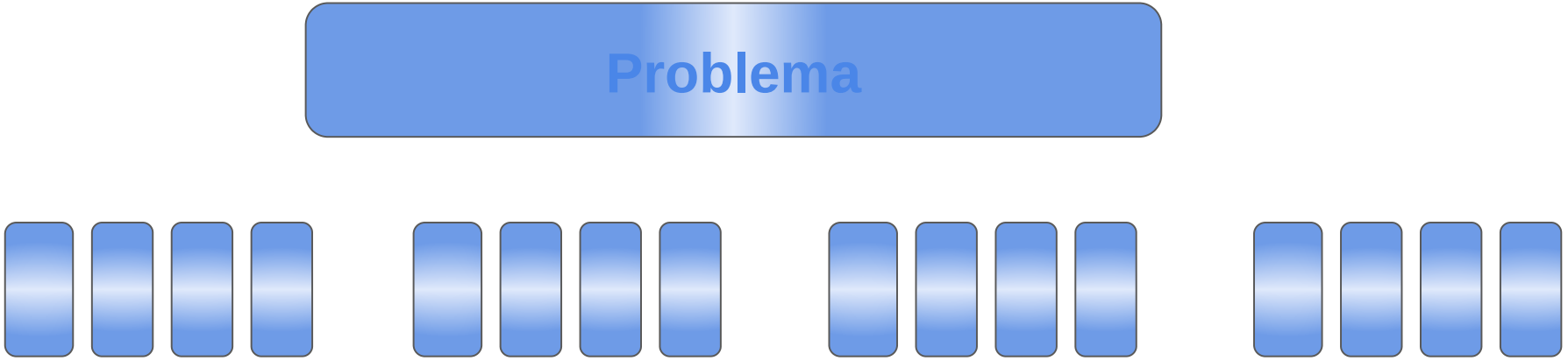


Processamento paralelo

Single task

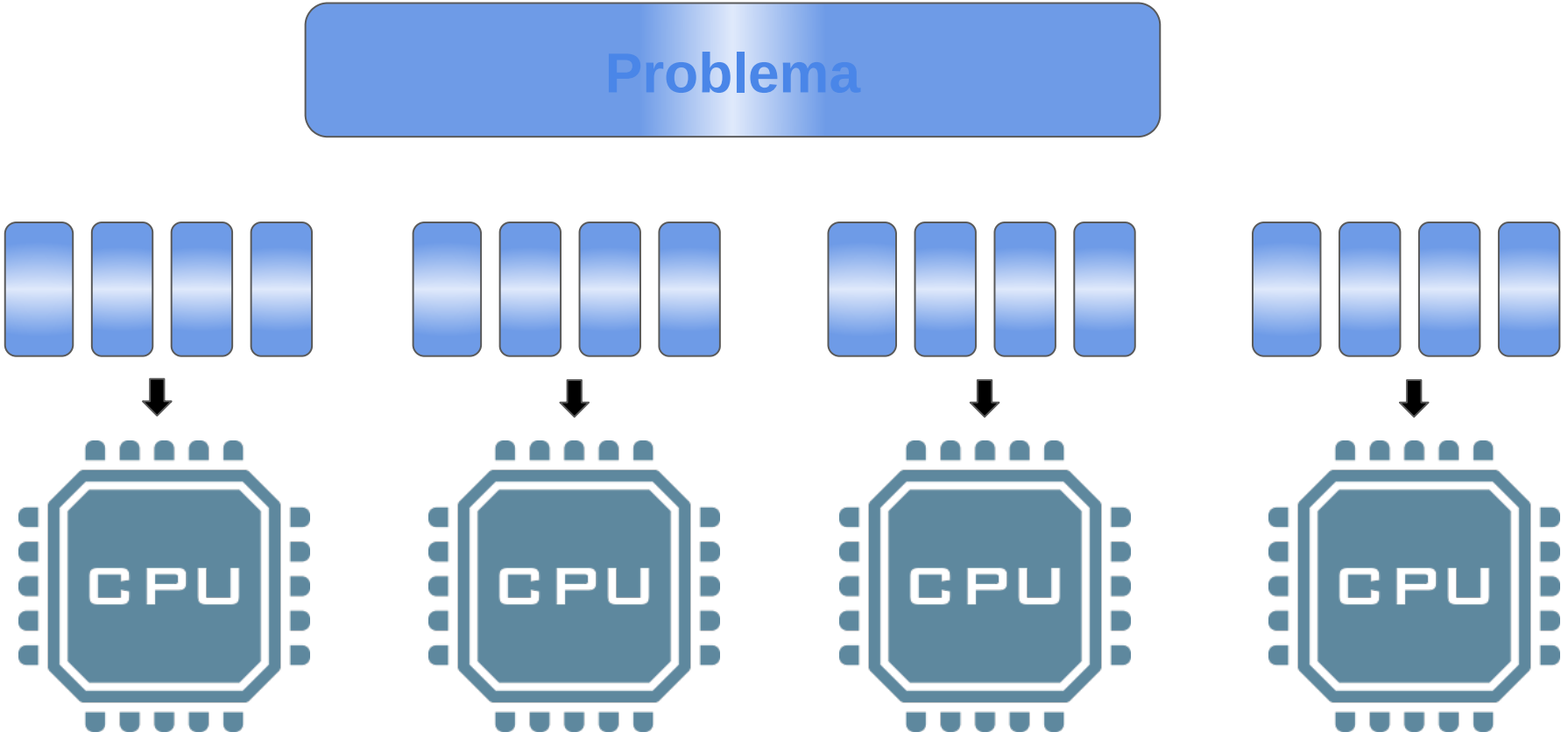


Processamento paralelo

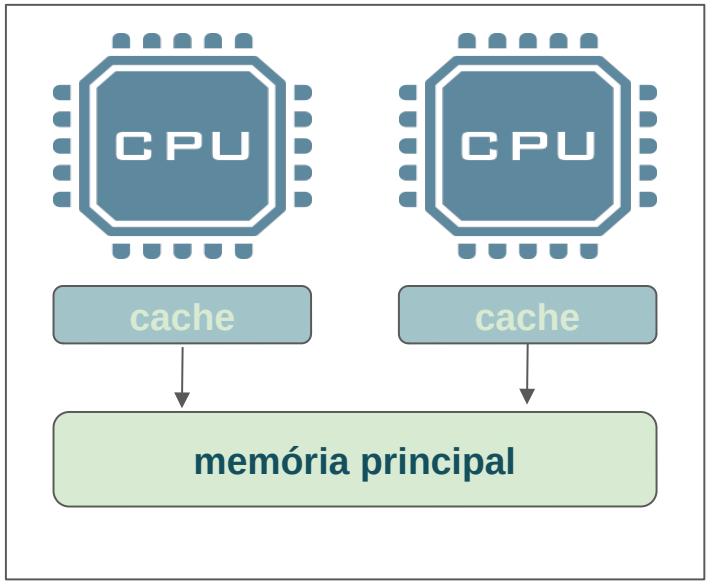


Processamento paralelo

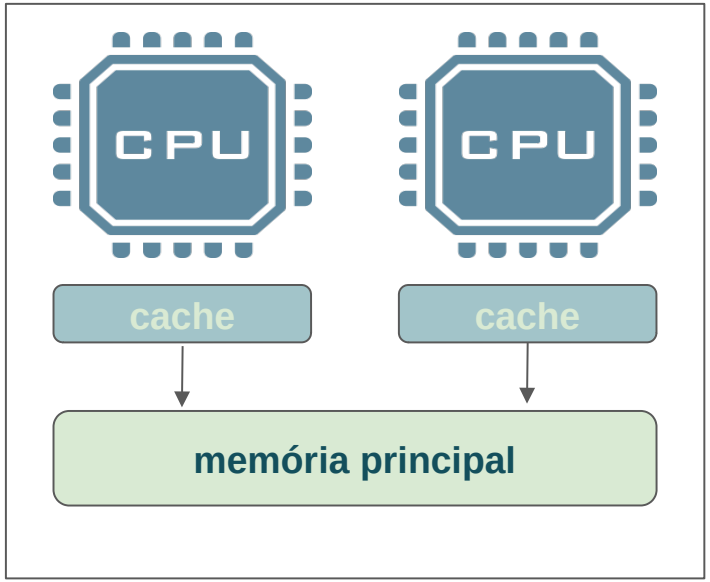
Single task

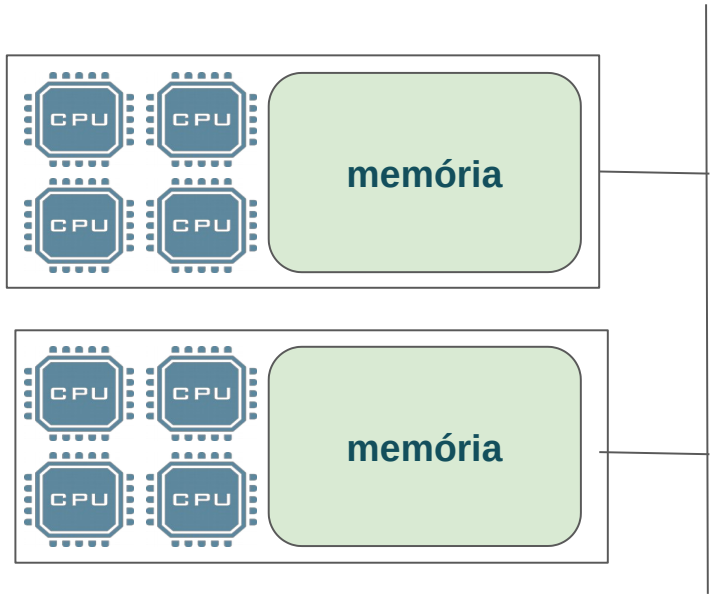


Esquemas de memória

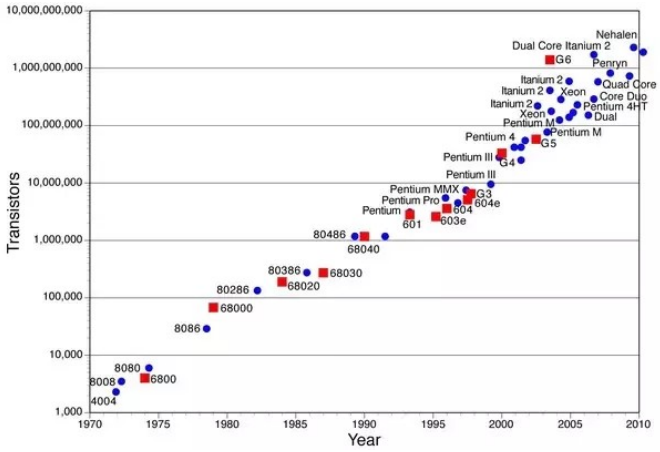


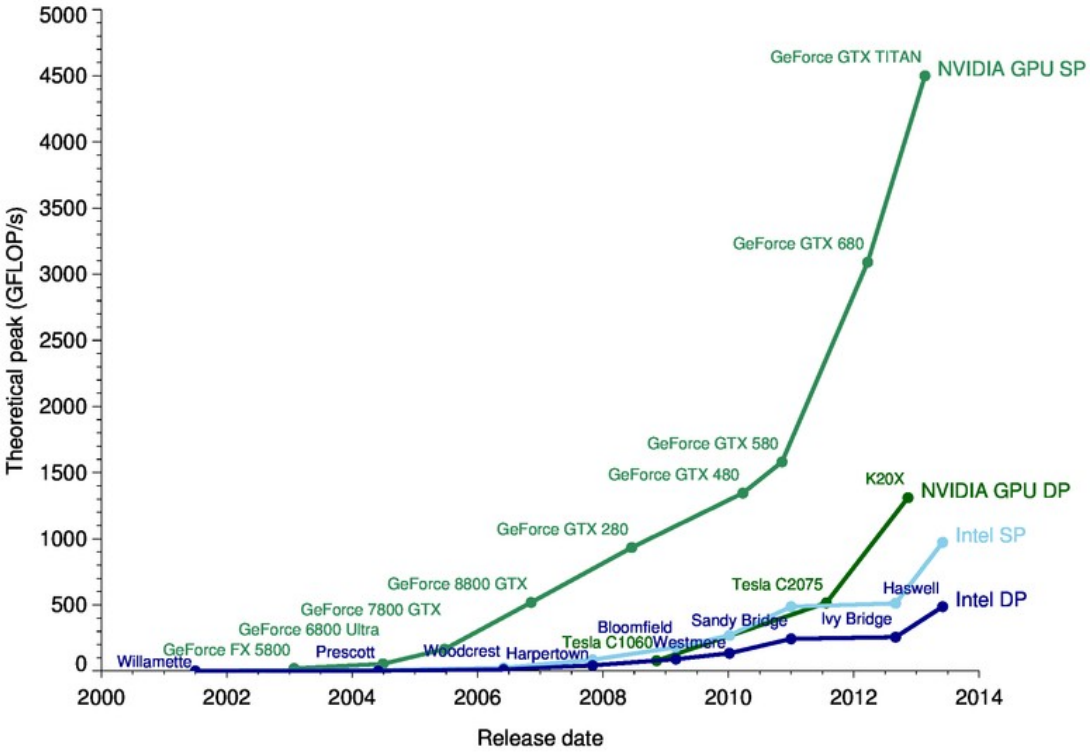
Esquemas de memória





Evolução do processamento paralelo...



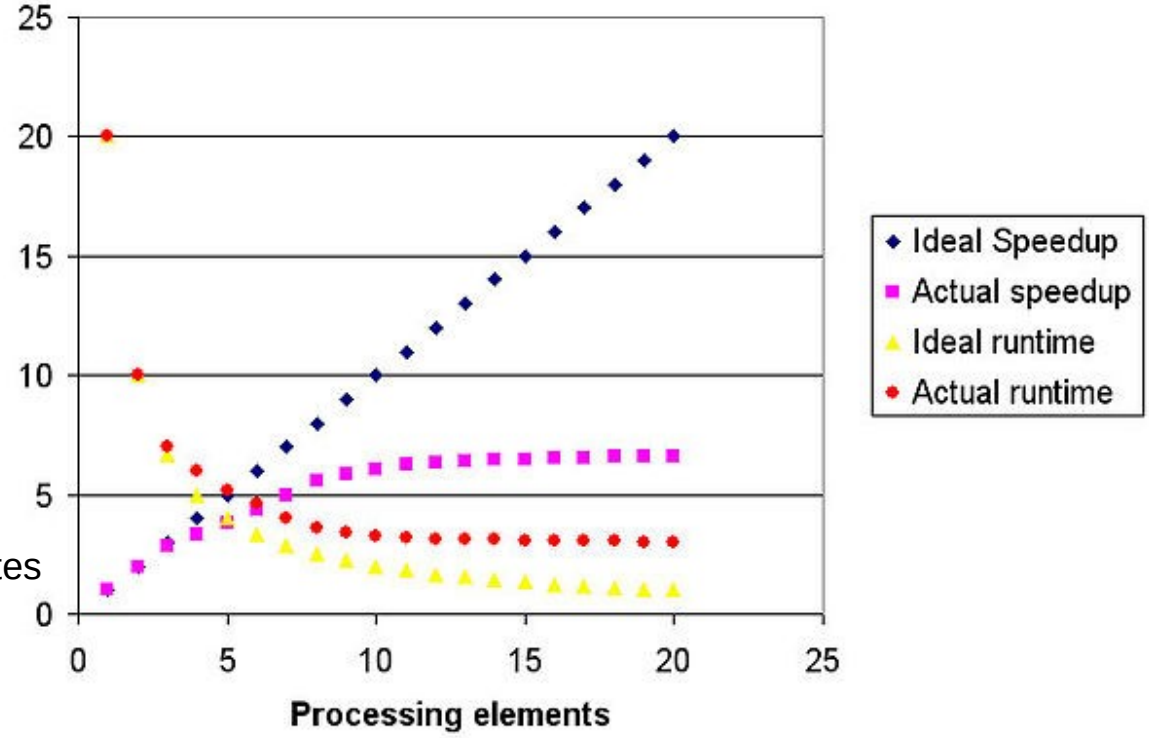


- Espaço físico;
- Temperatura;
- Velocidade dos barramentos;
- Overhead de infraestrutura;
- Custo de crescimento vertical.

Discussão

Quanto mais cores melhor? O aumento de recursos disponíveis sempre será a solução para lidar com problemas complexos em menor tempo?

