Desafio do Módulo 2

Entrega 29 set em 23:59 **Pontos** 40 **Perguntas** 15
Disponível até 29 set em 23:59 **Limite de tempo** Nenhum
Tentativas permitidas 2

Instruções

O Desafio do Módulo 4 está disponível!

1. Instruções para realizar o desafio

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 do Módulo 4".

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.

Caso o teste seja iniciado e não enviado até o final do prazo de entrega, a plataforma enviará a tentativa não finalizada automaticamente, independente do progresso no teste. Fique atento ao seu teste e ao prazo final, pois novas tentativas só serão concedidas em casos de questões médicas.

O gabarito será disponibilizado partir de quinta-feira, **29/09/2022,** às 23h59.

Bons estudos!

2. O arquivo abaixo contém o enunciado do desafio

Enunciado do Desafio – Módulo 2 – Engenheiro(a) de Dados Cloud.pdf

Este teste foi indisponível 29 set em 23:59.

Histórico de tentativas

	Tentativa	Tempo	Pontuação
MAIS RECENTE	Tentativa 1	9.415 minutos	40 de 40

Pontuação desta tentativa: **40** de 40
Enviado 26 set em 9:23
Esta tentativa levou 9.415 minutos.

Pergunta 1

2,67 / 2,67 pts

A melhor forma de reparticionar o DataFrame “df” para 25 partições, dado que ele tem 100, é

Correto!

- (●) df.coalesce(25).
- ( ) df.repartition(25).
- ( ) df.partitionBy(25).
- ( ) df.partition(25).

Pergunta 2

2,67 / 2,67 pts

O Spark dispõe de pelo menos três algoritmos diferentes para realizar joins: Broadcast Join (BHJ) Shuffled Hash Join (SHJ) Sort Merge Join (SMJ). Suponha que tenhamos duas tabelas, df1 e df2, e que durante o processamento devemos realizar uma operação de join entre elas. Considere que df1 é consideravelmente maior que df2, que nenhuma outra operação é realizada anterior ao join e que id é uma coluna ordenável que serve como chave da junção. Observação: os dados de ambas as tabelas estão particionados.

- ( ) Deve-se utilizar o Sort Merge Join em todas as situações.
- ( ) Salvar as tabelas em buckets particionados e ordenados pela coluna id e realizar um Sort Merge Join.

Correto!

- (●) Broadcast Join, se a tabela menor for pequena o suficiente, e Shuffled Hash Join caso contrário.
- ( ) Shuffled Hash Join, se a tabela menor for pequena o suficiente, e Broadcast Join caso contrário.

Pergunta 3

2,67 / 2,67 pts

A diferença entre os modos de deploy “client” e “cluster” são:

Correto!

- (●) No modo “cluster” tanto o driver como os executores do Spark estão presente em algum dos nós do cluster, enquanto no modo “client” o driver fica na máquina que enviou a aplicação e os executores presentes nos nós.
- ( ) No modo “cluster” o driver apenas se comunica com um manager existente, enquanto no modo “client” o driver do Spark faz o papel de cluster manager.
- ( ) No modo “cluster” o driver do Spark faz o papel de cluster manager, enquanto no modo “client” o driver apenas se comunica com um manager existente.
- ( ) No modo “cluster” o driver fica na máquina que enviou a aplicação e os executores presentes nos nós do cluster, enquanto no modo “client” tanto o driver como os executores do Spark estão presentes nos nós.

Pergunta 4

2,67 / 2,67 pts

A vantagem de se utilizar a persistência de dados na memória é:

- ( ) O plano físico gerado pelo Catalyst Optimizer é otimizado pela presença física dos dados em disco.

Correto!

- [x] É eliminada a necessidade de realizar as operações salvas pelo Lazy Evaluation todas as vezes em consultas futuras aos dados.
- [ ] As operações salvas pelo Lazy Evaluation são executadas de forma mais rápida em consultas futuras aos dados.
- [ ] Não há vantagens significativas de se utilizar a persistência em memória.

