DESM2 - Desafio do Módulo 2

Entrega 13 fev em 21:00 **Pontos** 40 **Perguntas** 15 **Disponível** até 13 fev em 21:00
Limite de tempo Nenhum

Instruções

O Desafio do Módulo 2 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 2".

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 a partir de sexta-feira, **17/02/2023,** às 23h59.

Bons estudos!

2. O arquivo abaixo contém o enunciado do desafio

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

Histórico de tentativas

	Tentativa	Tempo	Pontuação
MAIS RECENTE	Tentativa 1	59 minutos	40 de 40

As respostas corretas estarão disponíveis em 17 fev em 23:59.

Pontuação deste teste: **40** de 40
Enviado 6 fev em 21:23
Esta tentativa levou 59 minutos.

Pergunta 1

2,67 / 2,67 pts

Quantos estabelecimentos existem?

- (●) 20996744
- ( ) 4753429
- ( ) 20996747
- ( ) 4753426

Pergunta 2

2,67 / 2,67 pts

Na tabela de estabelecimentos, quantas colunas existem e quantas são identificadas pelo spark como números? (Use *inferSchema* ao ler os arquivos).

- ○ 30 e 30.
- ○ 30 e 0.
- ○ 30 e 12.
- ◉ 30 e 13.

Pergunta 3

2,67 / 2,67 pts

O formato Parquet **(https://parquet.apache.org/ (https://parquet.apache.org/))** é um formato de armazenamento útil em engenharia de dados projetado para armazenamento e recuperação dos dados eficiente. Ele usa mecanismos de compressão de dados para que os dados ocupem menos espaço.

Usando estabelecimentos_df.write.parquet("estabelecimentos.parquet") , compare o tamanho do(s) arquivo(s) parquet com os arquivos CSV originais. A economia de espaço foi da ordem de:

- ○ ~2 vezes menos espaço
- ○ ~ 1,5 vezes menos espaço

~ 5 vezes menos espaço

~ 2,5 vezes menos espaço

Pergunta 4

2,67 / 2,67 pts

Vamos usar Spark SQL para obter algumas contagens. Primeiro, vamos ver quantos estabelecimentos não tem logradouro cadastrado.

Use

estabelecimentos_df.createTempView("estabelecimentos")

para criar uma tabela temporária de logradouros e depois execute uma consulta SQL como “SELECT COUNT(*) FROM estabelecimentos WHERE LOGRADOURO IS NULL” para contar quantos estabelecimentos não tem logradouro informado. Quantos são?

0

343

828

919

Pergunta 5

2,67 / 2,67 pts

Em muitos casos, as UDFs (funções definidas pelo usuário) são uma forma muito conveniente de implementar uma lógica dentro de uma função python e chamá-la dentro de uma consulta SQL. Vamos fazer isso para computar quantos estabelecimento têm um logradouro cujo endereço é uma AVENIDA.

Crie uma função em Python “def is_avenida(logradouro):” que recebe um logradouro e decide se ele é uma avenida ou não. Dica: Use *startswith*() e fique atento ao resultado da questão anterior, pois ele pode impactar a lógica da sua função. Esteja atento a maiúsculas e minúsculas.

Em seguida, use Spark SQL para executar uma consulta como "SELECT COUNT(*) FROM estabelecimentos WHERE is_avenida(LOGRADOURO) == True” para contar quantos estabelecimentos ficam localizados em uma avenida.

- (●) 52587
- ( ) 0
- ( ) 20944157
- ( ) 219837

Pergunta 6

2,67 / 2,67 pts

Quantos CEPs distintos existem entre os estabelecimentos?

- 29129.
- 788134.
- 889886
- 238212

Pergunta 7

2,67 / 2,67 pts

Quantos CNAEs existem na tabela de CNAES?

- 3159
- 1359
- 9135
- 5319

Pergunta 8

2,67 / 2,67 pts

Vários CNAEs são de cultivo. Quantos estabelecimentos possuem um CNAE relacionado a cultivo?

Dicas: use a operação JOIN para criar uma dataframe que tem os estabelecimentos e as descrições de seus CNAES. Crie uma UDF que verifica se a descrição do CNAE é sobre cultivo; e usar Spark SQL para escrever uma consulta como “SELECT COUNT(*) from estabelecimentos_with_cnae WHERE is_cnae_cultivo(DESCRICAO_CNAE) == True”.

- ○ 11100
- ○ 95311
- ○ 32100
- ◉ 200243

Pergunta 9

2,67 / 2,67 pts

Todas estes provedores de nuvem oferecem serviços gerenciados para uso do Apache Spark, EXCETO:

- ○ Nuvem da Microsoft (Azure).
- ◉ Nenhuma opção é verdadeira.
- ○ Nuvem da Amazon (AWS).
- ○ Nuvem do Google (GCP).

Pergunta 10

2,67 / 2,67 pts

Qual dessas características se aplica à API de Dataframes do Spark, mas não à API de RDDs?

- imutabilidade.
- executa a computação de forma distribuída, explorando os recursos de um cluster.
- computação é “preguiçosa” (lazy).
- lida com dados estruturados.

Pergunta 11

2,67 / 2,67 pts

Quantos estabelecimentos são filiais? Consulte o dicionário dos dados e use Spark SQL.

- 0
- 19903662
- 1093079

1093082

Pergunta 12

2,67 / 2,67 pts

Qual dessas tarefas você diria que NÃO está no escopo de engenharia de dados?

- Definição da estrutura de armazenamento dos dados.
- Criação de modelos de inteligência artificial.
- Criação de fluxos de ETL.
- Normalização dos dados.

Pergunta 13

2,67 / 2,67 pts

Caso você queira construir uma aplicação que lê mensagens do Twitter e utiliza um modelo de aprendizado de máquina para classificar o conteúdo dos tweets entre carregando sentimentos “positivos” e “negativos”, quais componentes do Spark você utilizaria?

- Spark SQL e Spark GraphX.

- Spark GraphX e Spark ML.
- Spark Streaming e Spark GraphX.
- Spark Streaming e Spark ML.

Pergunta 14

2,67 / 2,67 pts

Quais formatos o Spark consegue NÃO possuir uma API para ler de forma nativa?

- CSV.
- Parquet.
- XML.
- JSON.

Pergunta 15

2,62 / 2,62 pts

O Apache Spark é tudo isso, EXCETO…

- [x] … um sistema de arquivos distribuído.
- [ ] … um sistema de código aberto.
- [ ] … uma plataforma para big data.
- [ ] … um framework.

Pontuação do teste: **40** de 40