Exercício Programa
Análise de Dados com Apache Spark

Prof. Dr. Daniel Cordeiro e Norton Trevisan Roman
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Entrega: **25 de julho de 2021**

Descrição Geral

O conceito de *Resilient Distributed Dataset* (RDD) e a categorização de operações nesses *datasets* em transformações e ações mostraram-se níveis de abstração mais gerais do que MapReduce e, principalmente, permitiram a criação de sistemas distribuídos de larga escala de melhor desempenho. O uso do Apache Spark facilita (e muito!) o desenvolvimento de aplicações distribuídas para análises de **grandes** volumes de dados. Graças a este projeto de código aberto desenvolvido sob a tutela da Fundação Apache, qualquer desenvolvedor pode escrever aplicações distribuídas escaláveis que podem utilizar milhares de máquinas simultaneamente.

Este exercício programa consiste da implementação de uma interface rica para a análise desses dados por um cientista de dados e da implementação de mecanismos de análise distribuídos implementados usando o Spark.

Usaremos o histórico de dados meteorológicos coletados pelo *National Climactic Data Center (NCDC)* para desenvolver um sistema de análise de informações meteorológicas distribuído escrito usando o paradigma de programação introduzido pelo Spark.

Os dados a serem utilizados (cerca de 20 GB) estão disponibilizados publicamente em `https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ncdc:C00516` e em `https://aws.amazon.com/public-datasets/gsod/` (use-o se você usar a Amazon). Leia atentamente a descrição dos dados na página e nos arquivos `.txt` disponibilizados junto com os dados.

Neste EP vocês deverão processar os dados e fazer algumas análises sobre eles. Essas podem se limitar a estatísticas descritivas simples (como médias, desvios-padrão, dentre outras), além de usar um método de regressão para estimar medidas futuras, como o método dos quadrados mínimos, por exemplo.

Análise

Para auxiliá-los, abaixo são descritas algumas possibilidades.

Média e desvio padrão

O programa deve ser capaz de calcular pelo menos a média e o desvio padrão de um conjunto de dados. A implementação de outras funções estatísticas será recompensada `:-)`.

Dada uma coleção $X = \{x_1, \ldots, x_n\}$ de n amostras de uma determinada medida, a média desta coleção é dada por

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Para tal amostra, se $n > 1$, o desvio padrão é dado por

$$\sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

Método dos quadrados mínimos

Dados valores $y_1, \ldots, y_n$, cada um associado a uma abcissa $x_1, \ldots, x_n$, podemos interpretar y_i como o valor de uma função no ponto x_i, ou seja, $y = f(x)$, para $x = x_1, \ldots, x_n$. O método dos quadrados mínimos é uma maneira de aproximar uma função assim por uma função linear, dada por $y = a + bx$.

Os valores de a e b são determinados a partir das médias $\overline{x}$ e $\overline{y}$ dos valores $x_1, \ldots, x_n$ e $y_1, \ldots, y_n$ da seguinte maneira:

$$b = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sum_{i=1}^{n} x_i(x_i - \overline{x})} \qquad \text{e} \qquad a = \overline{y} - b\overline{x}$$

Objetivo

Você deve implementar uma interface de usuário que permita aplicar as funções estatísticas escolhidas a cada tipo de informação disponível no *dataset* (temperatura, velocidade do vento, umidade, pressão etc.).

A interface (que pode ser uma API, um cliente em linha de comando ou uma interface gráfica) deve permitir que o usuário especifique:

- O tipo de informação que será analisada;
- Qual o período de tempo que será considerado na análise; e
- Como o resultado deve ser agrupado (ex: média de cada ano, média de cada mês, média de cada dia da semana etc.)

Além disso, sua interface deve permitir que o cientista de dados faça a predição dos valores do tipo de informação que este escolher (dentre as disponíveis no *dataset*), usando o método dos quadrados mínimos. O método dos quadrados mínimos recebe duas listas (X e Y) de mesmo tamanho, e devolve dois números (y_o e y_1) calculados da seguinte maneira:

- Aplique o método dos quadrados mínimos sobre os pontos dados por X e Y para as entradas válidas de Y, para obter valores a e b que aproximem esses pontos por uma reta;
- Devolva $y_0 = a + bx_{\min}$ e $y_1 = a + bx_{\max}$, onde $x_{\min}$ e $x_{\max}$ são respectivamente o menor e o maior valor na lista X.

Por fim, seu sistema deve gerar um gráfico com os valores da informação escolhida pelo usuário (no eixo y), ao longo do período escolhido (eixo x), que mostre claramente o valor da estatística, os desvios padrão e a reta determinada pelo método dos quadrados mínimos.

Instruções

Seu programa deve ser implementado usando apenas o Apache Spark, ou seja, não é permitida a utilização de outros projetos da Apache para auxiliar o desenvolvimento.

A execução dos experimentos pode ser realizada tanto localmente, em uma instalação do Apache Spark no(s) seu(s) computador(es), quanto em um provedor de Computação em Nuvem. Apesar de não ser um requisito para o EP, aproveite a oportunidade para configurar uma conta em uma plataforma de Computação em Nuvem e experimente suas possibilidades.

Vários provedores de Computação em Nuvem dão créditos para novos usuários e definem alguns recursos que podem ser utilizados de graça, desde que dentro de certos limites (o chamado *free tier*). Dentre eles, Amazon e Google também dão acesso a uma plataforma pré-configurada para computação usando Spark. Veja os sites `https://aws.amazon.com/pt/emr/features/spark/` e `https://cloud.google.com/dataproc/` para mais informações.

Note que o uso indiscriminado de uma plataforma de computação em nuvem **pode gerar custos financeiros** (que serão debitados do cartão de crédito associado à conta quando os créditos acabarem). Os possíveis custos adicionais incorridos da má utilização da plataforma são de inteira responsabilidade dos grupos. Se estiver em dúvida, **utilize uma instalação local**.

Observações

Para este trabalho, vocês devem se organizar em grupos de 4 (quatro) ou 5 (cinco) pessoas.

Dúvidas em relação ao EP devem ser discutidas no fórum da disciplina no ediscipliinas: `https://edisciplinas.usp.br/`. Todos são **fortemente encorajados** a participar das discussões e ajudar seus colegas.

Entregue junto com o código-fonte do programa um **relatório detalhado** que:

1. Descreva como as funções estatísticas foram implementadas usando o Spark;
2. Mostre vários exemplos de entradas (consultas) e os respectivos resultados obtidos; e
3. Explique em detalhes todos os passos necessários para a execução do programa.

A entrega será feita unica e exclusivamente via edisciplinas, até a data final marcada. Um (e apenas um) dos integrantes do grupo deve fazer a postagem de um arquivo zip, tendo como nome o número USP desse integrante:

`número_usp.zip`

Dentro do zip devem constar seu código fonte (organizado em diretórios e subdiretórios, conforme sua implementação), além do relatório final de entrega. **Não esqueçam de colocar os integrantes do grupo na capa do relatório**.

A responsabilidade de postagem é exclusivamente sua. Por isso, submeta e certifique-se de que o arquivo submetido é o correto (fazendo seu download, por exemplo). Problemas referentes ao uso do sistema devem ser resolvidos com antecedência.