SPARK - DISTRIBUIÇÃO E PROCESSAMENTO DE DADOS

FICHA DO CURSO

Curso: Spark - Distribuição e Processamento de Dados

Modalidade: EaD Autoinstrucional

Carga Horária Total: 36 horas

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 Pyton.

Conteudista: Carlos Manuel Dias Viegas.

1. 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.

2. Programa

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.

3. Procedimentos Didáticos e Metodológicos

O conteúdo está dividido entre 9 (nove) módulos/semanas, contando com uma carga horária total de 36 horas. A cada semana os(as) cursistas deverão desenvolver as seguintes atividades:

a) Estudar o material pré-aula como forma de preparo para a videoaula;
b) Assistir às videoaulas programadas para a semana;
c) Realizar o estudo individual dos materiais indicados, tais como leituras complementares, resolução de exercícios e acesso aos vídeos adicionais;
d) Realizar as atividades avaliativas semanais, respondendo aos questionários aplicados.

O curso foi previamente ofertado na modalidade síncrona a um público selecionado do Poder Judiciário. A presente versão encontra-se no formato autoinstrucional, não havendo, portanto, interação com um tutor e/ou com outros/as cursistas. O prazo para a conclusão das atividades e do curso, como um todo, dependerá do empenho, dedicação e disponibilidade do/a cursista.

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.

Ademais, 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.

4. Carga horária do(a) cursista

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

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

Atividade semanal	**Carga horária (horas)**	**Fração (%)**
Estudo individual de conteúdo pré-aula	00:30	12,5
Assistir às videoaulas	02:00	37,5
Estudo individual pós-aula	00:30	12,5
Atividades Avaliativas	01:00	25,0
Total	04:00	100

5. Conteúdo Programático

O curso está dividido em 9 módulos, sendo que cada módulo configura uma semana de atividades. O conteúdo programado é apresentado na Tabela 2.

Tabela 2 – Conteúdo programático previsto e respectivo período.

Semana	**Conteúdo**
1	**Apache Hadoop** Introdução ao Curso Introdução ao Ecossistema Hadoop

	Sistema de Arquivos HDFS e Modelo de Programa MapReduce Gerenciamento de Recursos com Yarm Implantação do Hadoop – Partes 1 a 4 Implantação do Hadoop – Perguntas e Respostas
2	**Apache Hadoop** Roteiro da Segunda Semana Desenvolvimento de Aplicações com MapReduce Execução e Monitoramento de Tarefas Visão geral dos Módulos Adicionais Hadoop - Hbase Intalação do Hbase Apache Sqoop Apache Mahout
3	**Apache Spark** Introdução ao Ecossistema Spark – Partes 1 e 2 Comparação Spark vs Hadoop Abstração de dados - RDD, Dataframe e Dataset Instalação e Configuração do Spark e Integração com Apache Hadoop
4	**Apache Spark** Programação com pySpark: DataFrame – Partes 1 a 3 API Pandas no Spark Spark Web UI - Interface de usuário e DAG
5	**Apache Spark - SparkSQL** Programação com pySpark: Dataset – Partes 1 e 2 Programação com pySpark: SparkSQL Manipulação de Dados com SparkSQL
6	**Apache Spark – MLlib** Fundamentos de Machine Learning – Partes 1 a 3 Machine Learning no Spark e Criação de pipelines

	Aplicações práticas – Partes 1 e 2
7	**Apache Spark - Streaming** Apache Spark Streaming – Partes 1 a 4
8	**Apache Spark - Spark R** Introdução ao Spark R
9	**Apache Spark - GraphX** Fundamentos de Grafos – Partes 1 e 2 Análise de Grafos com GraphX e GraphFrames – Partes 1 a 3 Análise de Grafos com GraphX e GraphFrames – Aplicação – Partes 1 e 2

6. Avaliação dos(as) cursistas no curso

O desempenho no curso será determinado pela nota em atividade avaliativa aplicada aos(às) cursistas. Requer-se, no mínimo, 70% de aproveitamento para obtenção de certificado. A nota final será determinada pela média aritmética simples das notas obtidas nas atividades semanais.

7. 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 a avaliar a percepção dos(as) cursistas quanto ao curso, incluindo: materiais disponibilizados, qualidade das videoaulas, carga horária e demais recursos pedagógicos oferecidos pelo curso.

8. Referências Bibliográficas

- Parsian, Mahmoud. **Data Algorithms with Spark: Recipes and Design Patterns for Scaling Up Using Pyspark**. Sebastopol, CA, USA: O'Reilly Media, 2022. Print.

- Damji, Jules; Wenig, Brooke; Das, Tathagata; Lee, Denny. **Learning Spark: Lightning-Fast Data Analytics**. Sebastopol, CA, USA: O'Reilly Media, 2020. Print.
- Chambers, Bill; Zaharia, Matei. **Spark: The Definitive Guide: Big Data Processing Made Simple**. Sebastopol, CA, USA: O'Reilly Media, 2018. Print.
- White, Tom. **Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale**. Sebastopol, CA, USA: O'Reilly Media, 2015. Print.
- Hamstra, Mark; Zaharia, Matei; Karau, Holden. **Learning Spark: Lightning-Fast Data Analysis**. Sebastopol, CA, USA: O'Reilly Media, 2015. Print.
- **Documentação Apache Hadoop**. Disponível em: <https://hadoop.apache.org/docs/current/>
- **Documentação Apache Spark**. Disponível em: <https://spark.apache.org/docs/latest/>
- **Spark with Python (PySpark) Tutorial For Beginners**. Spark by Examples. Disponível em: <https://sparkbyexamples.com/pyspark-tutorial/>
- **Documentação SparkR (R on Spark)**. Disponível em: <https://spark.apache.org/docs/latest/sparkr.html>
- 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.

9. Informações sobre o Conteudista

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.

10. 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.