Pergunta 5

2,67 / 2,67 pts

Sobre o Spark SQL, podemos afirmar que

- [ ] Ele é exclusivo em relação a API de DataFrames, e ambas as formas de manipulação não interagem entre si.
- [ ] Ele dispõe de diversas funções implementadas que podem ser usadas, mas não permite o registro de udfs.
- [ ] Ele só pode ser usado em um shell criado pelo comando spark-sql.

Correto!

- [x] Ele é intercambiável em relação a API de DataFrames, e ambas as formas de manipulação estão relacionadas.

Pergunta 6

2,67 / 2,67 pts

Qual código abaixo é usado para gerar **tabelas temporárias** no Spark?

- [ ] df.createOrReplaceTempTable(“table_name”).

- spark.sql("CREATE TEMPORARY TABLE table_name (schema)").
- df.write.saveAsTempTable("table_name").

Correto!

- (●) Não há tabelas temporárias no Spark.

Pergunta 7

2,67 / 2,67 pts

Para limpar **todos** os dados persistidos na memória, deve-se usar:

- spark.catalog.unpersistAll()
- spark.catalog.clearTables()
- spark.catalog.unpersistTables()

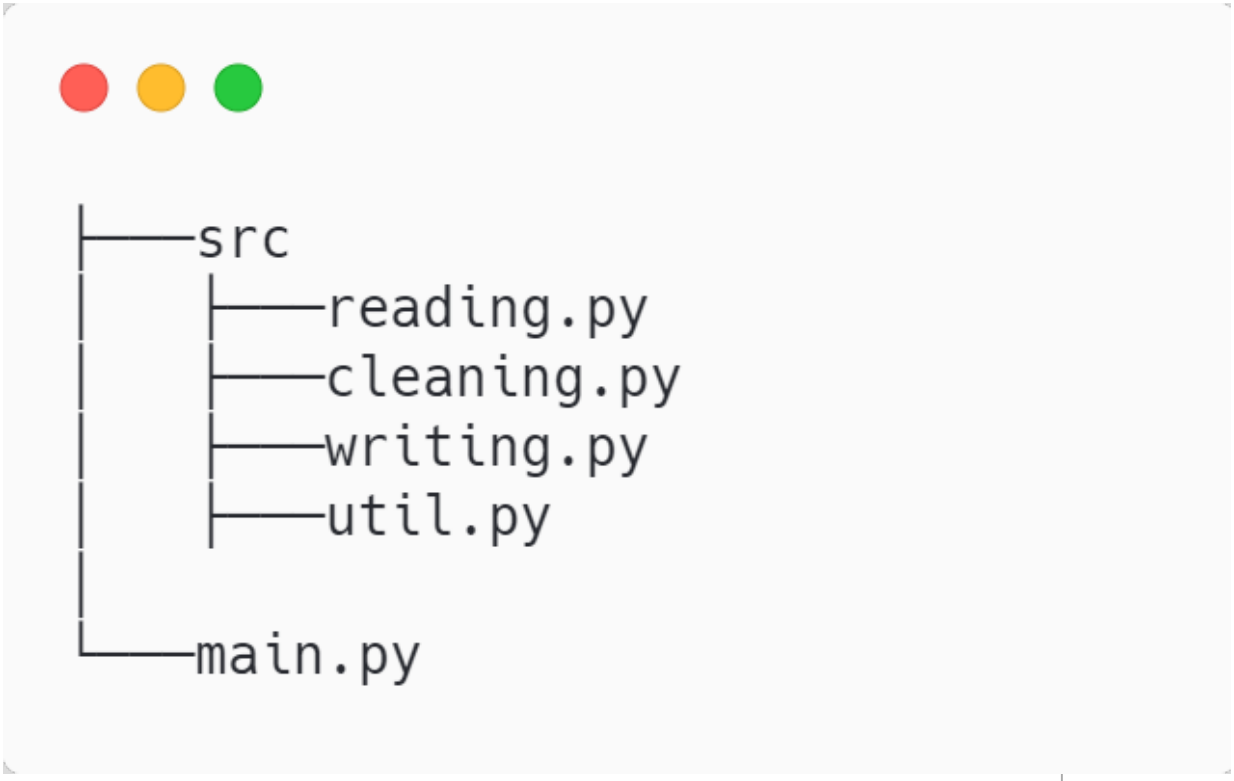
Correto!

