Untitled

Arthur Sena

03/01/2016

# Descriçao dos dados

Os dados são referentes aos usuários de ônibus de Curitiba, onde cada linha do arquivo segue o seguinte padrão:

## CODLINHA NOMELINHA CODVEICULO NUMEROCARTAO  
## [1,] "ARA" "SISTEMA ARAUCARIA" "00181" "0002624774"  
## [2,] "ARA" "SISTEMA ARAUCARIA" "00994" "0002575080"  
## [3,] "ARA" "SISTEMA ARAUCARIA" "00232" "0002849283"  
## [4,] "ARA" "SISTEMA ARAUCARIA" "00232" "0001502669"  
## [5,] "ARA" "SISTEMA ARAUCARIA" "00998" "0003535943"  
## [6,] "ARA" "SISTEMA ARAUCARIA" "00998" "0003496859"  
## DATAUTILIZACAO   
## [1,] "20/10/15 17:35:27,000000"  
## [2,] "20/10/15 18:25:34,000000"  
## [3,] "20/10/15 09:49:50,000000"  
## [4,] "20/10/15 13:07:02,000000"  
## [5,] "20/10/15 09:55:50,000000"  
## [6,] "20/10/15 12:37:20,000000"

Analisando os dados acima, vamos utilizar o HADOOP para responder algumas perguntas. O código MapReduce foi implementado utilizando Python.

# Perguntas

## 1) Quais as linhas que são mais utilizadas ?

Para responder essa pergunta precisamos fazer uma simples contagem em cima das linhas de ônibus.

**mapper.py**: O map apenas retorna uma tupla com a linha e o inteiro 1.

import sys  
   
 for line in sys.stdin:  
 if len(line) > 1:  
 cod\_linha = line.strip().split(',')[1].split(":")[1]  
 print "%s\t%s" % (cod\_linha,1)

**reducer.py**: O meu reducer conta quantas vezes aquela linha foi encontrada e retornar uma tupla com a linha e quantidade de vezes.

import sys  
  
 value\_temp = 0  
 cod\_linha\_temp = ''  
 cod\_linha\_atual = ''  
   
 for line in sys.stdin:  
 cod\_linha = line.split('\t')[0]  
 value = line.split('\t')[1]  
   
 if cod\_linha\_temp == '':  
 cod\_linha\_temp = cod\_linha  
 cod\_linha\_atual = cod\_linha  
 value\_temp = value\_temp + int(value)   
 elif cod\_linha\_temp == cod\_linha:  
 value\_temp = value\_temp + int(value)  
 else:  
 print "%s\t%s" % (cod\_linha\_temp,value\_temp)  
 cod\_linha\_temp = cod\_linha  
 value\_temp = 0  
 cod\_linha\_atual = cod\_linha  
   
 if cod\_linha\_atual == cod\_linha\_temp:  
 print "%s\t%s" % (cod\_linha\_temp,value\_temp)

Agora vamos rodar executar o Map e o Reduce com o comando abaixo

bin/hadoop jar contrib/streaming/hadoop-streaming-1.2.1.jar -file /home/arthur/Documents/DataAnalysis\_Problems/DataCourse2/Problem2/mapper.py -mapper /home/arthur/Documents/DataAnalysis\_Problems/DataCourse2/Problem2/mapper.py -file /home/arthur/Documents/DataAnalysis\_Problems/DataCourse2/Problem2/reducer.py -reducer /home/arthur/Documents/DataAnalysis\_Problems/DataCourse2/Problem2/reducer.py -input input/doc1-2015102121.txt -output output

Porque eu estou utilizando python, eu precisei usar o jar 'hadoop-streaming-1.2.1.jar' para executar. Ao fim da execução o Hadoop gerou o arquivo 'output/part-00000' contendo a resposta. Uma parte do arquivo pode ser vista logo abaixo:

## Linha Quantidade  
## 154 OPER S\\/LINHA 221661  
## 153 OP. CONTIGENCIA 32206  
## 192 SISTEMA ARAUCARIA 8576  
## 96 INTERBAIRROS IV 7056  
## 91 INTERB II ANTI H 5947  
## 97 INTERBAIRROS V 5674

Observando a tabela vemos as linhas mais frequentes

## 2) Quais linhas apresentam mais veiculos ?

## 3) Qual o número do cartão mais utilizado ? Quantas vezes ?