UNIVESIDADE POSITIVO

PÓS-GRADUAÇÃO EM BIGDATA

Contagem das palavras em PIG da base de dados lit2go

Alunos:

Aylson Martins

Jonas Batista

Disciplina:

Hadoop e MapReduce

Professor:

Alessandro Binhara

Curitiba, 04/04/2016

Proposta de análise

Este trabalho terá por objetivo extrair as 150 palavras mais utilizadas, da base de dados “lit2go” (base de dados com e-books), agrupado por ano e por país de origem utilizando a tecnologia PIG.

Orientação para execução do script

Devido às dificuldades de trabalhar com a análise dos dados no formato HTML na leitura e organização dos dados em PIG foi necessário à utilização do software “ClearFile.exe”, criado para este projeto em específico para extrair os dados do formato HTML e incluir no início de cada registro (texto para cada capítulo do livro) o ano e o país de origem dos livros (separando os campos pelo caractere “|”).

Para utilizar o software “ClearFile.exe” é necessário extrair os arquivos do “.zip” (“ClearFile.exe.zip”) e renomear o arquivo “ClearFile.exe.txt” para “ClearFile.exe”. Neste executável será possível selecionar o diretório de origem dos arquivos que deverão ser processados para retirar somente o texto dos livros, com o ano e país de origem e a seleção do diretório destino onde será criado o arquivo “dataOut.txt” com as informações disponíveis para análise através do script em PIG.

A utilização do script em PIG foi realizada no HUE. Para a utilização do script será necessário a criação do diretório de origem dos dados “/user/cloudera/lit2go/input/” com o arquivo “dataOut.txt” e a criação do diretório de destino “/user/cloudera/lit2go/output/”. Neste diretório de destino será criado o diretório “word\_count” com os resultados do script.

Segue abaixo o script em PIG utilizado para encontrar as 150 palavras mais utilizadas agrupadas por ano e país de origem:

Código fonte

*a = load '/user/cloudera/lit2go/input/dataOut.txt' USING PigStorage('|') AS (ano:chararray, pais:chararray, texto:chararray);*

*b = foreach a generate ano, pais, flatten(TOKENIZE((chararray)texto)) as word;*

*c = group b by (ano, pais, word);*

*d = foreach c { generate FLATTEN(group) AS (ano, pais, word), COUNT(b);};*

*e = FOREACH (group d BY (ano, pais)){ f = ORDER d by $3 desc; g = LIMIT f 150; GENERATE flatten(g);};*

*store e into '/user/cloudera/lit2go/output/word\_count';*

## Base de dados

Arquivo: https://dl.dropboxusercontent.com/u/10599684/pos-bigdata/lit2go.ok.tar.gz

Site: http://etc.usf.edu/lit2go/