Índice

Exercícios de laboratórios Hadoop em Cloudera

Cria uma pasta em seu ambiente para armazenar estes Hacks (Exercícios práticos de laboratórios), os Hacks serão inicialmente usados e criados para a prática dos conhecimentos adquiridos no curso e posteriormente podem ser uma base de conhecimento para consulta.

HACK70 – Preparando para rodar um Mapreduce com dados do HDFS.

Neste hacks vamos preparar rodar um programa no formato Mapreduce dentro do Ambiente Hadoop.

1. Acesse a sua máquina Virtual Cloudera
2. Abra um Console no Linux
3. Com os comandos que aprendemos sobre HDFS criaremos alguns diretório

```
hadoop fs -mkdir /user/cloudera/wordcount
hadoop fs -mkdir /user/cloudera/wordcount/input
```

4. Estes diretórios serão necessários para copiarmos os arquivos que vamos processar pelo programa em MapReduce.

HACK71 – Criando arquivos para analise.

1. Vamos criar 3 arquivos bem simples para analisarmos em nosso MapReduce.
2. Acesse seu diretorio Home (/home/cloudera/) e crie um diretório no file system do linux para armazenarmos os nossos arquivos e programas

```
cd ~
mkdir hacks-hadoop
```

3. Crie arquivos texto de exemplo para usar como entrada/teste . Você pode usar qualquer arquivo que você escolher; Por conveniência, os seguintes comandos de shell criam alguns pequenos arquivos de entrada para fins ilustrativos.

```
echo "Hadoop is an elephant" > file0
echo "Hadoop is as yellow as can be" > file1
echo "Oh what a yellow fellow is Hadoop" > file2
```

HACK72 – Copiando os arquivos para o HDFS.

1. Mova os arquivos para o diretório **/user/ cloudera/wordcount/ input** no HDFS

```
hadoop fs -put file* /user/cloudera/wordcount/input
```

Atenção: nas aulas ensinamos como usar o comando CopyFromLocal mas temos mais opções lendo em https://hadoop.apache.org/docs/r2.7.3/hadoop-project-dist/hadoop-common/FileSystemShell.html

HACK73 – Baixando um exemplo de MapReduce para Cloudera.

1. Acesse o Material de EAD e baixe os arquivos de exemplos de Wordcout da Cloudera no tópico “MapReduce com Distribuição Cloudera”
2. Baixe o mesmo e descompacte dentro do diretório /home/cloudera/hacks-hadoop na sua maquina virtual Cloudera.
3. Ao descompactar será criado um diretório /home/cloudera/hacks-hadoop/hadoop_tutorial com 3 diretórios (wordcount1 , wordcount2 e wordcount3)

HACK74 – Executando o WordCount MapReduce.

1. Pelo console no Linux acesse o diretório onde descompactamos os programas , vamos usar o workcount1

```
cd /home/cloudera/hacks-hadoop/hadoop_tutorial/WordCount1
```

2. Agora vamos executar o programa java/mapreduce que baixamos

3. Execute o aplicativo WordCount do arquivo JAR, passando os caminhos para os diretórios de entrada e saída no HDFS (o diretório output é criado pelo programa MR).

```
hadoop jar wordcount.jar org.myorg.WordCount
/user/cloudera/wordcount/input /user/cloudera/wordcount/output
```

4. Serão emitidos diversos Logs no Console e possivelmente WARNs também (dependendo da versão do ambiente) e processo demorá alguns minutos para execução , acompanhe.

HACK75 – Analisando os Resultados de Saída.

1. Um ou mais arquivos de saida são gerados no diretório output , Quando você olha para a saída, todas as palavras estão listadas em UTF-8 ordem alfabética (palavras em maiúsculas primeiro). O número de ocorrências de todos os arquivos de entrada foi reduzido a uma única soma para cada palavra.

2. Execute um cat para visualizar o conteúdo dos arquivos

```
hadoop fs -cat /user/cloudera/wordcount/output/*
```

Atenção: Caso já tenha um diretório com o nome **output** e queira executar novamente , você deve excluir o mesmo ou direcionar sua saida para outro diretorio , ex, Output2

```
hadoop fs -rm -r /user/cloudera/wordcount/output
```

HACK76 – Repetindo a operação com outros dados.

Vamos agora ver se todos entenderam repetindo a operação com algumas alterações, neste exercício quero que use seus conhecimentos adquiridos , não serão passados os passos a passos somente o que deve fazer.

1. Acesse uma pagina de noticias de algo de seu gosto

2. Copie 5 notícias ao menos, e adiciona cada uma em um arquivo diferente (ex. noticia1.txt , noticia2.txt , noticia3.txt , noticia4.txt , noticia5.txt) em seu filesystem no linux.

3. Crie uma nova pasta no HDFS e copie seus arquivos criados

4. Execute o WordCount1 da mesma forma que fizemos agora com arquivos diferentes e com mais dados.

5. Analise os resultados.

HACK77 – Usando o WordCount3.

Temos nos exemplos de WordCount 3 variações, o WordCount3 permite você formar um arquivo com as palavras que não quer considerar em seu WordCount.

1. Crie um arquivo com o nome stop_words.text , adicione palavas que não quer considerar em seu wordcount

```
a
o
,
)
(
.
as
no
na
e
```

2. Copie o arquivo de exclusão para este diretório

```
hadoop fs -mkdir stop_words* /user/cloudera/wordcount/
```

3. Acesse o diretório do WordCount3 no filesystem Linux

```
cd $HOME/hacks-hadoop/hadoop_tutorial/WordCount3
```

4. Execute o WordCount3

```
hadoop jar wordcount.jar org.myorg.WordCount
/user/cloudera/wordcount/input /user/cloudera/wordcount/output3 -skip
/user/cloudera/wordcount/stop_words.text
```

5. Analise os resultados

```
hadoop fs -cat /user/cloudera/wordcount/output2/*
```

HACK78 – Iniciando Eclipse e criando um novo projeto.

1. No seu Desktop existe um link para executar o Eclipse , execute-o
2. Será aberta a IDE de desenvolvimento com alguns fontes , vamos iniciar o desenvolvimento de um programa wordcount simples e do zero.
3. Acesse File → New → Project
4. Selecione a opção Java → Java Project e clique em New
5. Adicione na propriedade “Project Name” o nosso nome de Projeto: WordCount e clique em Next
6. Nasta tela podemos configurar as bibliotecas necessárias para compilar um programa MapReduce, clique na aba “Libraries”

7. Clique no botão “Add external Jars” e selecione o jar **hadoop-core.jar** que pode ser encontrado no diretório /usr/lib/hadoop/client-0.20

   Atenção: precisaremos de diversas outras bibliotecas e vamos adicionar sob demanda a partir de adicição de códigos em nosso programa.

8. Clique em Finish.

HACK79 – Criando a Class WordCount.

1. No package explorer do Eclipse expanda o projeto “WordCount” e clicando em “src” com o botão direito selecione **New → Class**
2. Na propriedade **name** informa : **WordCount**
3. Clique em Finish
4. Abra o arquivo Java criado (WordCount.java)

HACK80 – Implementando a Classes WordCount.

1. Abra o arquivo WordCount.java ele esta com o conteúdo igual a seguir:

```
public class WordCount {

}
```

2. Acompanhe a implementação e faca as copias acompanhando o instrutor

HACK99 – Monitorando os Jobs de MapReduce.

1. Use o “Yarn Recources Management” para monitorar e visualizar todos os Jobs executados em seu cluster.
2. Para acessar use a URL abaixo.

http://quickstart.cloudera:8088/cluster