



CURSO DE CIÊNCIA DE DADOS APLICADA AO PODER JUDICIÁRIO

SPARK PARA CIÊNCIA DE DADOS

Apresentação do curso

PROF. CARLOS M. D. VIEGAS

Apresentação pessoal

- **Prof. Carlos M. D. Viegas**

- Professor Adjunto

- Departamento de Engenharia de Computação e Automação (DCA)
- Universidade Federal do Rio Grande do Norte (UFRN)

- Formação:

- Doutor em Engenharia Informática (2015)
 - Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
- Mestre em Engenharia Elétrica e de Computação (2009)
 - Universidade Federal do Rio Grande do Norte, Natal/RN, Brasil
- Engenheiro de Computação (2006)
 - Universidade Federal do Rio Grande do Norte, Natal/RN, Brasil

- Áreas de interesse/experiência:

- Redes de Computadores, Segurança da Informação, **Engenharia de Dados** e Sistemas de Comunicação de Tempo Real



Spark para Ciência de Dados

• Caracterização do Curso

- Modalidade
 - Ensino à distância
- Carga horária total
 - 36 horas
- Carga horária semanal
 - 4 horas

- Duração do curso
 - 9 semanas
- Início da oferta
 - Março/2023 ➡ 10/03/2023
- Fim da oferta
 - Maio/2023 ➡ 25/05/2023



Spark para Ciência de Dados

- **Objetivos**

- Capacitar o(a) cursista a utilizar as soluções **Apache Hadoop** e **Apache Spark** para o desenvolvimento de aplicações para resolução de problemas na área da Ciência de Dados
- Ao final do curso, o(a) cursista terá como habilidades:
 - Capacidade para planejar e preparar a infraestrutura de dados de uma organização
 - Conhecimento das técnicas e ferramentas para o desenvolvimento de aplicações com Apache Spark para o processamento de dados em larga escala



Spark para Ciência de Dados

• Programa/Ementa

- Apresentação do Ecossistema Apache Hadoop
- Instalação e configuração do ambiente Apache Hadoop
- Estudo do sistema de arquivos HDFS (*Hadoop Distributed File System*) e do modelo de programação MapReduce
- Criação de cluster para processamento de dados
- Gerenciamento de recursos e escalonamento de tarefas com YARN
- Desenvolvimento de aplicações com MapReduce em linguagem Python
- Integração do Ecossistema Hadoop com módulos adicionais: bancos de dados e outras fontes de dados
- Introdução ao Apache Spark
- Instalação, configuração e integração do ambiente Apache Spark
- Abstrações de dados RDD (*Resilient Distributed Dataset*), Dataframe e Datasets
- Comparação Spark vs Hadoop
- Desenvolvimento de Aplicações com pySpark
- SparkSQL
- Estudo das bibliotecas Spark MLlib, Spark GraphX e aplicações práticas
- Noções de SparkR

Spark para Ciência de Dados

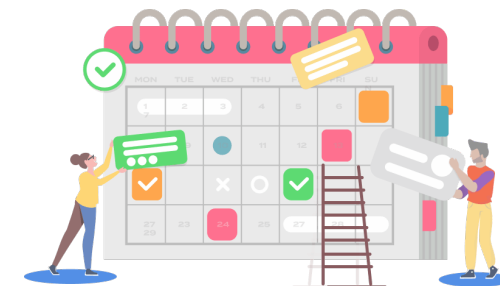
- Conteúdo programático

- Semanas 1 e 2: Apache Hadoop

- Introdução ao Ecossistema Hadoop
 - Sistema de arquivos HDFS
 - Modelo de programação MapReduce
 - Gerenciamento de recursos com Yarn
 - Instalação e configuração do Hadoop (standalone e multi-node)
 - Análise de logs para diagnóstico e resolução de problemas
 - Desenvolvimento de aplicações com MapReduce
 - Execução e monitoramento de tarefas
 - Visão geral de módulos adicionais Hadoop:
 - Hive, Hbase, Sqoop e Mahout



