Big Data e Computação em Nuvem

Insper - Pós-Graduação em Data Science e Decisão - 2024.4

Professor:	Michel Fornaciali	**Prof.ª Assistente:**	Thanuci Silva
E-mail:	michelsf@insper.edu.br	**E-mail:**	thanucis@insper.edu.br

Competências: Desenvolver sistemas para o processamento efetivo de dados em larga escala, considerando estruturas de dados e algoritmos apropriados, ambientes e frameworks necessários ao processamento em larga escala.

Habilidades:

- 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, Spark e arquiteturas em nuvem
- Planejar, estruturar e utilizar bancos de dados estruturados SQL;

Material do curso:

O material do curso está disponível no *Blackboard*.

Principais bibliografias:

- Data Analytics with Spark using Python. Aven, J. 2018. Addison-Wesley.
- Barroso, Luiz André, Jimmy Clidaras, and Urs Holzle. *The datacenter as a computer: An introduction to the design of warehouse-scale machines*. Synthesis lectures on computer architecture 8.3 (2013). Disponível em https://www.morganclaypool.com/doi/abs/10.2200/S00516ED2V01Y201306CAC024

Programa da disciplina:

- Big Data e computação em nuvem.
- Big Data e fundamentos de processamento e distribuído.
- Fundamentos do Dask e Apache Spark.
- Estruturas de dados do Apache Spark.
- Programação PySpark para Apache Spark e Spark SQL.
- Arquitetura, elementos e gestão de serviços em nuvem.
- Sistemas de arquivos e bancos de dados para Big Data em nuvem.
- Utilizando bancos de dados estruturados em nuvem.
- Machine learning utilizando Spark MLib.
- Projeto de computação em larga escala em nuvem.

Notas: Listas de exercícios (30%), Checkpoint (20%), Projeto Final (50%).

- Lista 1: início (25/10); fim (25/11)
- Lista 2: início (06/11); fim (25/11)
- Checkpoint em 23/11 – projeto final
- Apresentação do projeto final: 09/dez

Previsão de conteúdo das aulas:

- **Aula 01 (19/10)**: Conceitos introdutórios + programação funcional + Dask
- **Aula 02 (21/10)**: MapReduce e Introdução ao PySpark
- **Aula 03 (25/10)**: Práticas em PySpark
- **Aula 04 (30/10)**: Novas transformações e ações, otimizações de código em PySpark
- **Aula 05 (06/11)**: Banco de Dados (teoria e prática)
- **Aula 06 (08/11)**: DataFrames PySpark
- **Aula 07 (09/11)**: DataFrames PySpark em dados reais
- **Aula 08 (11/11)**: Union / Cache
- **Aula 09 (18/11)**: SparkML
- **Aula 10 (23/11)**: Práticas em SparkML
- **Aula 11 (25/11)**: Cloud
- **Aula 12 (04/12)**: Tempo reservado para projeto
- **Aula 13 (09/12)**: Apresentações do projeto

Bons estudos!