Mestrado em Engenharia Informática 2022

# FUNDAMENTOS DE *BIG DATA*

FUNDAMENTOS DE *BIG DATA*
COM *PYSPARK*

Regina Sousa

Daniela Oliveira

# Objetivos da aula

- Introdução ao conceito de *Big Data*;
- Apresentação do *dataset* original;
- Download e análise do *dataset* original fornecido pelos docentes;
- Definição do *use case* para o trabalho de grupo a realizar;
- Procura de 2 ou mais *datasets* complementares de acordo com o *use case* definido;
- Pesquisa de ferramentas de *Big Data*;
- Análise do estado da arte para o desenvolvimento do 1º trabalho individual (facultativo).

# BIG DATA

*Big data* é o termo utilizado para o estudo e desenvolvimento de aplicação com dados que, por alguma razão requerem processamento complexo, para os métodos de processamento tradicional. Existem 3 V's que são usados para descrever as características de Big Data.

**Volume**

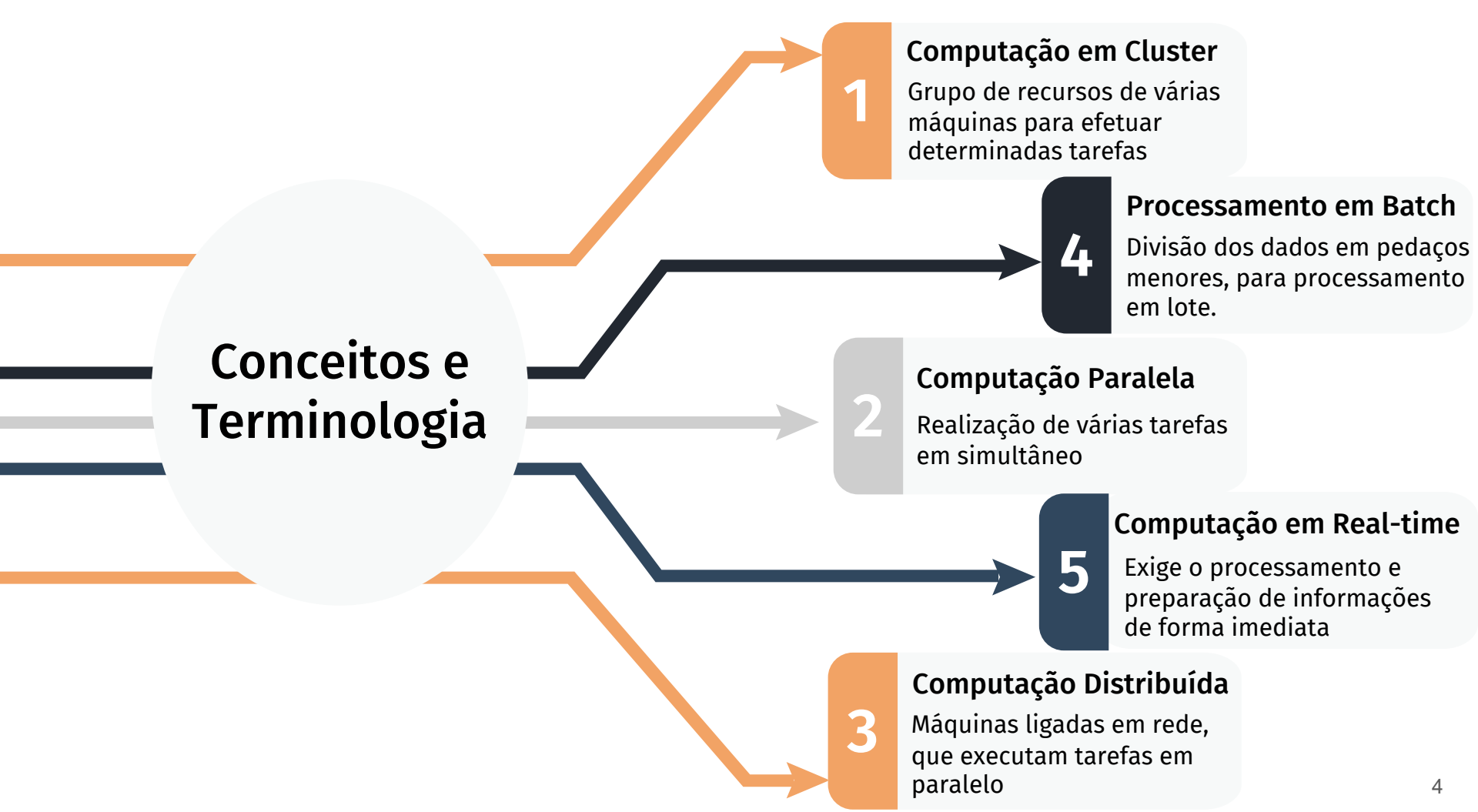
Tamanho do Conjunto de Dados (poderá ter um valor variável e até crescente.)

