Exercício Programa 2
Análise de Dados com MapReduce

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

Entrega: **2 de julho de 2017**

Descrição Geral

O paradigma de programação MapReduce facilita (e muito!) o desenvolvimento de aplicações distribuídas para análises de **grandes** volumes de dados. Graças ao projeto de código aberto Hadoop, 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 MapReduce.

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 MapReduce.

Os dados a serem utilizados (cerca de 20 GB) estão disponibilizados publicamente em `ftp://ftp.ncdc.noaa.gov/pub/data/gsod/` 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 alguns cálculos (como médias, desvios-padrão, e outros), além de usar um método de estimação para as medidas.

Análise

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 $x_1, \ldots, x_n$ de amostras de uma 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 mínimos quadrados

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. O método dos mínimos quadrados é uma maneira de aproximar uma tal função 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:

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

Objetivo

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

A interface deve permitir que o usuário especifique:

- o tipo de informação que será analisada;
- qual o período de tempo que será considerado na análise;
- 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 gráfica deve permitir que o cientista de dados faça a predição dos próximos valores usando o método dos mínimos quadrados.

O método dos mínimos quadrados recebe duas listas de mesmo tamanho (x e y) e devolve dois números (y_o e y_1) calculados da seguinte maneira. Aplique o método dos mínimos quadrados sobre os pontos dados por x e y, para as entradas válidas de y, para obter valores a e b que aproximem os pontos dados

por x e y por uma reta. Devolva $y_0 = a + bx_{\text{min}}$ e $y_1 = a + bx_{\text{max}}$, 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 estatística escolhida (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 mínimos quadrados.

Instruções

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

O EP será testado em um computador equipado apenas com o sistema operacional GNU/Linux e com a distribuição Ubuntu 17.04 (Zesty Zapus).

A execução dos experimentos pode ser realizada tanto localmente, em uma instalação do Apache Hadoop 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 MapReduce. Veja os sites `https://aws.amazon.com/pt/elasticmapreduce/` e `https://cloud.google.com/dataproc/` para mais informações.

Nosso curso está inscrito no programa AWS Educate `http://aws.amazon.com/education/awseducate`, um programa educacional do provedor de Computação em Nuvem Amazon. Os alunos que se inscreverem nesse programa ganharão também US$ 100,00 em créditos para utilizar na Amazon.

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 é de inteira responsabilidade dos grupos. Se estiver em dúvida, utilize uma instalação local.

Observações

O EP deve ser feito em *equipes de 3 ou 4 pessoas*.

Dúvidas em relação ao EP devem ser discutidas no fórum da disciplina. Todos são **fortemente encorajados** a participar das discussões e ajudar seus colegas.

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

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