MapReduce – Hands-On

Objetivo: O objetivo deste hands-on é compilar e executar a aplicação WordCount usando MapReduce nas linguagens Java e Python.

Compilação e execução do WordCount em Java

1. Baixe os arquivos WordCountDriver.java, WordCountMapper.java, WordCountReducer.java e shakespeare.txt da pasta "05_MapReduce/WordCountJava" do repositório no GitHub (https://github.com/GerInfraBigDataPUCRS/CursoDataScience2018/).

2. Crie um diretório chamado shakespeare no HDFS.

```
$HADOOP_HOME/bin/hadoop fs -mkdir shakespeare
```

3. Faça o upload do arquivo shakespeare.txt para a pasta criada no HDFS

```
$HADOOP_HOME/bin/hadoop fs -put shakespeare.txt Shakespeare
```

4. Compile o código fonte do programa WordCount.

```
$ javac -classpath `$HADOOP_HOME/bin/hadoop classpath` *.java
```

5. Crie o arquivo jar

```
$ jar cvf wc.jar *.class
```

6. Execute o programa WordCount

```
$ $HADOOP_HOME/bin/hadoop jar wc.jar WordCountDriver -D mapreduce.job.reduces=2
shakespeare wordcount
```

7. Verifique o conteúdo do diretório "wordcount" contendo o resultado da execução

```
$ $HADOOP_HOME/bin/hadoop fs -ls wordcount
```

8. Inspecione o conteúdo dos arquivos

```
$ $HADOOP_HOME/bin/hadoop fs -tail wordcount/part-r-00000
```

9. Crie um programa MapReduce que ordene o resultado do WordCount de acordo com o número de ocorrências das palavras no documento. (Dica: não é necessário implementar nenhum algoritmo de ordenação!)

Execução do WordCount em Python

1. Baixe os arquivos wordmapper.py, wordreducer.py e shakespeare.txt da pasta "05_MapReduce/WordCountPython" do repositório no GitHub (https://github.com/GerInfraBigDataPUCRS/CursoDataScience2018/). Iremos utilizar o mesmo arquivo de entrada do exemplo em Java (shakespeare.txt).

2. Antes de executar o código no cluster Hadoop, teste a execução local do código.

```
$ head -n 100 shakespeare.txt | ./wordmapper.py | sort | ./wordreducer.py
```

3. Agora execute o código usando o MapReduce Streaming API.

```
$HADOOP_HOME/bin/hadoop jar \
$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-2.9.1.jar \
-input shakespeare \
-output wordcountpy \
-mapper wordmapper.py \
-reducer wordreducer.py \
-file wordmapper.py \
-file wordreducer.py
```

4. Crie um programa MapReduce em Python que faça a ordenação do resultado do WordCount (palavras em ordem crescente de frequência). (Dica: não é necessário implementar nenhum algoritmo de ordenação!)