… e suas limitações.

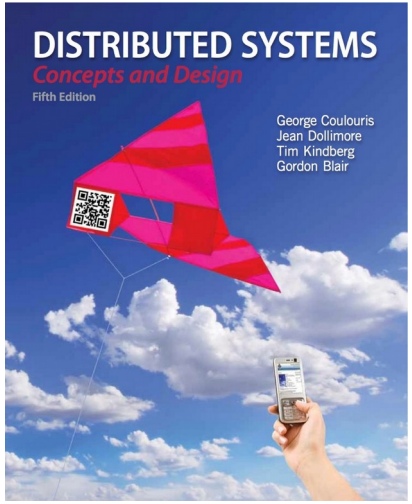


Outras limitações

- Espaço físico (em CPDs, gabinetes e placas;
- Temperatura;
- Velocidade dos barramentos;
- Overhead de infraestrutura;
- Custo de crescimento vertical.

Roteiro

- Introdução: a arquitetura básica
- Processamento paralelo
- **Sistemas distribuídos**
- Big Data
- Ferramentas de Big Data
  - Hadoop MapReduce
  - Spark
- Big data analytics

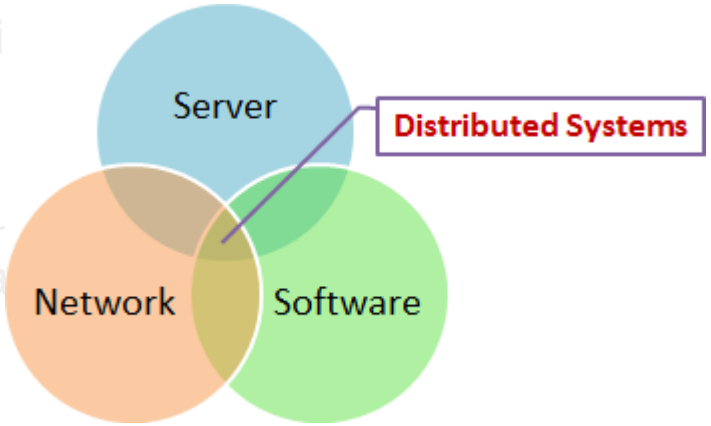


Sistemas distribuídos

Um sistema distribuído é uma coleção de computadores autônomos, ligados por uma **rede de computadores, e equipados com software de sistema distribuído**.

Sistema no qual os componentes de hardware ou software, localizados em computadores interligados em rede, **se comuni**
apenas enviando mensagens entre si.