- (●) spark.catalog.clearCache()

Pergunta 8

0 / 2,67 pts

Uma aplicação do Spark foi desenvolvida em uma estrutura composta por cinco scripts Python. Os scripts estão organizados na seguinte estrutura:



Qual é o comando spark-submit adequado para deploy dessa aplicação localmente?

esposta correta

- spark-submit --py-files src.zip main.py
- spark-submit –py-files main.py src.zip
- spark-submit --py-files src.zip src/main.py

ocê respondeu

- spark-submit –py-files src.zip,main.py

ão respondida

ergunta 9

0 / 2,67 pts

Qual das opções de leitura abaixo fornece a leitura adequada da tabela de estabelecimentos? Considere que o schema está criado e a variável abaixo:

path = 'gs://desafio-final/estabelecimentos/*'

Obs.: utilize a opção de “escape” apresentada abaixo para a leitura dos dados na resolução das demais questões.

esposta correta

```
df = (
  spark.read
  .format('csv')
  .option('encoding', 'ISO-8859-1')
  .option('sep', ';')
  .option("escape", "\"")
  .schema(schema)
  .load(path)
)
```

```
df = (
  spark.read
  .format('csv')
  .options('encoding', 'ISO-8859-1')
  .options('sep', ';')
  .options("escape", "\"")
  .schema(schema)
  .load(path)
)
```

```
options_dict = {
  'encoding':'ISO-8859-1',
  'sep':';',
  "escape": "\"",
  'format':'csv'
}

df = (
  spark.read
  .option(**options_dict)
  .schema(schema)
  .load(path)
)
```

```
options_dict = {
  'encoding':'ISO-8859-1',
  'sep':';',
  "escape": "\"",
  'format':'csv'
}

df = (
  spark.read
  .options(**options_dict)
  .schema(schema)
  .load(path)
)
```

Pergunta 10

2,67 / 2,67 pts

Qual o código do CNAE mais presente nas empresas ativas? Quantas empresas utilizam esse CNAE?

- CNAE 4781400, 2.415.486 empresas.
- CNAE 4781400, 991.316 empresas.
- CNAE 4772500, 825.859 empresas.
- (Correto!) CNAE 4781400, 1.781.558 empresas.

Pergunta 11

2,67 / 2,67 pts

Usando a função to_date(), qual a forma CORRETA do argumento format para converter as colunas de data para o tipo correto?

- (Correto!) yyyyMMdd
- dd/MM/yyyy
- ddMMyyyy
- yyyy-MM-dd

Pergunta 12

2,67 / 2,67 pts

Quantos CNPJs não ativos existem no estado de São Paulo?

- 14.088.503
- (Correto!) 7.966.472
- 6.020.626
- 6.122.031

Pergunta 13

2,67 / 2,67 pts

Quantas empresas de “Consultoria em tecnologia da informação” existem em Belo Horizonte?

- 8957.
- Correto! 6930.
- 13891.
- 5754.

Pergunta 14

2,67 / 2,67 pts

Qual o CNAE primário do IGTI?

Dica: O IGTI está localizado em Belo Horizonte.

- Correto! 8532500, Educação superior – graduação e pós-graduação.
- 8531700, Educação superior – graduação.
- 8542200, Educação profissional de nível tecnológico.
- 8533300, Educação superior – pós-graduação e extensão.

Pergunta 15

2,62 / 2,62 pts

Quantas empresas foram abertas desde 2020?

- 6.312.750.
- 1.036.664.

Correto!

- (●) 6.314.456.
- ( ) 5.277.792.

Pontuação do teste: **40** de 40

A pontuação deste teste foi ajustada manualmente em +5,34 pontos.