Bootcamp: Cientista de Dados

Trabalho Prático

Módulo 2: Desenvolvendo Soluções Utilizando Apache Spark

Objetivos de Ensino

Bem-vindos(as) ao trabalho prático do módulo sobre Spark! Neste trabalho, você vai exercitar os conceitos trabalhados na primeira parte do módulo, e vai:

- ✓ Se acostumar a escrever e executar aplicações que usam o Spark;
- ✓ Construir aplicações Spark interativas usando o pyspark ou uma plataforma interativa como o *jupyter-lab*;
- ✓ Computar estatísticas descritivas usando o Spark;
- ✓ Manipular dados a partir da API de DataFrames.

É recomendado que você leia os capítulos 1, 2 e 3 da apostila e assista às aulas relacionadas a eles. Em particular, o capítulo 2 da apostila contém instruções para instalar o Spark na sua máquina.

Divirta-se!

Enunciado

Dados do mercado financeiro são interessantes e ricos: cada ação negociada na bolsa de valores tem um preço que varia a cada dia. Você foi contratado como cientista de dados de uma empresa de Wall Street para criar modelos preditivos que, a partir da variação diária do preço das ações, consigam subsidiar e melhorar decisões de compra e venda de ações. Você disse que, como todo bom cientista de

dados, gostaria de explorar os dados para entender suas características antes de criar qualquer modelo preditivo.

Os dados estão disponíveis em https://www.kaggle.com/camnugent/sandp500/ por meio do arquivo *all_stocks_5yr.csv*. O arquivo contém, para cada dia e ação do S&P 500 (lista de 500 maiores empresas americanas), os seguintes dados:

- *Date* - no formato yy-mm-dd
- *Open* - Preço da ação na abertura do mercado no dia, em dólares.
- *High* - Maior preço alcançado naquele dia.
- *Low* - Menor preço alcançado naquele dia.
- *Close* - Preço da ação no fechamento do mercado no dia.
- *Volume* - Número de ações vendidas / compradas.
- *Name* - O nome da ação.

Apesar do volume de dados ser pequeno, você decidiu usar o Apache Spark para processar os dados para aprender a ferramenta, e tendo em vista que a sua empresa disse que, em breve, obterá dados por minuto, e não por dia, e de todas as ações do planeta, não apenas dos Estados Unidos. Neste caso, uma ferramenta desenhada para lidar com *big data* será necessária, e você já quer estar com o código pronto.

Atividades

O aluno deve extrair as principais estatísticas descritivas do conjunto de dados, usando a API de Dataframe do Spark. Consulte a aula sobre DataFrames e materiais como:

- https://www.datasciencemadesimple.com/descriptive-statistics-or-summary-statistics-of-dataframe-in-pyspark/

- https://docs.databricks.com/spark/latest/dataframes-datasets/introduction-to-dataframes-python.html
- https://medium.com/analytics-vidhya/spark-group-by-and-filter-deep-dive-5326088dec80
- https://towardsdatascience.com/the-most-complete-guide-to-pyspark-dataframes-2702c343b2e8

As perguntas objetivas contêm perguntas específicas que o aluno deve responder por meio de aplicações Spark.