Um sistema distribuído é um conjunto de compu
apresenta a seus usuários como um sistema

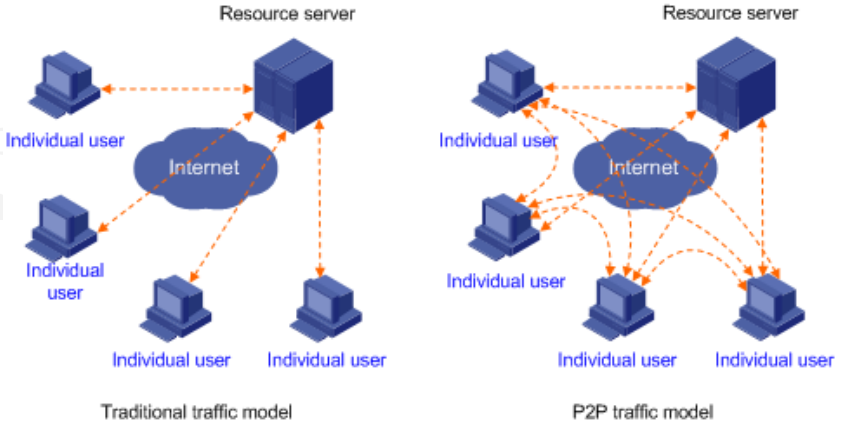


Sistemas distribuídos

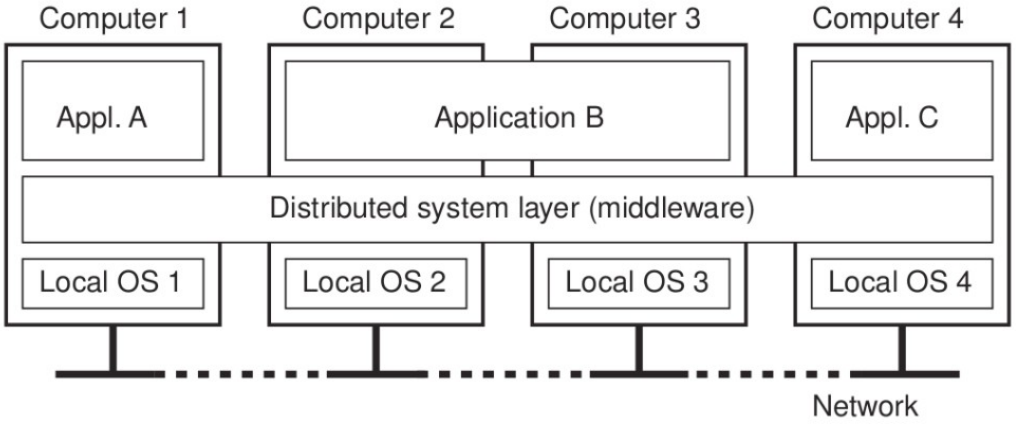
Um sistema distribuído é uma coleção de computadores autônomos, ligados por uma **rede de computadores, e equipados com software de sistema distribuído**.

Sistema no qual os componentes de hardware ou software, localizados em computadores interligados em rede, **se comunicam e coordenam suas ações apenas enviando mensagens entre si**.

Um sistema distribuído é um conjunto de comp
apresenta a seus usuários como um sistem



Sistemas distribuídos



Um sistema distribuído é um conjunto de computadores independentes que **se apresenta a seus usuários como um sistema único e coerente.**

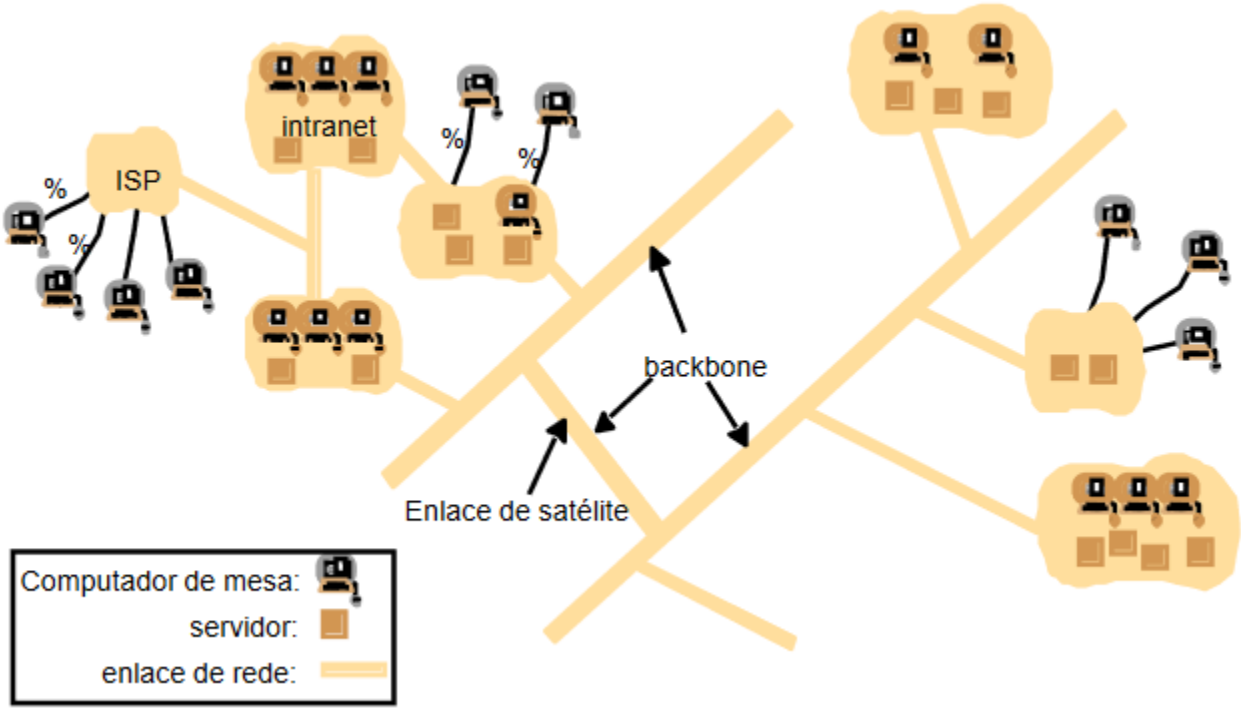
Principais desafios de um SD

Concorrência – execução concorrente de programas/recursos compartilhados;

Inexistência de relógio global – programas cooperam trocando mensagens e coordenam suas ações a partir de uma noção compartilhada de tempo;

Falhas independentes – qualquer componente do sistema pode falhar, e as falhas não são imediatamente percebidas pelos demais componentes do sistema.

Principais desafios de um SD



Transparências: Heterogeneidade

O acesso a serviços e a execução de aplicativos é feito através de um **conjunto heterogêneo de componentes**:

- Redes,
- hardware de computador,
- sistemas operacionais,
- linguagens de programação,
- implementações de diferentes desenvolvedores.

Ex.: Todos os computadores utilizam os protocolos da Internet para se comunicar.

Você consegue pensar em algum exemplo de aplicação desta transparência?

Transparências: Escalabilidade

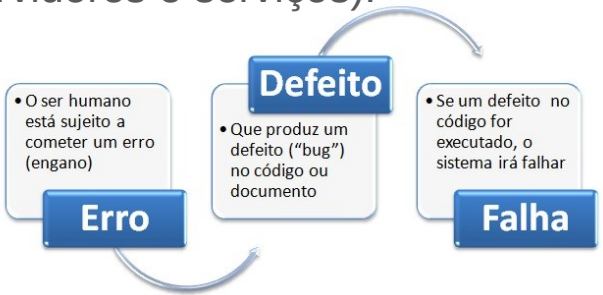
O sistema distribuído **permanece eficiente** quando há um **aumento significativo** no número de recursos e no número de usuários.

Você consegue pensar em algum exemplo de aplicação desta transparência?

Transparências: Falhas

Componentes (software e hardware) de um sistema distribuído **podem falhar**, podendo gerar erros nos resultados do sistema.

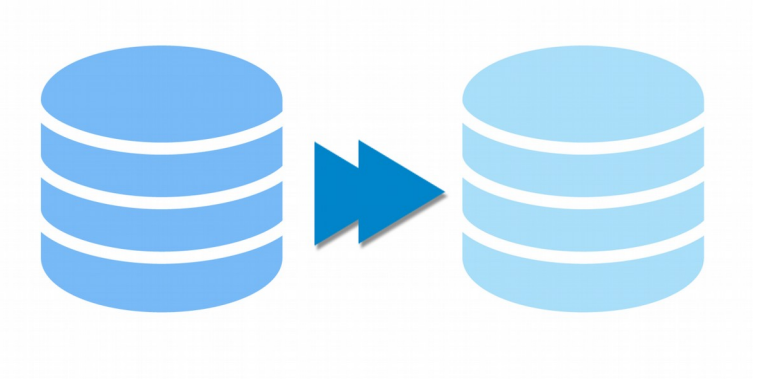
- **Detecção de Falhas**: Detectar falhas (sinc.) ou suspeitas (assinc.).
- **Mascaramento de Falhas**: ocultar falhas (retransmitir ou replicar dados).
- **Tolerância à Falhas**: Executar serviços apesar das falhas.
- **Recuperação de Falhas**: recuperar um estado consistente após uma falha.
- **Redundância**: componentes redundantes (rotas, servidores e serviços).



Você consegue pensar em algum exemplo de aplicação desta transparência?

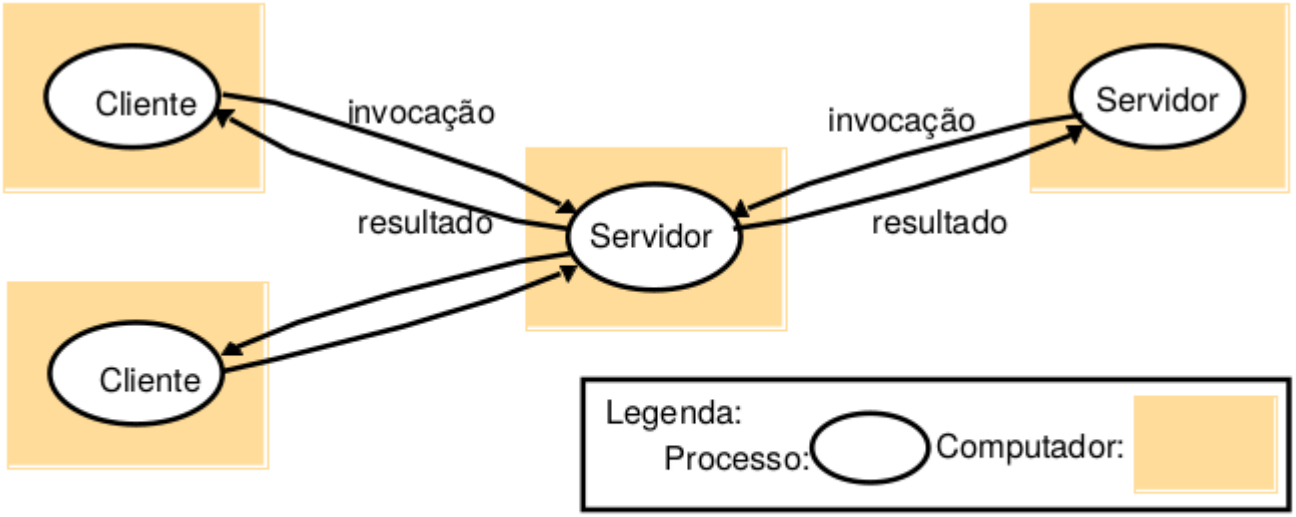
Transparências: Replicação

Permite que **várias instâncias** dos recursos sejam usadas para aumentar o desempenho e a confiabilidade, **sem conhecimento das réplicas** por parte dos usuários ou dos programadores de aplicativos;



