Mineração de Dados Complexos
Curso de Aperfeiçoamento
INF-0617 - Big Data

Map Reduce Local
INF-0617 - Big Data

Rafael Fernando Ribeiro
Thiago Gomes Marçal Pereira

Prof. Lucas Wanner

Outubro de 2018

Apresentação do Problema

Neste trabalho, fomos apresentados com a medição de temperatura de diversas estações de medição, bem como outras medidas realizadas. E dados de localização geográfica, data e horário das medições.

O objetivo era a obtenção da maior diferença de temperatura no período de uma hora.

Solução

O projeto consistiu em uma implementação de MapReduce, em Python.

Para o Mapper, a partir dos dados passados à aplicação, foram feitas algumas análises iniciais antes de se proceder com a seleção dos dados enviados ao Reducer. Essas análises consistiram basicamente em verificar a validade dos dados. Para tal, inicialente foram ignorados os dados que continham valores inválidos “-99.0” ou “-9999.0” para os valores das temperaturas e também linhas que contivessem temperatura mínima maior que a máxima. Após algumas execuções e conferência dos dados, também passamos a considerar alguna valores de *“threshold”*, e ignorar linhas que contivessem algum valor “-9999.0” em alguma das primeiras 27 colunas, pois isso indicava que os valores não eram completamente confiáveis. Havia por exemplo valores inválidos substituídos por 0, ou alguma medição próxima a 0. Os valores de *“threshold”* foram obtidos de forma empírica após algumas execuções.

Tendo-se removido os valores inválidos, criamos um Map, onde a chave era o ID da estação e os valores (separados por espaço) eram: latitude, longitude, data, hora e a diferença de temperatura.

Para o reducer, utilizamos um dicionário com a chave sendo o ID da estação, e nos campos, outros dicionários, com a chave sendo cada campo, e os valores sendo o relativo ao de maior valor de diferença de temperatura para aquela estação. Ao final, tendo-se os maiores valores de cada estação, buscamos o maior valor entre as estações.

O resultado obtido foi:

Station at (-110.29, 46.88) measured 18.3 degree(s) difference at 2017-01-10 0400

Ou seja, uma diferença de temperatura de 18,3° no dia 10/01/2017 às 4 da manhã, na estação indicada por essas posições. A entrada está no arquivo *CRNH0203-2017-MT_Lewistown_42_WSW.txt* e olhando as entradas, elas parecem válidas.

Para melhoria futura, poderíamos melhorar a forma como o Mapper faz a análise de dados inválidos, economizando checagens que foram sendo adicionadas com a realização de testes. Outra maneira possível seria talvez a não necessidade de passar as informações de localização em todas as entradas da estação, visto que essa informação não muda. Isso poderia talvez economizar tempo na transmissão dos dados.