CRT020: Databricks Certified Associate Developer for Apache Spark 2.4

Espera-se que os candidatos estejam familiarizados com os seguintes conceitos arquiteturais e seu relacionamento entre si:

- Driver
- Executor
- Core/Slots
- Jobs
- Stages
- Tasks
- Partitions
- Shuffling
- Wide vs Narrow Transformations

Espera-se que os candidatos tenham um comando das seguintes APIs, mas a memorização da API não é necessária.

- SparkSession
- DataFrameReader / DataFrameWriter
- DataFrame / Dataset
- Row / Column
- Spark SQL functions

Apache Spark™ Programming

- Use as APIs centrais do Spark para operar em dados
- Articular e implementar casos de uso típicos para o Spark
- Crie pipelines de dados e consulte grandes conjuntos de dados usando o Spark SQL e DataFrames
- Analise os trabalhos do Spark usando as UIs de administração dentro do Databricks
- Criar trabalhos de fluxo estruturado
- Trabalhar com dados relacionais usando as APIs GraphFrames
- Entenda como funciona um pipeline de Aprendizado de Máquina
- Entenda o básico dos componentes internos da Spark

Apache Spark™ for Machine Learning and Data Science

- Entenda quando e onde usar o Spark
- Articular a diferença entre um RDD, um DataFrame e um conjunto de dados
- Explicar o aprendizado de máquina supervisionado versus não supervisionado e os aplicativos típicos de ambos
- Construa um Pipeline Aprendizado de Máquina usando uma combinação de Transformadores e Estimadores
- Salvar / restaurar modelos
- Aplique modelos para transmitir dados
- Executar ajuste de hyperparameter com validação cruzada
- Analisar o desempenho da consulta do Spark usando a interface do usuário do Spark
- Treinar modelos com bibliotecas de terceiros, como o XGBoost
- Realize a pesquisa de hiperparâmetros em paralelo usando algoritmos de um único nó, como o scikit-learn
- Obter familiaridade com Árvores de Decisão, Random Forests, Gradient Boosted Trees, Regressão Linear, Filtragem Colaborativa e K-Means
- Explicar opções para colocar modelos em produção