Big Data – Hands-On

Objetivo: O objetivo deste hands-on é pesquisar sobre repositórios de dados para aplicações de Big Data.

1. Repositórios de dados

Atualmente, existem diversos repositórios com dados abertos para download e utilização.

- AWS Public Dataset Program - https://aws.amazon.com/opendata/public-datasets/
- Google Cloud Pubilc Datasets - https://cloud.google.com/public-datasets/
- Registry of Research Data Repositories - https://www.re3data.org/
- Wikipedia downloadable dataset – http://en.wikipedia.org/wiki/Wikipedia:Database_download
- Million Song Dataset - https://labrosa.ee.columbia.edu/millionsong/
- U.S. Government's open data - https://www.data.gov/
- Open Data Brazilian Portal - http://dados.gov.br/dataset
- https://github.com/awesomedata/awesome-public-datasets
- Kaggle Datasets - https://www.kaggle.com/datasets
- DataHub - https://datahub.io/
- NYC Taxi & Limousine Commission Trip Record Data - http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
- YFCC100M - https://multimediacommons.wordpress.com/yfcc100m-core-dataset/

Acesse os repositórios de dados acima e escolha 3 datasets. Para cada dataset, apresente as seguintes informações:

- Nome do dataset
- Descrição do conteúdo do dataset
- Área (governamental, meteorologia, transporte, saúde, etc)
- Estrutura de organização dos dados
- Tamanho do dataset
- URL para download
- Dois exemplos de aplicação usando o dataset