10/03/2023
a
23/03/2023



Spark para Ciência de Dados

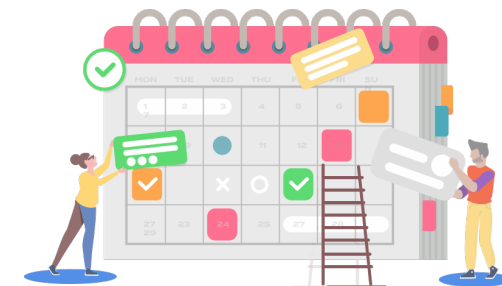
- Conteúdo programático

- Semanas 3 e 4: Apache Spark

- Introdução ao Ecossistema Spark
 - Comparação Spark vs Hadoop
 - Abstração de dados: RDD, Dataframe e Dataset
 - Instalação e configuração do Spark (standalone e cluster)
 - Integração com Apache Hadoop
 - Programação com pySpark: RDD
 - Spark Web UI: Interface de usuário e DAG
 - Programação com pySpark: Dataframe e Dataset
 - API Pandas no Spark
 - Interação com fontes de dados e Aplicações práticas



23/03/2023
a
13/04/2023*



Spark para Ciência de Dados

- **Conteúdo programático**

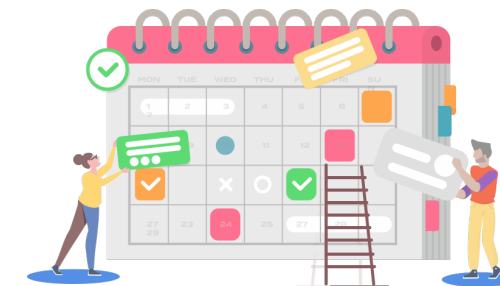
- Semana 5: Apache Spark – SparkSQL
 - Programação com pySpark: SparkSQL
 - Manipulação de dados com SparkSQL
 - Aplicações práticas
- Semana 6: Apache Spark – MLlib
 - Fundamentos de Machine Learning
 - Machine Learning no Spark
 - Criação de pipelines com Machine Learning
 - Aplicações práticas



14/04/2023
a
27/04/2023*



28/04/2023
a
04/05/2023



Spark para Ciência de Dados

- Conteúdo programático

- Semana 7: Apache Spark – Streaming

- Modelo de programação Spark Structured Streaming
 - Criação de Streams com Dataframe e Dataset
 - Operações sobre Streams de dados
 - Aplicações práticas



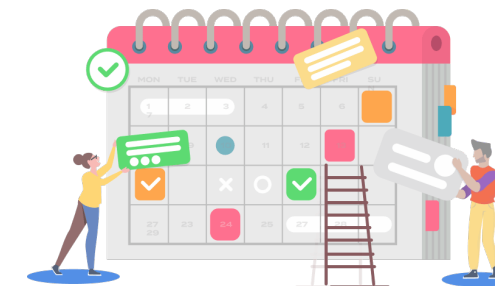
05/05/2023
a
11/05/2023

- Semana 8: Apache Spark – Spark R

- Introdução ao Spark R
 - Exemplos de programação em R para Spark



12/05/2023
a
18/05/2023



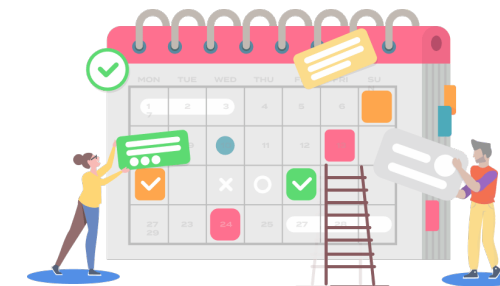
Spark para Ciência de Dados

- Conteúdo programático

- Semana 9: Apache Spark – GraphX
 - Fundamentos de grafos
 - Análise de grafos com GraphX e GraphFrames