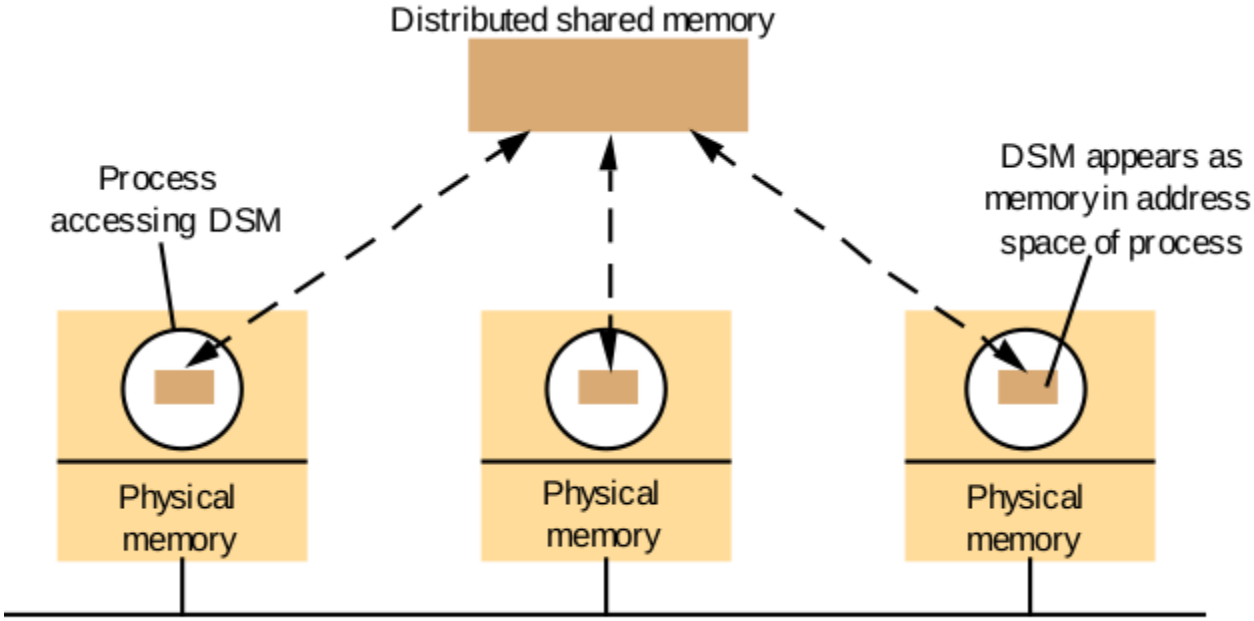
Você consegue pensar em algum exemplo de aplicação desta transparência?

Arquiteturas: Cliente-Servidor

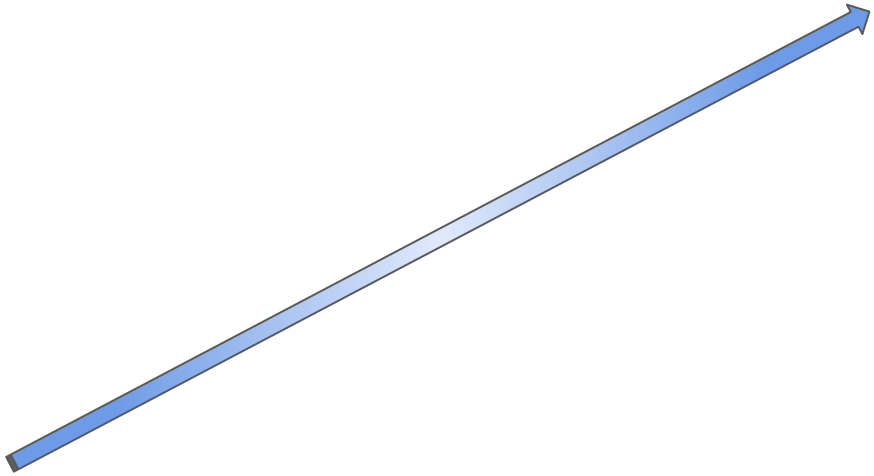


Clientes realizam pedidos a servidores.

