Bootcamp: Cientista de Dados

Desafio do módulo

Módulo 3	Processamento de Dados Utilizando o Ecossistema Hadoop.

Objetivos

Exercitar os seguintes conceitos trabalhados no Módulo:

- ✓ Iniciar o Apache Hive.
- ✓ Carregar uma base de dados para o HDFS.
- ✓ Executar consultas na base de dados.
- ✓ Analisar os resultados.

Enunciado

Para essa atividade, o aluno deverá assistir atentamente as seguintes aulas, disponíveis no ambiente virtual de aprendizagem:

1. Instalando a máquina virtual.
2. Executando comandos básicos do Ecossistema Hadoop e do HDFS.
3. Prática: importação e manipulação de dados: Manipulação de dados com o Hive.

Atividades

Os alunos deverão desempenhar as seguintes atividades:

1. Iniciar os 5 serviços do Hadoop, por meio do comando `/usr/local/hadoop/sbin/start-all.sh`.

2. Por meio do comando `-mkdir`, criar um diretório chamado `Desafio` no HDFS.

3. Inserir no diretório `Desafio`, o arquivo `covidData.txt`. Esse arquivo se encontra no seguinte endereço do sistema de arquivos do sistema operacional da máquina virtual: `/usr/local/hadoop/Dados`. Para inserir o arquivo, utilize o comando `-put`.

4. Iniciar o Hive em `/usr/local/hive/bin/hive`. Se houver algum erro de Schema, seguir os passos para correção, apresentados no vídeo "Manipulação de dados com o Hive".

5. Criar um *database* chamado `dbDesafio`, por meio do comando `create database`.

6. Em seguida, por meio do comando `create table`, crie uma tabela chamada `DadosCovid`, que armazene os seguintes campos do arquivo `covidData.txt`:

```
dataOcorrencia String
siglaPais String
descPais String
regiao String
novosCasos int
casosAcumulados int
novosObitos int
obitosAcumulados int
```

   i) **Lembre-se que os campos do arquivo covidData.txt estão separados por vírgula e os registros por '\n'.**

   ii) Salve a tabela no seguinte destino do HDFS: `/Desafio`. Use o `STORED AS TEXTFILE LOCATION` para isso.

```
2020-02-26T00:00:00Z,AF,Afghanistan,EMRO,0,1,0,0
2020-02-27T00:00:00Z,AF,Afghanistan,EMRO,0,1,0,0
2020-02-28T00:00:00Z,AF,Afghanistan,EMRO,0,1,0,0
2020-02-29T00:00:00Z,AF,Afghanistan,EMRO,0,1,0,0
2020-03-01T00:00:00Z,AF,Afghanistan,EMRO,0,1,0,0
2020-03-02T00:00:00Z,AF,Afghanistan,EMRO,0,1,0,0
2020-03-03T00:00:00Z,AF,Afghanistan,EMRO,0,1,0,0
2020-03-04T00:00:00Z,AF,Afghanistan,EMRO,0,1,0,0
2020-03-05T00:00:00Z,AF,Afghanistan,EMRO,0,1,0,0
2020-03-06T00:00:00Z,AF,Afghanistan,EMRO,0,1,0,0
2020-03-07T00:00:00Z,AF,Afghanistan,EMRO,3,4,0,0
2020-03-08T00:00:00Z,AF,Afghanistan,EMRO,0,4,0,0
2020-03-09T00:00:00Z,AF,Afghanistan,EMRO,0,4,0,0
2020-03-10T00:00:00Z,AF,Afghanistan,EMRO,0,4,0,0
2020-03-11T00:00:00Z,AF,Afghanistan,EMRO,3,7,0,0
2020-03-12T00:00:00Z,AF,Afghanistan,EMRO,0,7,0,0
2020-03-13T00:00:00Z,AF,Afghanistan,EMRO,0,7,0,0
2020-03-14T00:00:00Z,AF,Afghanistan,EMRO,3,10,0,0
2020-03-15T00:00:00Z,AF,Afghanistan,EMRO,6,16,0,0
```

7. Faça a importação dos dados que estão no HDFS para a nova tabela, usando o comando `LOAD DATA INPATH`.

8. Execute uma sentença SQL que conte todos os registros da tabela `DadosCovid`. Para isso utilize a função `count(*)` do SQL. Anote o resultado.

9. Execute uma sentença SQL que verifique quantas comunicações os países Uruguay e Brazil fizeram **cada um** durante o período de apuração do arquivo. Lembre-se, cada linha do arquivo é uma comunicação. Para isso, utilize a cláusula `where`. Anote o resultado.

10. Execute a seguinte sentença: `select avg(novosCasos) from DadosCovid where descPais = "France";` Anote o resultado.

11. Execute uma sentença que apure quantos novos casos e quantos novos óbitos foram comunicadas no dia 26/05/2020, considerando todos os países. Anote o resultado.

12. Execute o seguinte comando: `describe extended DadosCovid`. Copie o resultado apresentado em tela.

13. Execute a sentença: `select concat(dataOcorrencia, " ", siglaPais, " ", descPais) from DadosCovid where novosCasos = 501;` Anote o resultado.

14. Execute a sentença: `select região, count(1) from DadosCovid group by regiao order by regiao;` Anote os resultados.

Respostas Finais

Os alunos deverão desenvolver a prática e, depois, responder às seguintes questões objetivas: