

# CURSO DE SPARK

## FICHA DO CURSO

### 1. Resumo do curso

**Curso:** Curso de Spark – Distribuição e Processamento de Dados

**Modalidade:** Ensino à distância

**Duração do curso:** 9 semanas

**Carga horária total:** 36 horas

**Carga horária semanal:** 4 horas

**Carga horária síncrona:** 2,5 horas por semana

**Carga horária assíncrona:** 1,5 horas por semana

**Início da oferta:** Março/2023

**Fim da oferta:** Maio/2023

**Pré-requisitos:** Conhecimentos do sistema operacional GNU/Linux ou Unix (noções de utilização de terminal shell, de variáveis de ambiente de sistema, de comunicação em rede com utilização de ssh) e conhecimentos em linguagem Python.

### 2. 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. Realizar tarefas de implantação, configuração, e integração de dados em *cluster*.

Ao final do curso, o(a) cursista terá como habilidades principais a capacidade para planejar e preparar a infraestrutura de dados de uma organização, podendo projetar, construir, integrar e manter banco de dados ou outras fontes de dados, bem como conhecer as técnicas e ferramentas para o desenvolvimento de aplicações com Apache Spark para o processamento de dados em larga escala.

### 3. 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; e Noções de SparkR.

### 4. Procedimentos metodológicos e didáticos

Os procedimentos metodológicos que serão adotados no curso envolvem aulas expositivas (online), plantões de dúvidas (online), práticas interativas, exercícios complementares e trabalhos de implementação.

As aulas expositivas estão programadas para ocorrerem às sextas-feiras de cada semana a partir das 08h30, e os plantões de dúvidas ocorrerão às quintas-feiras de cada semana também a partir das 08h30. As aulas expositivas e plantões de dúvidas ocorrerão de forma síncrona por meio da ferramenta Webex. O detalhamento de conteúdo e o respectivo período de aulas e plantões de dúvidas estão indicados na Seção 7.

A plataforma utilizada para acompanhamento, comunicação e divulgação dos materiais do curso será o ambiente de aprendizado Moodle do Centro de Formação e Aperfeiçoamento de Servidores do Poder Judiciário (CEAJUD).

As atividades práticas que serão desenvolvidas pelos(as) cursistas envolvem a configuração do sistema operacional GNU/Linux para a instalação e (também) configuração das ferramentas Apache Hadoop, Apache Spark e módulos adicionais. Os(as) cursistas, no decorrer do curso, irão realizar atividades práticas de implantação, configuração e programação, focando na resolução de problemas similares ou correlatos aos do Poder Judiciário. A cada semana, os(as) cursistas serão expostos a problemas e utilizarão o conhecimento adquirido durante as aulas para a resolução dos mesmos.

De maneira resumida, a cada semana (em um total de 9 semanas) 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.

## 5. Requisitos técnicos

Para a realização das atividades práticas de implantação, configuração e programação do Apache Hadoop e Apache Spark no decorrer do curso, é necessário que cada cursista possua um computador pessoal com uma configuração mínima de 8 GB de RAM e 40 GB disponível em disco.

O curso será direcionado à implantação das soluções Apache Hadoop e Apache Spark em ambiente GNU/Linux. Neste sentido, é requisitado que o sistema GNU/Linux esteja instalado nativamente no computador ou que se recorra ao uso de soluções de virtualização que permitam a instalação do referido sistema operacional.

## 6. Carga horária do(a) cursista

A Tabela 1 a seguir apresenta distribuição da carga horária semanal total do(a) cursista, de 4 horas por semana, nas atividades semanais previstas na vigência do curso.