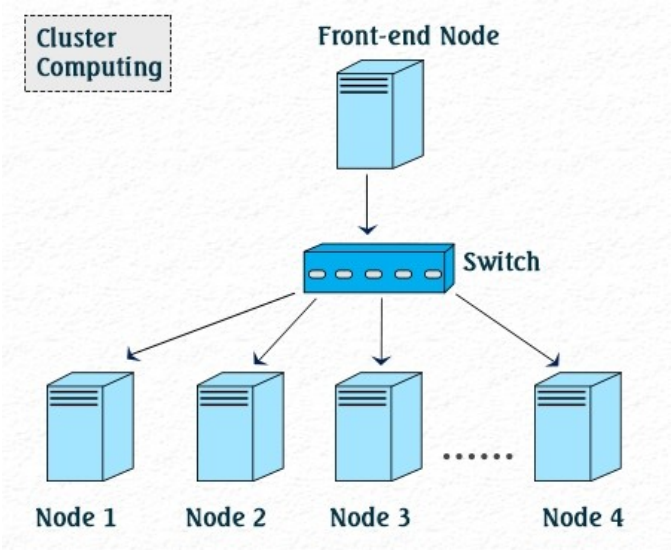
Esquema de memória compartilhada



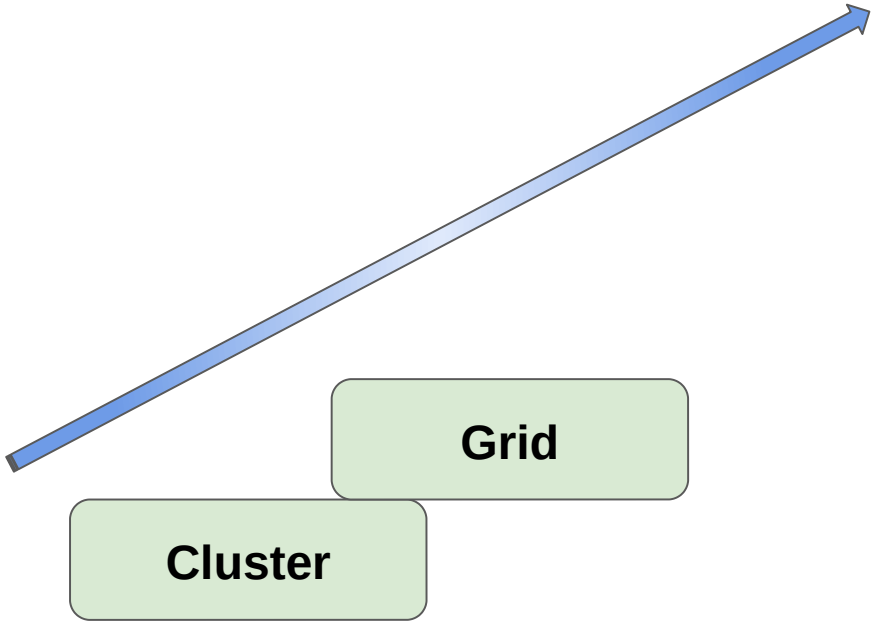
Evolução dos sistemas distribuídos

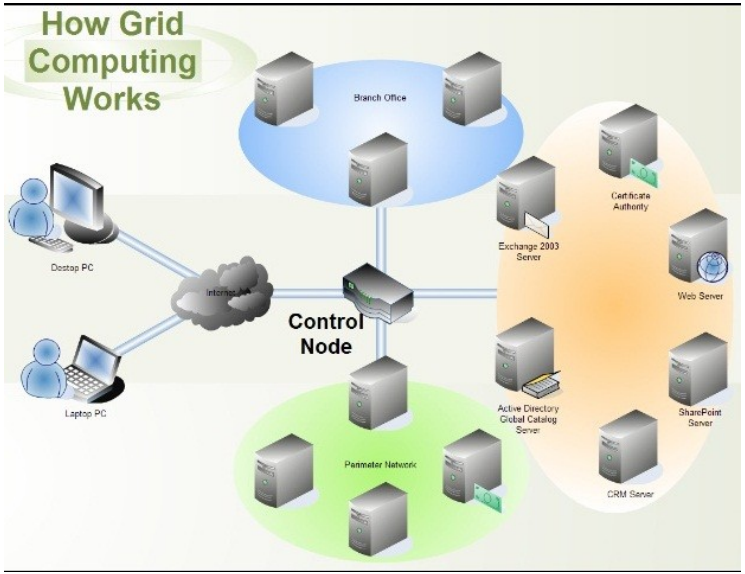


Cluster

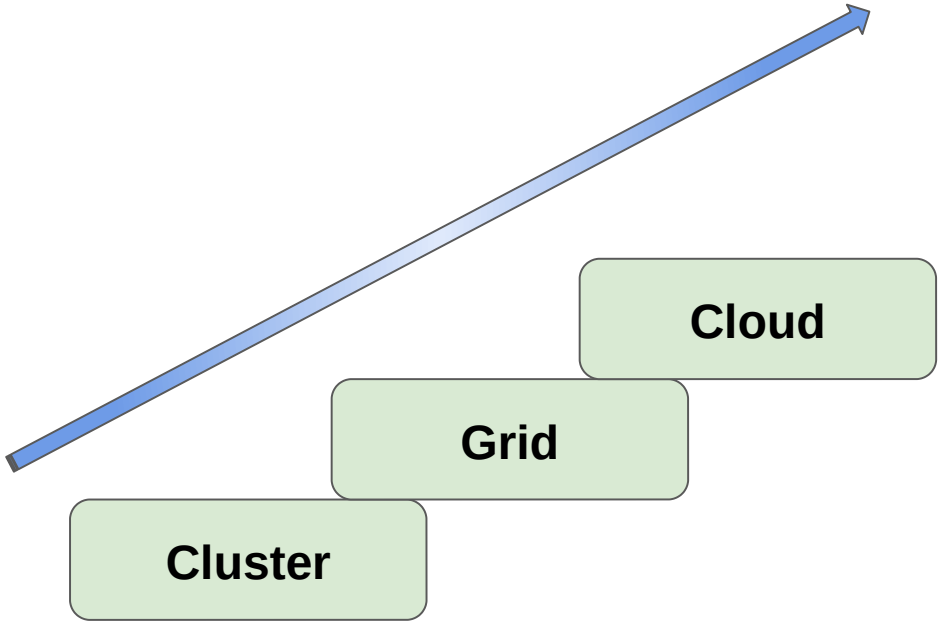


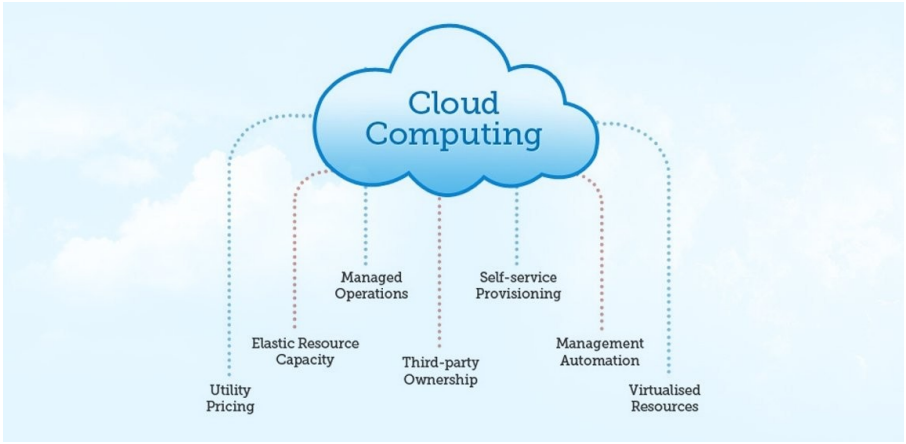
Evolução dos sistemas distribuídos



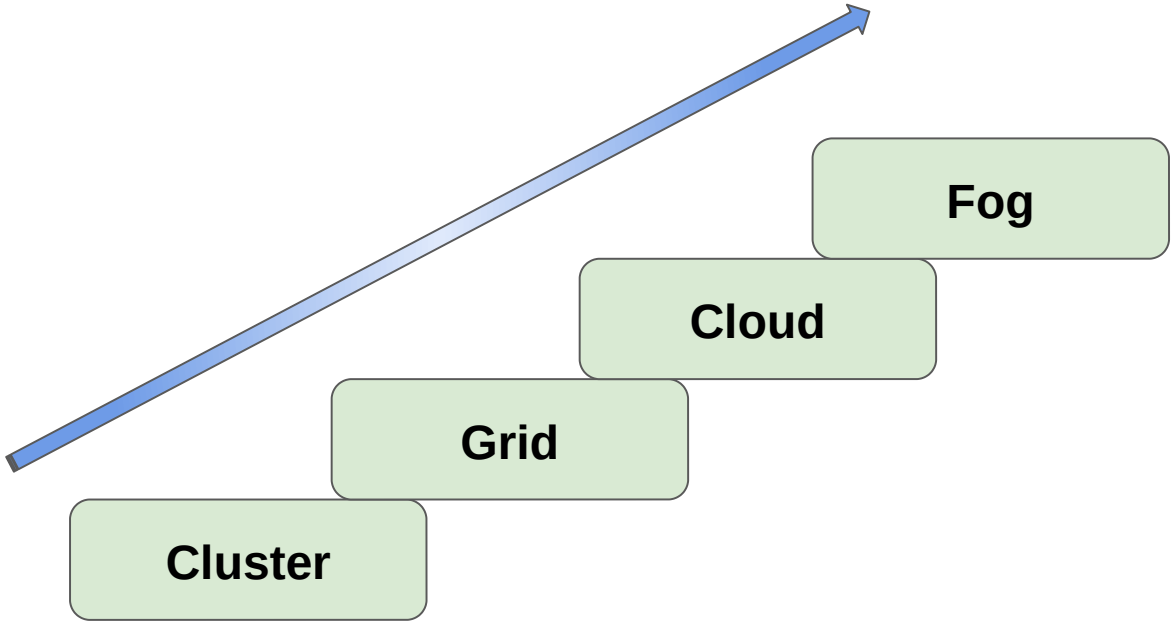


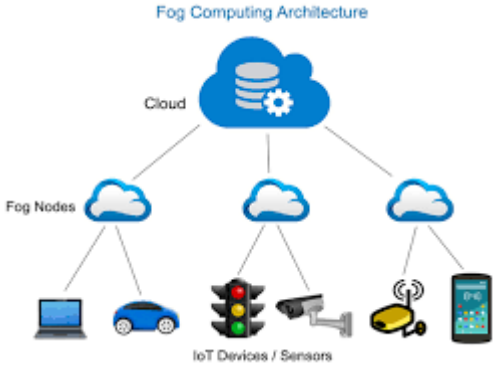
Evolução dos sistemas distribuídos





Evolução dos sistemas distribuídos

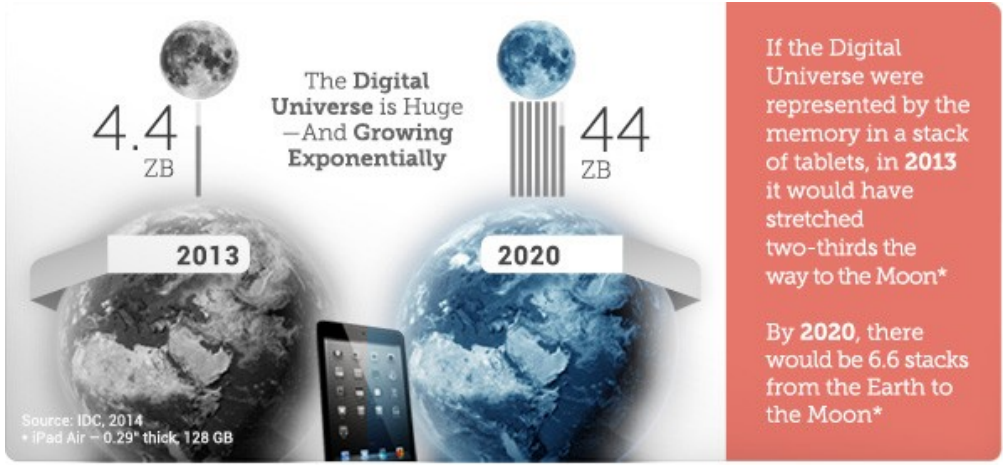


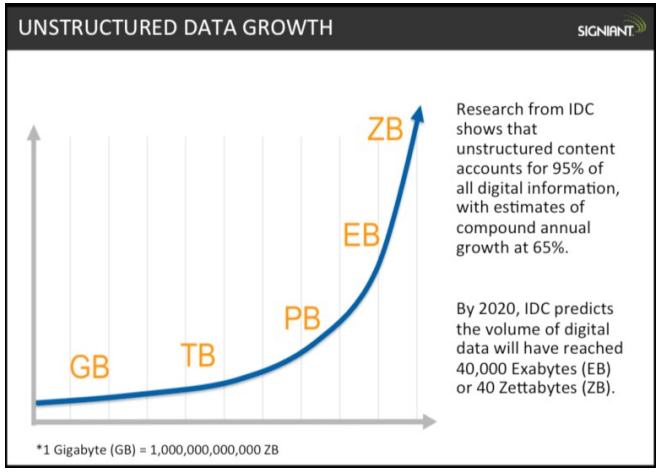


Roteiro

- Introdução: a arquitetura básica
- Processamento paralelo
- Sistemas distribuídos
- **Big Data**
- Ferramentas de Big Data
  - Hadoop MapReduce
  - Spark
- Big data analytics

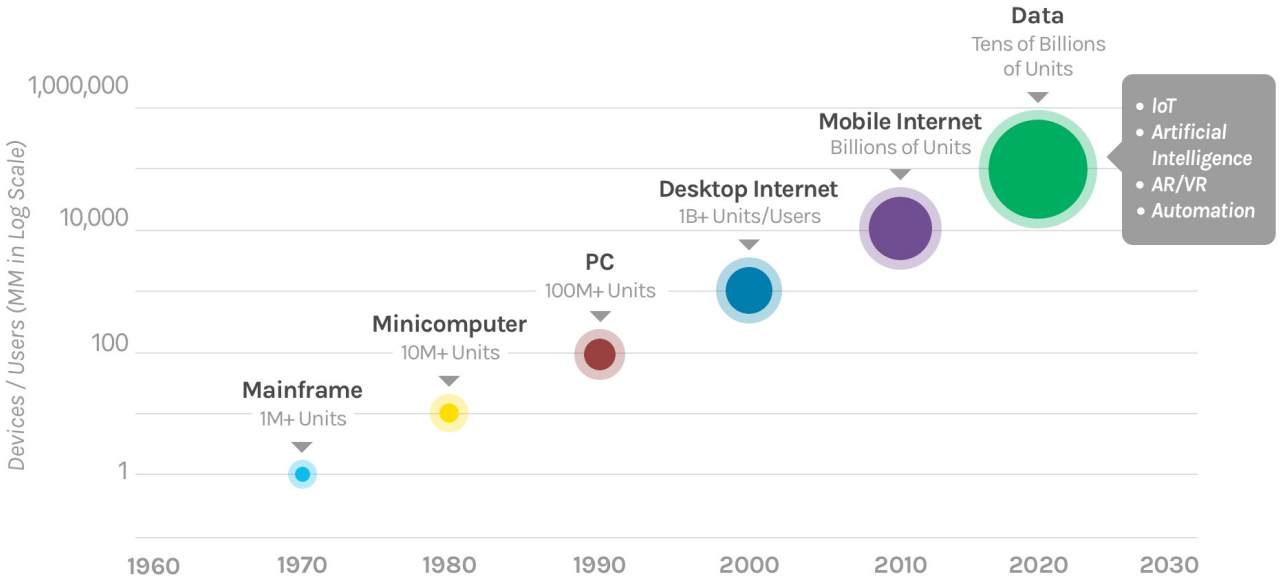
Mas afinal, o que é Big Data?



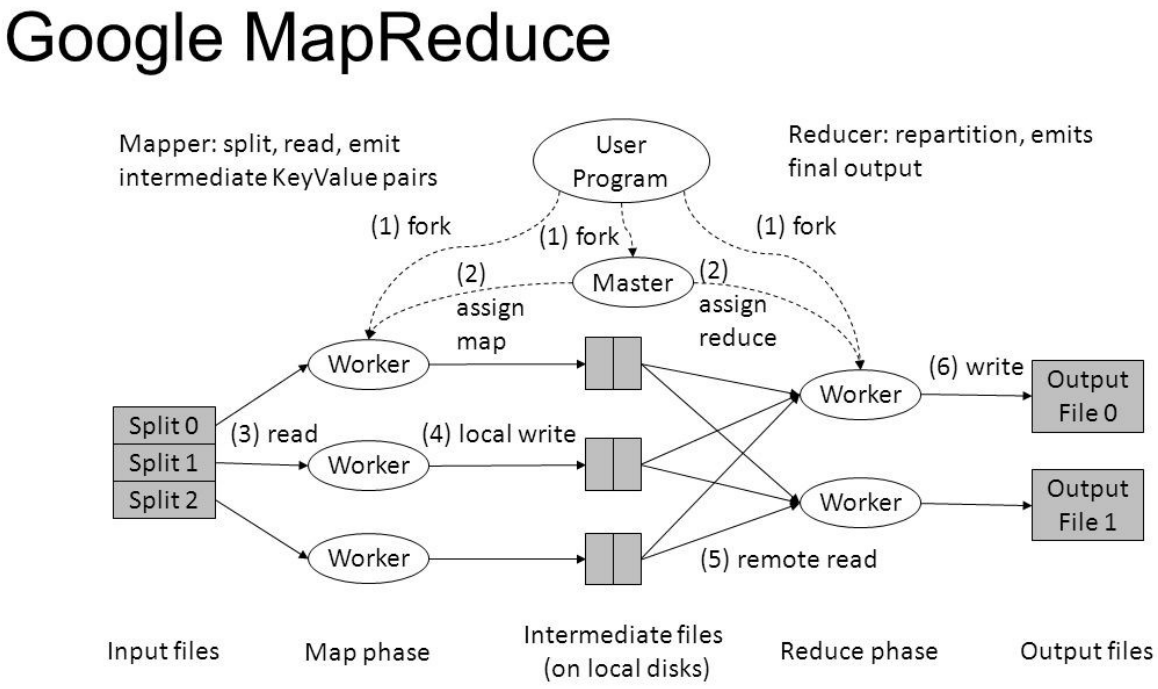


E o que a indústria tem a ver com isso?

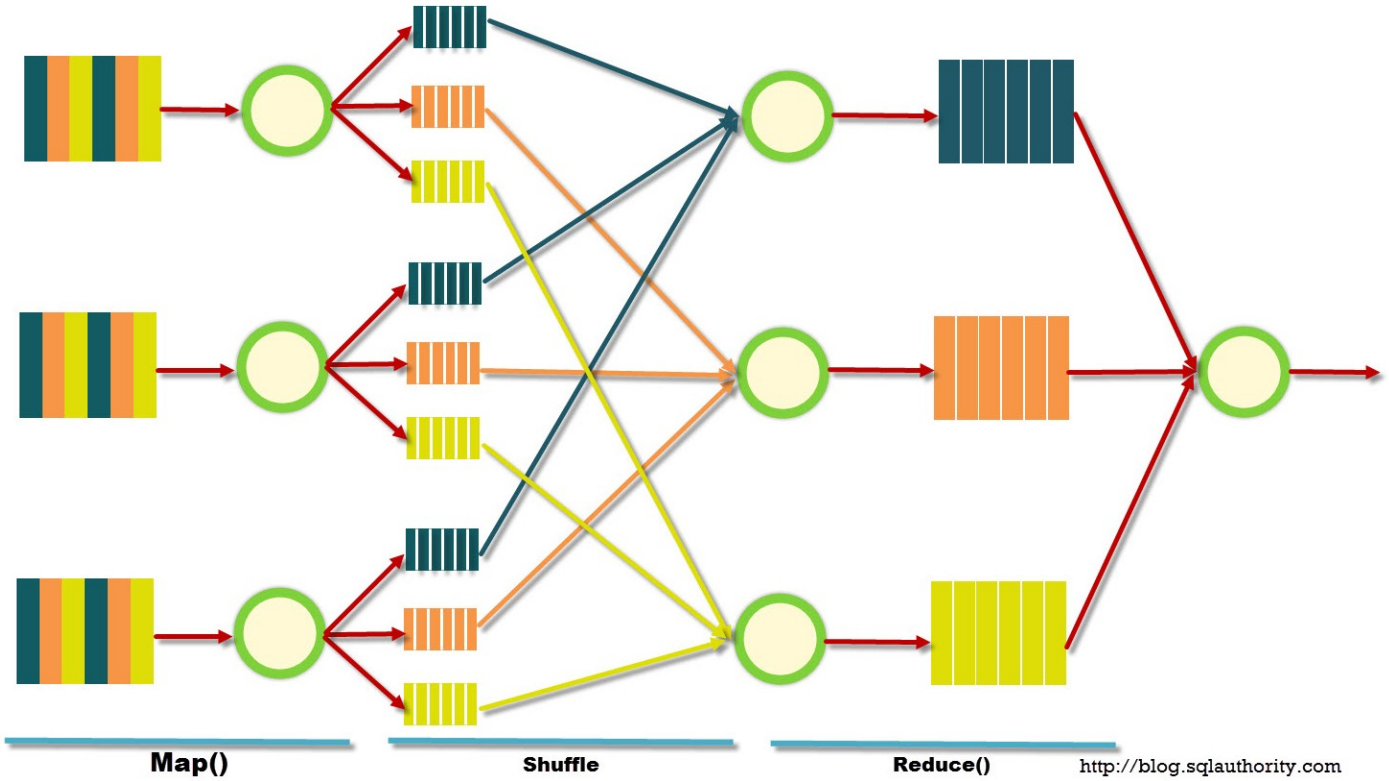
Mas afinal, o que é Big Data?



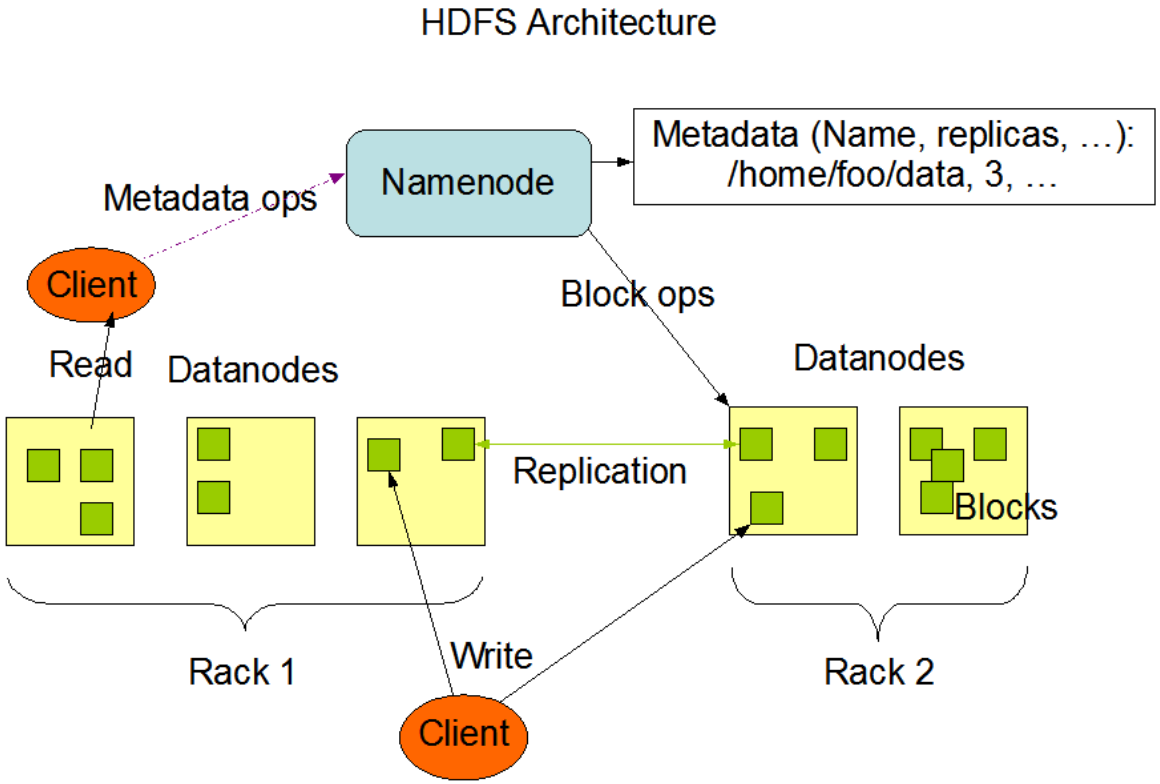
No princípio era o MapReduce…



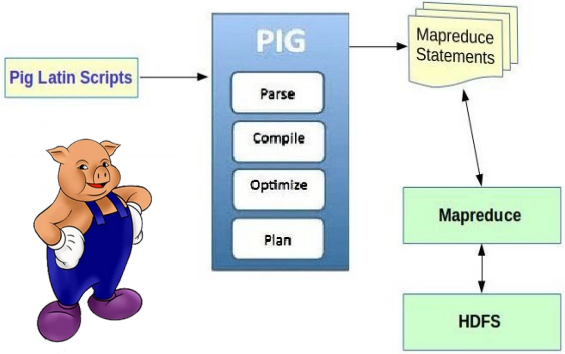
Hadoop MapReduce

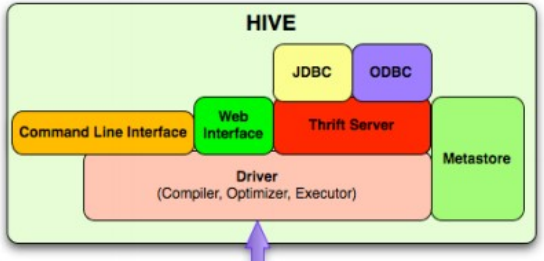


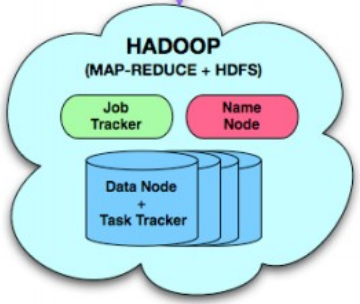
Hadoop MapReduce

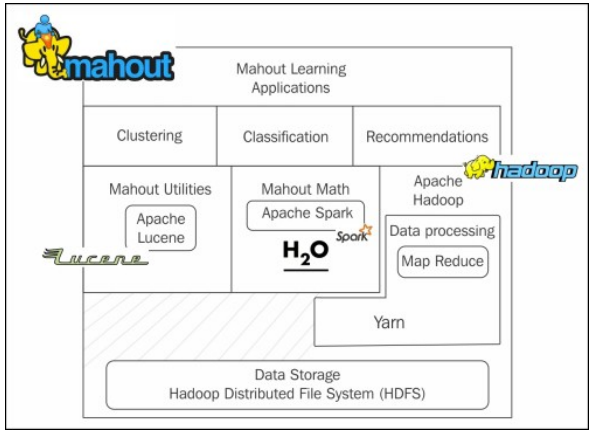


Outras ferramentas além do hadoop

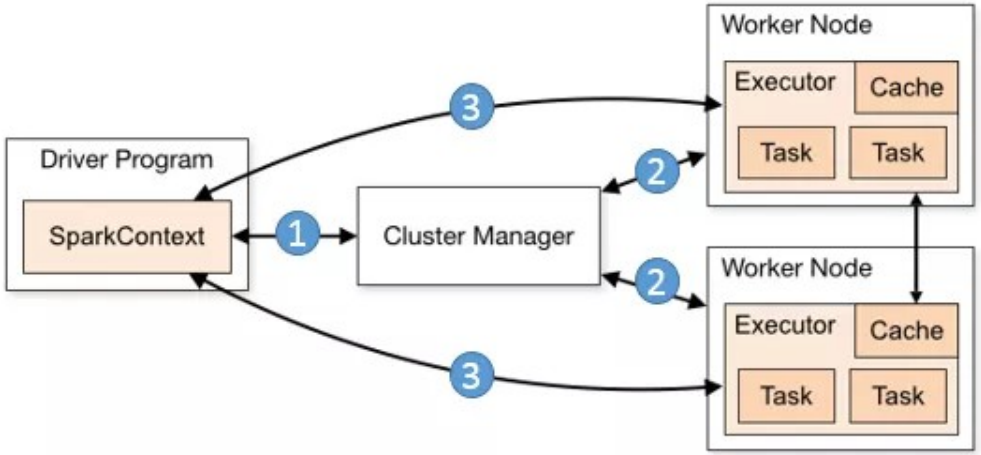


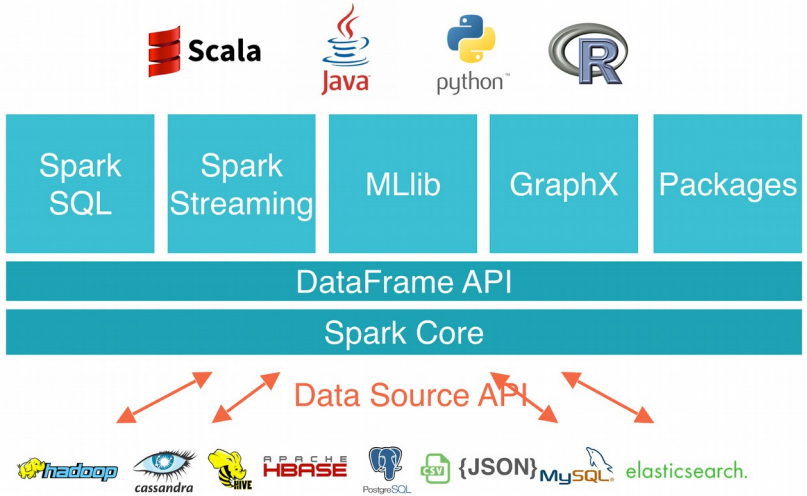






Spark: além do MapReduce



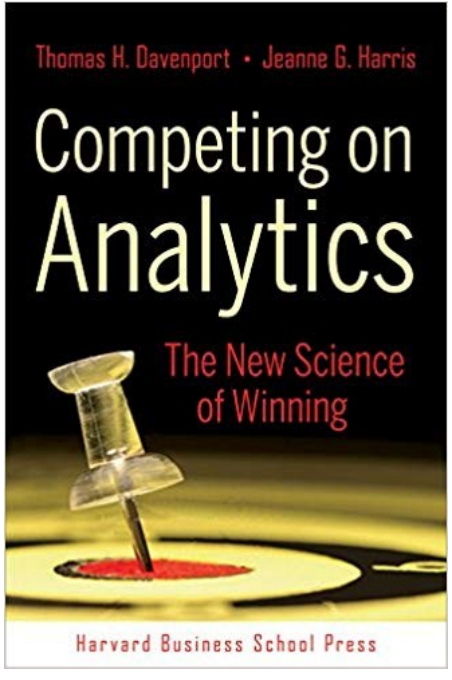


Spark: além do MapReduce

Transformations	Actions
map(func)	take(N)
flatMap(func)	count()
filter(func)	collect()
groupByKey()	reduce(func)
reduceByKey(func)	takeOrdered(N)
mapValues(func)	top(N)
…	…

Roteiro

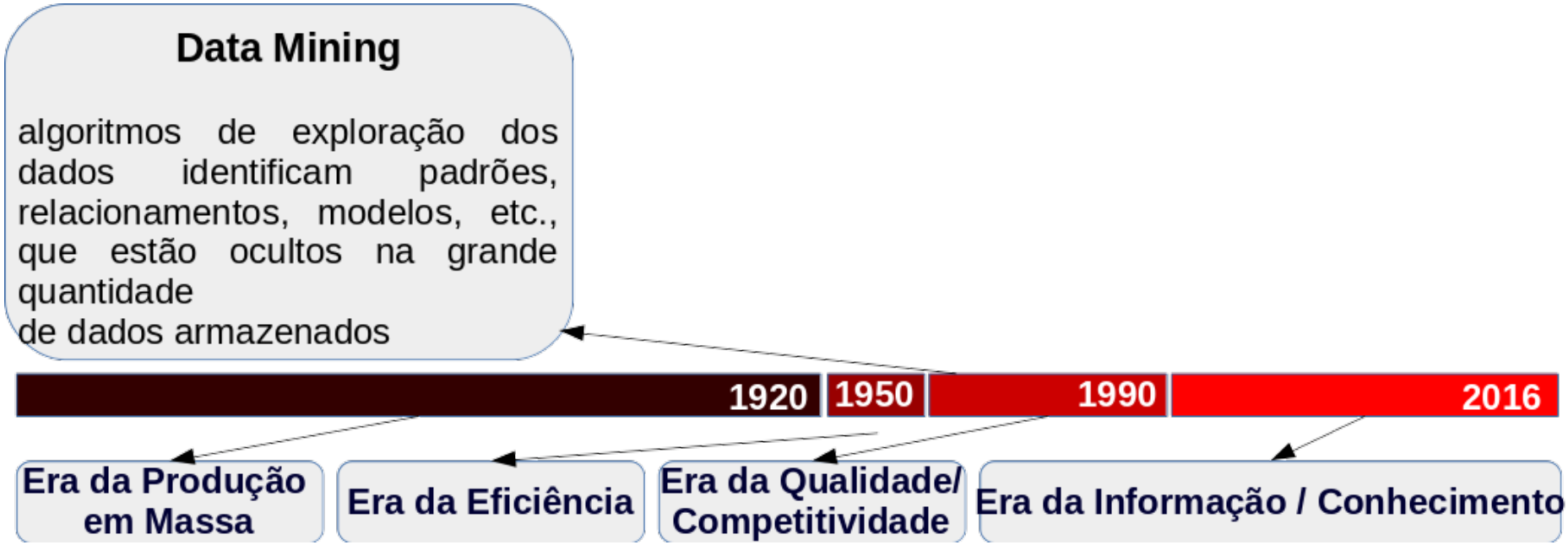
- Introdução: a arquitetura básica
- Processamento paralelo
- Sistemas distribuídos
- Big Data
- Ferramentas de Big Data
  - Hadoop MapReduce
  - Spark
- **Big data analytics**



