CURTA DURAÇÃO

ANÁLISE DE BIG DATA COM SPARK E PYTHON

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO
BUSINESS SCHOOL

CARGA HORÁRIA: 48 horas

COORDENAÇÃO:
Prof.ª Drª Alessandra de Ávila Montini

APRESENTAÇÃO

O crescente volume de dados gerados atualmente implica na necessidade de processar e realizar a análise dos dados de maneira eficiente, rápida e simplificada. As aplicações criadas utilizando o framework Apache Spark atendem esses requisitos, uma vez que essas aplicações rodam 10-100 vezes mais rápido do que os trabalhos executados no Hadoop MapReduce, além de possibilitar o uso de diferentes linguagens de programação.

Portanto, as empresas estão adaptando suas infraestruturas para essa tecnologia e buscando especialistas que dominem o desenvolvimento de soluções utilizando o Spark.

Este curso tem por objetivo, através do uso do Apache Spark e da linguagem de programação Python, realizar o processamento e análise de diversos conjuntos de dados, bem como entender o funcionamento dos componentes envolvidos neste framework.

DIFERENCIAL

- A FIA é líder em educação executiva.
- O LabData é um dos pioneiros no lançamento dos cursos de Big Data e Analytics no Brasil.
- Laboratórios de alta qualidade
- Participação gratuita dos alunos do LabData nas batalhas de dados e Hackatons
- Participação gratuita das palestras do Labdata com profissionais de grandes multinacionais
- Todas as aulas são práticas.

OBJETIVO

Após realizar esse curso, os alunos serão capazes de:

- Descrever os mecanismos fundamentais do Apache Spark
- Utilizar as principais APIs do Spark para realizar operações com os dados
- Aprender e utilizar a linguagem de programação Python para criar as aplicações com o Spark
- Implementar casos de uso típicos para o Apache Spark.
- Construir pipelines dos dados e realizar consultas em grandes volumes de dados utilizando o Spark SQL e DataFrames
- Analisar os trabalhos executados pelo Spark através da interface gráfica administrativa, bem como os logs
- Entender o funcionamento interno do Spark
- Processar fluxos de dados em tempo real com escalabilidade, alta vazão e tolerante a falha através do Spark Streaming
- Conhecer as funcionalidades das bibliotecas de aprendizagem de máquina disponível no Spark

PERFIL DO ALUNO

Profissionais de diversas áreas que desejam entender o framework Apache Spark.

CORPO DOCENTE

O corpo docente conta com professores **altamente capacitados, com experiência no mundo corporativo.** Nos critérios de seleção do corpo docente, são priorizadas as qualificações e experiências profissionais nas distintas matérias, de maneira que o curso permita não somente a transmissão de conhecimentos, mas também experiências enriquecedoras para os alunos.

METODOLOGIA

Aulas expositivas, resolução de exercícios práticos e estudo de casos.

CONHEÇA O LABDATA

- Convido você a assisir o vídeo do LabData e conhecer nossos laboratórios.

Acesse o QR code.

MATRIZ CURRICULAR

Introdução ao Spark e Python

- Visão geral do ecossistema Spark
- Componentes básicos do Spark
- Aplicações com Spark
- Estudo de caso
- Configuração do ambiente Spark

Fundamentos RDD

- Revisão dos principais conceitos do Python
- Introdução ao RDDs (*Resilient Distributed Dataset*)
- Transformações, Ações e DAG (*Directed Acyclic Graph*)
- API de programação RDD
- Realizar consultas interativas utilizando RDDs

Spark SQL e DataFrame

- Conceitos de Spark SQL e DataFrame
- APIs DataFrame e SQL
- Otimização de consultas Catalyst
- ETL (Extrair, Transformar e Carregar)
- Realizando consultas com DataFrame e SQL
- **Caching**
- Visualização

Funcionamento Interno do Spark

- Jobs, Stages e Tasks
- Desempenho do Job
- Uso de caching e melhores práticas
- Resolução de problemas utilizando a interface do Spark
- Visualizar como os jobs são divididos e executados dentro do Spark
- Análise dos logs dos executores
- Visualizar execução dos DAG (*Directed Acyclic Graph*)
- Visualizar consultas SQLs
- Observar a execução das tarefas (Tasks)
- Entendendo o desempenho
- Medindo a memória utilizada

Spark Streaming

- Fontes de fluxos de dados em tempo real
- APIs disponíveis no Spark Streaming
- Confiabilidade e recuperação de falhas
- Otimização do desempenho
- Operações realizadas nos fluxos de dados
- Lendo dados de diferentes fontes (por exemplo: TCP, Kafka)
- Visualização contínua
- Visualizando os trabalhos (jobs) de fluxos na interface do Spark

Spark Streaming

- Extrair, transformar e carregar dados de múltiplas fontes de dados (JSON, base de dados relacionais, etc) com DataFrames
- Extrair dados estruturados de fontes de dados não estruturados através de transformações e ações

- Tratar valores faltando no conjunto de dados
- Aplicar as melhores práticas para análise de dados com Spark
- Realizar análise exploratória de dados utilizando DataFrames
- Visualizando dados através de bibliotecas populares de visualização do Python

Introdução ao Aprendizado de Máquina

- Princípios básicos de Aprendizagem de Máquina
- Padrões de API de Aprendizagem de Máquina do Spark
- Utilizar algoritmos de aprendizagem de máquina através de **pipelines** e DataFrames
- Exemplos práticos para classificação e clusterização

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

BUSINESS SCHOOL

INFORMAÇÕES

Tel: (11) 3732-3535 | faleconosco@fia.com.br