Tabela 1 – Carga horária semanal do(a) cursista.

| Atividade semanal                      | Carga horária (horas) | Fração (%) |
|----------------------------------------|-----------------------|------------|
| Estudo individual de conteúdo pré-aula | 00:30                 | 12,5       |
| Participação na aula ao vivo           | 01:30                 | 37,5       |
| Estudo individual pós aula             | 00:30                 | 12,5       |
| Participação no plantão de dúvidas     | 01:00                 | 25,0       |
| Participação no fórum                  | 00:30                 | 12,5       |
| Total                                  | 04:00                 | 100        |

## 7. Conteúdo programático

Tabela 2 – Conteúdo programático previsto e respectivo período.

| Unidade | Conteúdo                         | Período                    |
|---------|----------------------------------|----------------------------|
| 1       | <b>Apache Hadoop</b>             | De 10/03 a 16/03/2023:     |
|         | Introdução ao Ecossistema Hadoop | - 10/03 aula síncrona      |
|         | Sistema de arquivos HDFS         | - 16/03 plantão de dúvidas |

|              |                                                                                                                                                                                                                                                                   |                                                                                                             |
|--------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|
|              | Modelo de programação MapReduce<br>Gerenciamento de recursos com Yarn<br>Instalação e configuração do Hadoop (standalone e multi-node)<br>Análise de logs para diagnóstico e resolução de problemas                                                               |                                                                                                             |
| 2            | <b>Apache Hadoop</b><br>Desenvolvimento de aplicações com MapReduce<br>Execução e monitoramento de tarefas<br>Visão geral de módulos adicionais Hadoop: Hive, Hbase, Sqoop e Mahout                                                                               | De 17/03 a 23/03/2023:<br>- 17/03 aula síncrona<br>- 23/03 plantão de dúvidas                               |
| 3            | <b>Apache Spark</b><br>Introdução ao Ecossistema Spark<br>Comparação Spark vs Hadoop<br>Abstração de dados: RDD, Dataframe e Dataset<br>Instalação e configuração do Spark (standalone e cluster)<br>Integração com Apache Hadoop<br>Programação com pySpark: RDD | De 24/03 a 30/03/2023:<br>- 24/03 aula síncrona<br>- 30/03 plantão de dúvidas                               |
| 4            | <b>Apache Spark</b><br>Spark Web UI: Interface de usuário e DAG<br>Programação com pySpark: Dataframe e Dataset<br>API Pandas no Spark<br>Interação com fontes de dados e Aplicações práticas                                                                     | De 31/03 a 13/04/2023:<br>- 31/03 aula síncrona<br>- 06/04 plantão de dúvidas<br>- 13/04 plantão de dúvidas |
| 5            | <b>Apache Spark - SparkSQL</b><br>Programação com pySpark: SparkSQL<br>Manipulação de dados com SparkSQL<br>Aplicações práticas                                                                                                                                   | De 14/04 a 27/04/2023:<br>- 14/04 aula síncrona<br>- 20/04 plantão de dúvidas<br>- 27/04 plantão de dúvidas |
| 6            | <b>Apache Spark – MLlib</b><br>Fundamentos de Machine Learning<br>Machine Learning no Spark<br>Criação de pipelines com Machine Learning<br>Aplicações práticas                                                                                                   | De 28/04 a 04/05/2023:<br>- 28/04 aula síncrona<br>- 04/05 plantão de dúvidas                               |
| 7            | <b>Apache Spark - Streaming</b><br>Modelo de programação Spark Structured Streaming<br>Criação de Streams com Dataframe e Dataset<br>Operações sobre Streams de dados<br>Aplicações práticas                                                                      | De 05/05 a 11/05/2023:<br>- 05/05 aula síncrona<br>- 11/05 plantão de dúvidas                               |
| 8            | <b>Apache Spark - Spark R</b><br>Introdução ao Spark R<br>Exemplos de programação em R para Spark                                                                                                                                                                 | De 12/05 a 18/05/2023:<br>- 12/05 aula síncrona<br>- 18/05 plantão de dúvidas                               |
| 9            | <b>Apache Spark - GraphX</b><br>Fundamentos de grafos<br>Análise de grafos com GraphX e GraphFrames                                                                                                                                                               | De 19/05 a 25/05/2023:<br>- 19/05 aula síncrona<br>- 25/05 plantão de dúvidas                               |
| Encerramento |                                                                                                                                                                                                                                                                   |                                                                                                             |

## 8. Avaliação dos(as) cursistas no curso

### 8.1 Avaliação de Desempenho

Os(as) cursistas serão avaliados por meio de questionários semanais. Algumas questões serão teóricas e outras de cunho prático, cuja resposta virá da realização de determinada atividade prática.