**Velocidade**

Velocidade da geração de dados bem como disponibilização para processamento.

**Variedade**

Variedade na quantidade de fontes de dados e formato em que os dados vêm.


Conceitos e Terminologia
1
Computação em Cluster
Grupo de recursos de várias máquinas para efetuar determinadas tarefas
4
Processamento em Batch
Divisão dos dados em pedaços menores, para processamento em lote.
2
Computação Paralela
Realização de várias tarefas em simultâneo
5
Computação em Real-time
Exige o processamento e preparação de informações de forma imediata
3
Computação Distribuída
Máquinas ligadas em rede, que executam tarefas em paralelo

# Big Data Processing Systems

**Hadoop/MapReduce**

Open Source

Scalable & Fault Tolerant

Batch Processing

**Apache Spark**

Cluster Computing

Open Source

Batch & Real-Time

# Conjunto de Dados

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | **Date_reported** | **Country_code** | **Country** | **WHO_region** | **New_cases** | **Cumulative_cases** | **New_deaths** | **Cumulative_deaths** |
| 2 | 03/01/20 | AF | Afghanistan | EMRO | 0 | 0 | 0 | 0 |
| 3 | 04/01/20 | AF | Afghanistan | EMRO | 0 | 0 | 0 | 0 |
| 4 | 05/01/20 | AF | Afghanistan | EMRO | 0 | 0 | 0 | 0 |
| 5 | 06/01/20 | AF | Afghanistan | EMRO | 0 | 0 | 0 | 0 |
| 6 | 07/01/20 | AF | Afghanistan | EMRO | 0 | 0 | 0 | 0 |
| 7 | 08/01/20 | AF | Afghanistan | EMRO | 0 | 0 | 0 | 0 |
| 8 | 09/01/20 | AF | Afghanistan | EMRO | 0 | 0 | 0 | 0 |

| Field name | Type | Description |
|---|---|---|
| Date_reported | Date | Date of reporting to WHO |
| Country_code | String | ISO Alpha-2 country code |
| Country | String | Country, territory, area |
| WHO_region | String | WHO regional offices: WHO Member States are grouped into six WHO regions -- Regional Office for Africa (AFRO), Regional Office for the Americas (AMRO), Regional Office for South-East Asia (SEARO), Regional Office for Europe (EURO), Regional Office for the Eastern Mediterranean (EMRO), and Regional Office for the Western Pacific (WPRO). |
| New_cases | Integer | New confirmed cases. Calculated by subtracting previous cumulative case count from current cumulative cases count.* |
| Cumulative_cases | Integer | Cumulative confirmed cases reported to WHO to date. |
| New_deaths | Integer | New confirmed deaths. Calculated by subtracting previous cumulative deaths from current cumulative deaths.* |
| Cumulative_deaths | Integer | Cumulative confirmed deaths reported to WHO to date. |

**Download link:** https://covid19.who.int/WHO-COVID-19-global-data.csv