

❗ O teste foi reavaliado; sua pontuação não foi afetada.

Desafio do Módulo 2

Entrega 12 jun em 17:00	Pontos 40	Perguntas 15
Disponível até 12 jun em 17:00	Limite de tempo Nenhum	

Instruções

Informações sobre o Desafio

O Desafio do Módulo 2 está disponível! Consulte a data de entrega no teste e em seu calendário.

Reserve um tempo para realizar a atividade, leia as orientações e enunciados com atenção. Em caso de dúvidas utilize o "Fórum de dúvidas do Desafio".

Para iniciá-lo clique em "Fazer teste". Você tem somente **uma** tentativa e não há limite de tempo definido para realizá-lo. Caso precise interromper a atividade, apenas deixe a página e, ao retornar, clique em "Retomar teste".

Clique em "Enviar teste" **somente** quando você concluí-lo. Antes de enviar confira todas as questões.

O gabarito será disponibilizado partir de sexta-feira, 12/06/2020, às 21h.

Bons estudos!

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,EMR0,0,1,0,0
2020-02-27T00:00:00Z,AF,Afghanistan,EMR0,0,1,0,0
2020-02-28T00:00:00Z,AF,Afghanistan,EMR0,0,1,0,0
2020-02-29T00:00:00Z,AF,Afghanistan,EMR0,0,1,0,0
2020-03-01T00:00:00Z,AF,Afghanistan,EMR0,0,1,0,0
2020-03-02T00:00:00Z,AF,Afghanistan,EMR0,0,1,0,0
2020-03-03T00:00:00Z,AF,Afghanistan,EMR0,0,1,0,0
2020-03-04T00:00:00Z,AF,Afghanistan,EMR0,0,1,0,0
2020-03-05T00:00:00Z,AF,Afghanistan,EMR0,0,1,0,0
2020-03-06T00:00:00Z,AF,Afghanistan,EMR0,0,1,0,0
2020-03-07T00:00:00Z,AF,Afghanistan,EMR0,3,4,0,0
2020-03-08T00:00:00Z,AF,Afghanistan,EMR0,0,4,0,0
2020-03-09T00:00:00Z,AF,Afghanistan,EMR0,0,4,0,0
2020-03-10T00:00:00Z,AF,Afghanistan,EMR0,0,4,0,0
2020-03-11T00:00:00Z,AF,Afghanistan,EMR0,3,7,0,0
2020-03-12T00:00:00Z,AF,Afghanistan,EMR0,0,7,0,0
2020-03-13T00:00:00Z,AF,Afghanistan,EMR0,0,7,0,0
2020-03-14T00:00:00Z,AF,Afghanistan,EMR0,3,10,0,0
2020-03-15T00:00:00Z,AF,Afghanistan,EMR0,6,16,0,0
```

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 regioao order by regioao; Anote os resultados.

Este teste foi indisponível 12 jun em 17:00.

Histórico de tentativas

	Tentativa	Tempo	Pontuação	Reavaliado
MAIS RECENTE	Tentativa 1	248 minutos	40 de 40	40 de 40

Pontuação deste teste: **40** de 40

Enviado 10 jun em 16:18

Esta tentativa levou 248 minutos.

Pergunta 1

2,67 / 2,67 pts

Quantos registros existem na tabela DadosCovid, criada no Hive?

☐ 8712

☒ 17705

☐ 14719

☐ 138

Correto!

Pergunta 2

2,67 / 2,67 pts

Ao realizar a contagem de registros na tabela DadosCovid, o Hive instanciou um job MapReduce. Qual foi o nome dado a esse job? Para descobrir, acesse a ferramenta administrativa do Hadoop.

☐ Contagem de elementos

☒ select count(*) from DadosCovid (Stage-1)

Correto!

☐ select count(*) from DadosCovid

☐ DadosCovid

Pergunta 3

2,67 / 2,67 pts

Quantas comunicações foram realizadas pelo país Uruguay?

Correto!

☒ 74

☐ 82

☐ 92

☐ 101

Pergunta 4

2,67 / 2,67 pts

Quantas comunicações foram realizadas pelo país Brazil?

Correto!

☐ 118

☒ 92

☐ 101

☐ 74

Pergunta 5

2,67 / 2,67 pts

Ao inserir o comando:

```
select avg(novosCasos) from DadosCovid where descPais = "France";
```

Qual é o resultado apresentado pelo Hive?

☐ 1147.5

☐ 0

☐ 64.19

☒ 1141.632

Correto!

Pergunta 6

2,67 / 2,67 pts

Considerando todos os países, quantos novos casos foram informados no dia 26/05/2020?

☒ 99023

☐ 1493

☐ 18829

☐ 1418

Correto!

Pergunta 7

2,67 / 2,67 pts

Considerando todos os países, quantos novos óbitos foram informados no dia 26/05/2020?

☐ 1357

☐ 16899

☒ 1493

☐ 99023

Correto!

Pergunta 8

2,67 / 2,67 pts

O que o Hive faz quando executa a sentença: `select count(1) from DadosCovid`?

☐

Retorna o número total de registros diretamente, sem executar funções Map e Reduce.

☒

Retorna o número total de registros, após executar funções Map e Reduce.

☐

Retorna a média de linhas.

☐

Retorna 1.

Correto!

Pergunta 9

2,67 / 2,67 pts

É correto afirmar que, quando utilizamos o comando `describe extended DadosCovid`, os seguintes dados são apresentados na tela:

Correto!

- ☐ count e avg
- ☒ line.delim e field.delim.
- ☐ count e distinct
- ☐ Map e Reduce.

Question 10

Pontuação original: 2,67 / 2,67 pts **Pontuação reavaliada: 2,67 / 2,67 pts**

⚠ Esta pergunta foi reavaliada.

QUESTÃO ANULADA

Usando o comando `describe extended DadosCovid`, podemos encontrar os `FieldSchemas` dos campos da tabela `DadosCovid`. O `FieldSchema` do campo `casosacumulados` está representado da seguinte forma:

Correto!

- ☒ `FieldSchema(name:obitosacumulados, type:int, comment:null)`
- ☐ `FieldSchema(name:obitosacumulados, type:String, comment:null)`
- ☐ `FieldSchema(name:obitosacumulados, type:int, comment:"Obitos acumulados até o momento")`
- ☐ O campo `casosacumulados` não possui `FieldSchema`.

Pergunta 11**2,67 / 2,67 pts**

Marque a alternativa que melhor representa o resultado do seguinte comando:

```
select concat(dataOcorrencia, " ", siglaPais, " ", descPais) from  
DadosCovid where novosCasos = 501;
```

Correto!

- ☒ 2020-03-11T00:00:00Z ES Spain
- ☐ 2020-03-14T00:00:00Z ES Spain
- ☐ 2020-05-07T00:00:00Z UA Ukraine
- ☐ 2020-04-13T00:00:00Z JP Japan

Pergunta 12**2,67 / 2,67 pts**

São características do Apache Hive, EXCETO:

Correto!

- ☐ O Hive transpila consultas SQL como jobs MapReduce.
- ☐ O Hive é uma estrutura de Data Warehouse.
- ☒ É uma ferramenta construída para transferir dados entre o cluster Hadoop e um banco de dados SQL.
- ☐ O Hive é construído sobre o Ecossistema Hadoop, utilizando o HDFS para armazenamento.

Pergunta 13**2,67 / 2,67 pts**

Quantos registros de comunicação foram feitos para a região EURO?

Correto!☒ 5431☐ 3417☐ 5341☐ 1816**Pergunta 14****2,67 / 2,67 pts**

Marque a opção que representa o número de jobs MapReduce que foram criados ao executar a sentença:

```
select região, count(1) from DadosCovid group by regioao order by regioao;
```

Correto!☒ 2☐ 3☐ 0☐ 1**Pergunta 15****2,62 / 2,62 pts**

Marque a alternativa que NÃO representa comandos do Hive:

Correto!☐ Show Databases e Describe Extended.☐ Show Tables e Describe.☒ HDFS e YARN.☐ Load Data Inpath e Create Table.Pontuação do teste: **40** de 40