Marcelo Paulon - 1411029

Descoberta do Conhecimento – 3º Trabalho

1) Criar uma máquina virtual (Virtual Box 5.1.26 ou superior) Linux (sugestão, **Ubuntu 16.04 with a non-root user with** sudo **privileges**) e instalar e configurar o Hadoop.

OK.

2) Criar uma aplicação Java chamada WordCount que conta o número de ocorrências de cada palavra nos arquivos wordcount_01.txt, wordcount_02.txt e wordcount_01.txt e grava um arquivo de saída com as palavras encontradas e as respectivas quantidades. O Hadoop deverá ser executado em *single-node* em modo *pseudo-distributed*.

Deve ser entregue a máquina virtual criada e um documento que explique passo a passo como a aplicação WordCount pode ser executada e onde poderão ser encontrados os arquivos Java e de entrada e saída utilizados no processamento.

Arquivos:

***$REPO*: /home/marcelo/PUC-Rio-INF1399-Inteligencia-Computacional/trabalho 3**

Arquivo Java: $REPO/src/main/java/com/wordcount/Main.java

JAR: $REPO/target/wordcount-1.0-SNAPSHOT.jar

Arquivos de entrada (HDFS): /user/marcelo/wordcount/input

Arquivos de saída (HDFS): /user/marcelo/wordcount/output

Execução da aplicação:

(O arquivo JAR já está na pasta, mas caso seja necessário gerá-lo novamente, basta executar "mvn package" no diretório $REPO)

1. Executar o comando "./run.sh", no diretório $REPO

OU, também no diretório $REPO, executar os seguintes comandos, em ordem:

1. "../../hadoop/sbin/stop-dfs.sh"

2. "../../hadoop/bin/hadoop namenode -format" (caso o programa peça confirmação, digitar "Y" e pressionar a tecla Enter)

3. "../../hadoop/sbin/start-dfs.sh"

4. "../../hadoop/bin/hadoop fs -mkdir /user /user/marcelo /user/marcelo/wordcount /user/marcelo/wordcount/input"

5. "../../hadoop/bin/hadoop fs -put wordcount_0* /user/marcelo/wordcount/input"

6. "../../hadoop/bin/hadoop fs -rm -r /user/marcelo/wordcount/output"

7. "../../hadoop/bin/hadoop jar target/wordcount-1.0-SNAPSHOT.jar com.wordcount.Main /user/marcelo/wordcount/input /user/marcelo/wordcount/output"

8. "../../hadoop/bin/hadoop fs -cat /user/marcelo/wordcount/output/*"

<u>Resultado:</u>

hadoop fs -cat /user/marcelo/wordcount/output/*

Data 1

Fuzzy 1

Link 1

MapReduce 1

Neural 1

Text 1

a 4

algorithm 1

an 1

analysis 1

analytics 1

and 4

associated 1

been 1

between 2

big 1

board 1

cluster 1

completely 2

computer 1

computing 1

concept 1

data 2

deriving 1

diagnosis 1

discovering 1

distributed 1

employed 1

evaluate 1

false 1

filtering 1

for 1

from 1

games 1

generating 1

handle 1

have 1

high-quality 1

implementation 1

in 1

information 1

is 5

large 1

logic 1

machine 1

may 1

medical 1

mining 2

model 1

network 1

networks 1

nodes 1

of 3

on 2

or 1

parallel 1

partial 1

patterns 1

playing 1

process 2

processing 1

programming 1

range 1

recognition 1

relationships 1

sets 2

social 1

speech 1

technique 1

text 2

the 4

to 2

translation 1

true 1

truth 2

used 2

value 1

video 1

vision 1

where 1

with 1

Referências:

http://hadoop.apache.org/docs/stable/index.html

http://www.tutorialspoint.com/map_reduce/index.htm;

https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html