19/05/2023
a
25/05/2023



Spark para Ciência de Dados

- Metodologia de ensino

- Aulas expositivas (online)
 - Encontros síncronos às **sextas-feiras** às 08h30
 - Plantões de dúvidas (online)
 - Encontros síncronos às **quintas-feiras** às 08h30
 - Práticas interativas
 - Exercícios complementares
 - Trabalhos de implementação
-
- A plataforma utilizada para acompanhamento, comunicação e divulgação dos materiais do curso será o ambiente de aprendizado **Moodle** (do CEAJUD)



Spark para Ciência de Dados

- **Atividades previstas para os(as) cursistas**

- Os(as) cursistas realizarão atividades práticas de implantação, configuração e programação, focando na resolução de problemas similares ou correlatos aos do Poder Judiciário
 - Os(as) cursistas serão expostos a problemas e utilizarão o conhecimento adquirido durante as aulas para a resolução dos mesmos
- A cada semana os(as) cursistas deverão desenvolver as seguintes atividades:
 1. Estudar o material pré-aula como forma de preparo para a aula
 2. Assistir às aulas programadas para a semana no horário definido
 3. Trabalhar nos exercícios disponibilizados pelo professor
 4. Realizar o estudo individual dos materiais indicados, tais como leituras complementares, resolução de exercícios e acesso a vídeos adicionais
 5. Participar do fórum do curso contribuindo com tópicos para a discussão ou respondendo e complementando tópicos em aberto relacionados ao conteúdo apresentado (opcional)
 6. Realizar as tarefas de avaliação semanal, respondendo aos questionários aplicados



Spark para Ciência de Dados

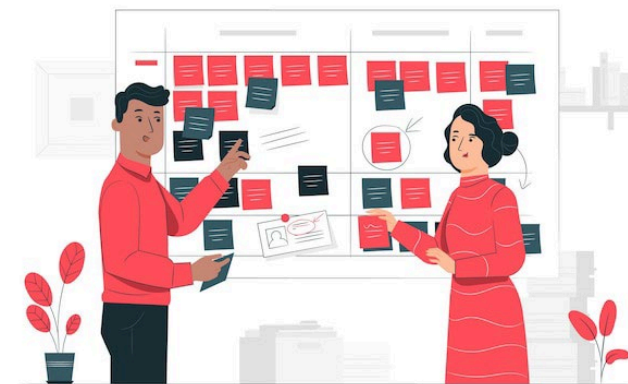
- Carga horária do(a) cursista

- 4 horas/semana

- Estudo individual de conteúdo pré-aula 00:30
 - Participação na aula ao vivo 01:30
 - Estudo individual pós aula 00:30
 - Participação no plantão de dúvidas 01:00
 - Participação no fórum 00:30

- Pré-requisitos

- Conhecimentos em sistemas GNU/Linux ou Unix
 - Conhecimentos em linguagem de programação Python
 - Computador pessoal com pelo menos 8 GB de memória RAM e 40 GB de espaço em disco



Spark para Ciência de Dados

- Avaliação dos(as) cursistas

- Avaliação de Desempenho

- Questionários semanais: questões teóricas e práticas
 - Média aritmética simples das notas obtidas nas tarefas semanais

- Avaliação de Participação

- A frequência de participação será determinada pela realização das tarefas semanais de avaliação do curso e/ou questões do material pré-aula/pós-aula
 - Receberão certificados de participação aqueles que obtiverem aproveitamento igual ou superior a 70% nessas atividades

- Avaliação de Reação

- Ao final do curso será aplicada uma avaliação de reação com o intuito de avaliar a percepção dos(as) cursistas quanto ao curso realizado no alcance dos objetivos



OBRIGADO

CONTATO: viegas@dca.ufrn.br

CURSO DE CIÊNCIA DE DADOS
APLICADA AO PODER JUDICIÁRIO



CNJ CONSELHO
NACIONAL
DE JUSTIÇA