O desempenho no curso será determinado pela média aritmética simples das notas obtidas nas tarefas semanais aplicadas aos/as cursistas.

## 8.2 Avaliação de Participação

A frequência de participação no curso será determinada pela realização das tarefas semanais de avaliação do curso ou questões do material pré-aula/pós-aula. Receberão certificados de participação aqueles que obtiverem aproveitamento igual ou superior a 70% nessas atividades.

## 8.3 Avaliação de Reação

Ao final do curso será aplicada uma avaliação de reação, por meio da qual os(as) cursistas responderão a um formulário de reação com questões relativas ao curso, seu conteúdo e didática do professor, com o intuito de avaliar a percepção dos(as) cursistas quanto ao curso realizado no alcance dos objetivos.

## 9. Referências Bibliográficas

- 1) Parsian, Mahmoud. **Data Algorithms with Spark: Recipes and Design Patterns for Scaling Up Using Pyspark**. Sebastopol, CA, USA: O'Reilly Media, 2022. Print.
- 2) Damji, Jules; Wenig, Brooke; Das, Tathagata; Lee, Denny. **Learning Spark: Lightning-Fast Data Analytics**. Sebastopol, CA, USA: O'Reilly Media, 2020. Print.
- 3) Chambers, Bill; Zaharia, Matei. **Spark: The Definitive Guide: Big Data Processing Made Simple**. Sebastopol, CA, USA: O'Reilly Media, 2018. Print.
- 4) White, Tom. **Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale**. Sebastopol, CA, USA: O'Reilly Media, 2015. Print.
- 5) Hamstra, Mark; Zaharia, Matei; Karau, Holden. **Learning Spark: Lightning-Fast Data Analysis**. Sebastopol, CA, USA: O'Reilly Media, 2015. Print.
- 6) **Documentação Apache Hadoop**. Disponível em: <<https://hadoop.apache.org/docs/current/>>
- 7) **Documentação Apache Spark**. Disponível em: <<https://spark.apache.org/docs/latest/>>
- 8) **Spark with Python (PySpark) Tutorial For Beginners**. Spark by Examples. Disponível em: <<https://sparkbyexamples.com/pyspark-tutorial/>>
- 9) **Documentação SparkR (R on Spark)**. Disponível em: <<https://spark.apache.org/docs/latest/sparkr.html>>
- 10) Luraschi, Javier; Kuo, Kevin; Ruiz, Edgar. **Mastering Spark with R: The Complete Guide to Large-Scale Analysis and Modeling**. Sebastopol, CA, USA: O'Reilly Media, 2019. Print.

## 10. Informações sobre o professor

O instrutor do curso é o Professor Carlos Manuel Dias Viegas, Professor Adjunto do Departamento de Engenharia de Computação e Automação (DCA) da Universidade Federal do Rio Grande do Norte (UFRN). O Professor Carlos Viegas é Doutor em Engenharia Informática (2015) pela Faculdade de Engenharia da Universidade do Porto, Porto, Portugal. Mestre em Engenharia Elétrica e de Computação (2009) e Engenheiro de Computação (2006) pela Universidade Federal do Rio Grande do Norte, Natal/RN, Brasil. Tem experiência nas áreas de redes de computadores, segurança da informação, engenharia de dados e sistemas de comunicação de tempo real. Para contato enviar e-mail para [viegas@dca.ufrn.br](mailto:viegas@dca.ufrn.br).