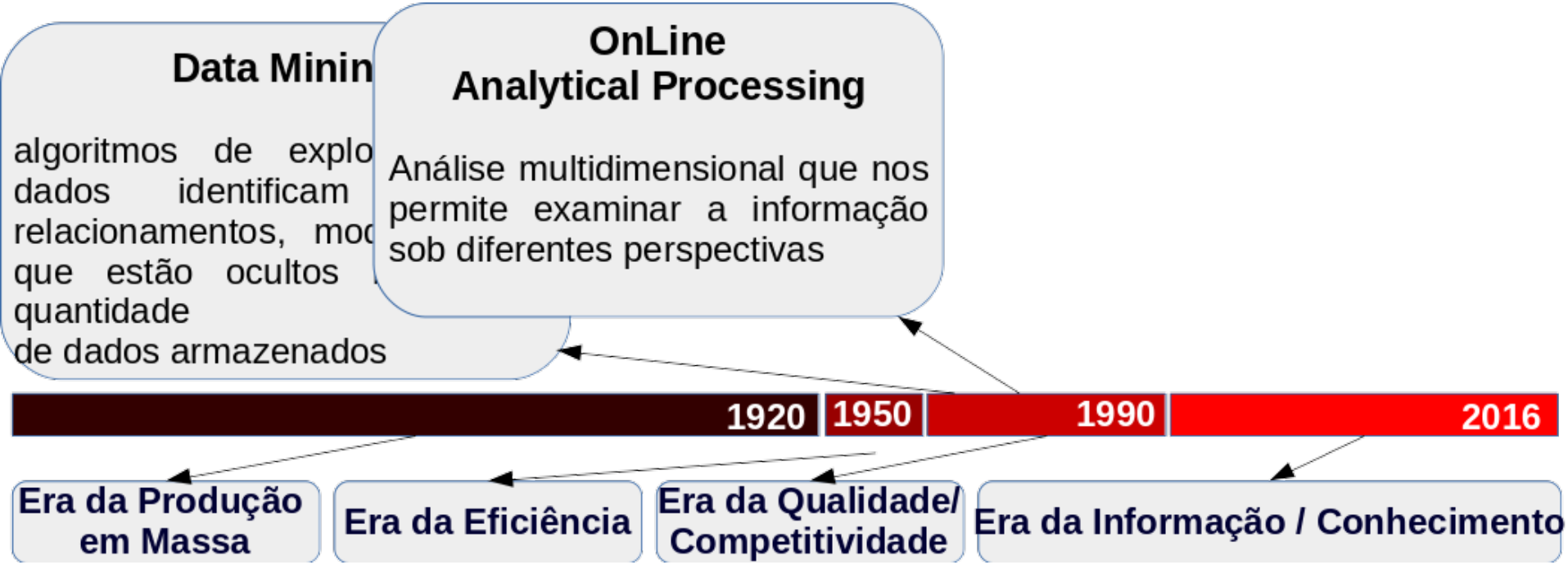
Big data analytics: ondas e eras

Era da Produção em Massa	Era da Eficiência	Era da Qualidade/ Competitividade	Era da Informação / Conhecimento
Atender o mercado quanto a quantidade	Produzir e vender	Satisfazer / Superar Expectativas	Tecnologia como fator de apoio à inovação
O homem é uma Extensão da máquina	Controlar e avaliar empregados	Recursos / Talentos Humanos são chaves	Obtenção/Tratamento/Uso/Análise dos dados para obter vantagens

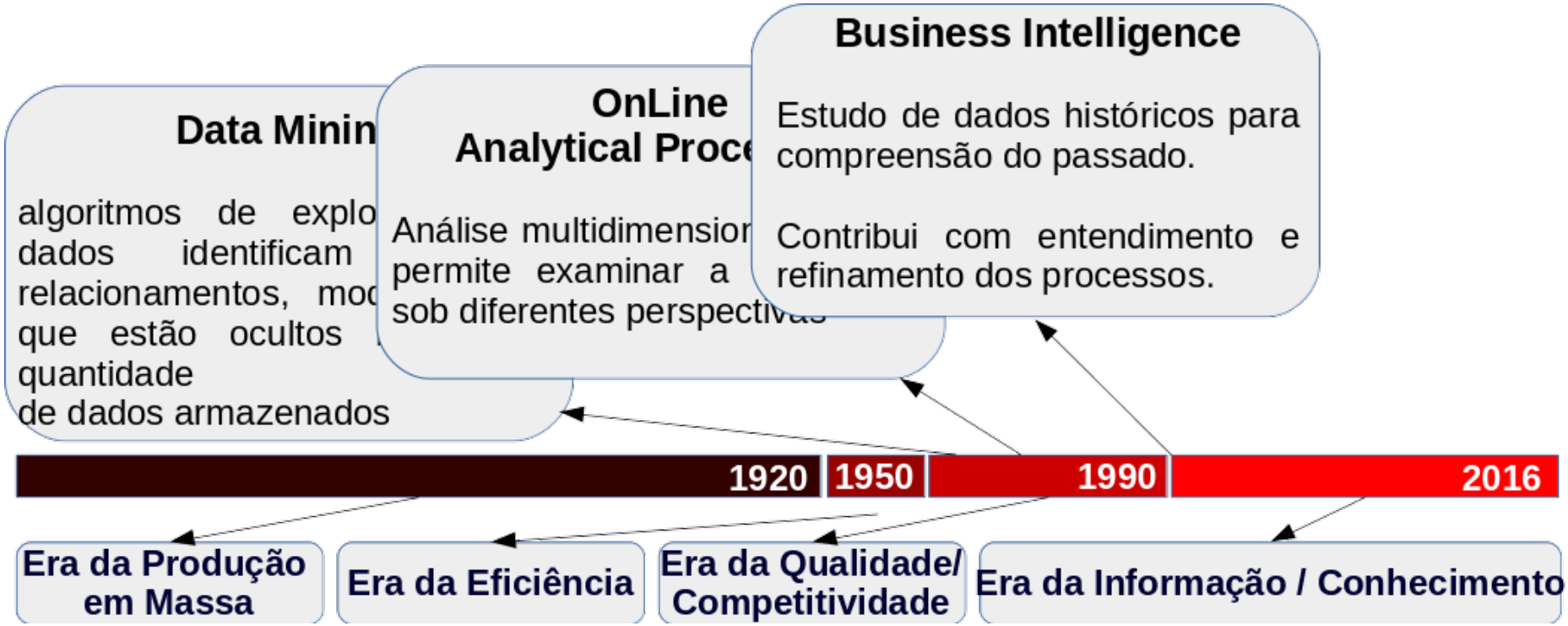
Big data analytics: informação x tempo



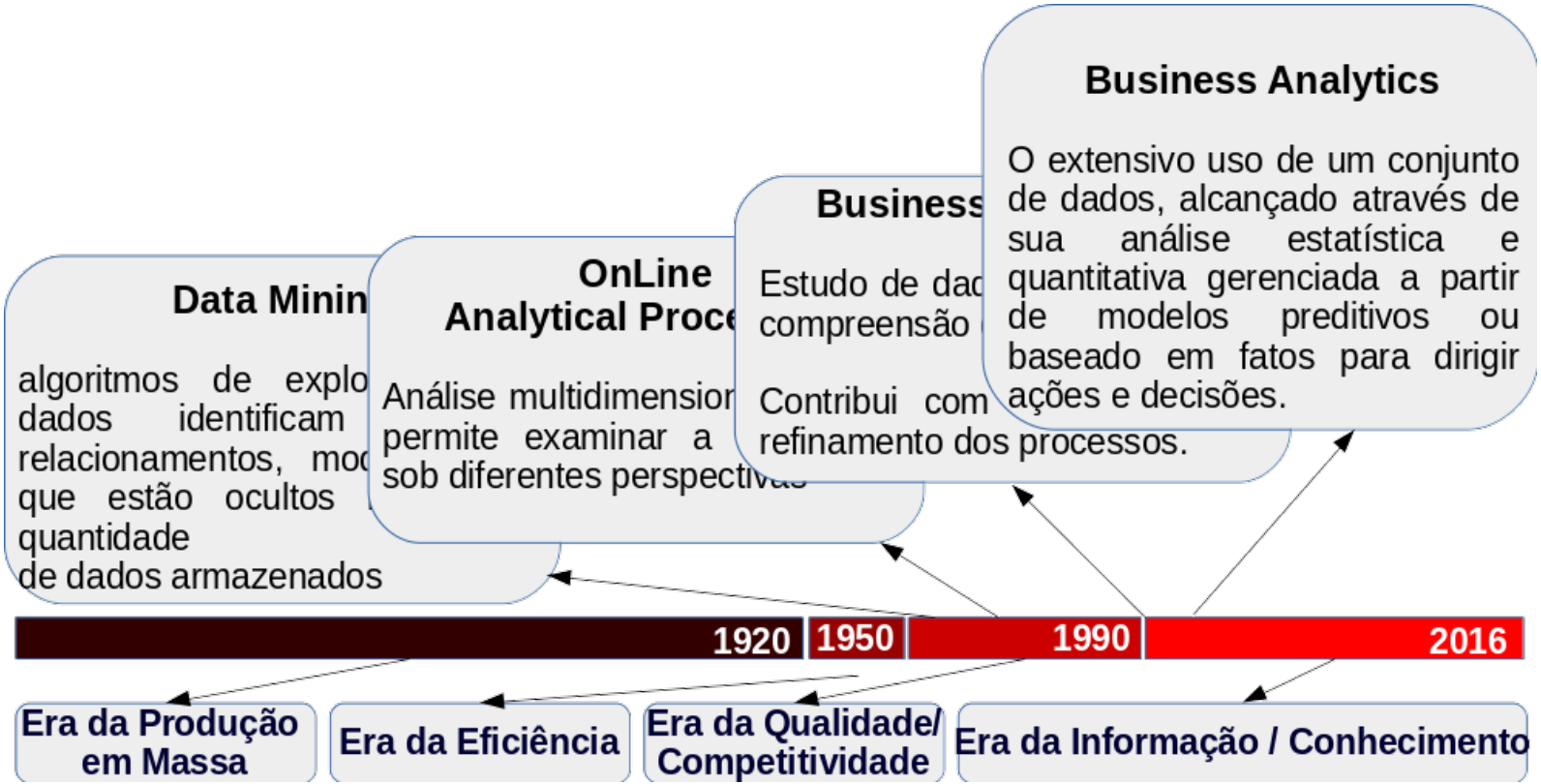
Big data analytics: informação x tempo



Big data analytics: informação x tempo



Big data analytics: informação x tempo



Obrigado. :)