Trabalho Prático do Módulo 2

Entrega 26 set em 23:59 **Pontos** 25 **Perguntas** 15
Disponível até 26 set em 23:59 **Limite de tempo** Nenhum

Instruções

O Trabalho Prático do Módulo 2 está disponível!

1. Instruções para realizar o trabalho prático

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 Trabalho Prático 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 partir de sexta-feira, **26/09/2022,** às 23h59.

Bons estudos!

2. O arquivo abaixo contém o enunciado do trabalho prático

Enunciado do Trabalho Prático - Módulo 2 - Engenheiro(a) de Dados Cloud.pdf

Este teste foi indisponível 26 set em 23:59.

Histórico de tentativas

	Tentativa	Tempo	Pontuação
MAIS RECENTE	Tentativa 1	1.487 minutos	25 de 25

Pontuação deste teste: **25** de 25
Enviado 20 set em 21:12
Esta tentativa levou 1.487 minutos.

Pergunta 1

1,67 / 1,67 pts

Qual é a característica do Apache Spark em relação ao tratamento das computações intermediárias que faz com que ele seja mais rápido do que o Hadoop MapReduce?

Correto!

- (●) Armazenamento dos resultados intermediários em memória.
- ( ) Utilização do lazy evaluation na realização das operações.
- ( ) Reorganização das computações no plano físico de execução.
- ( ) Realização de shuffles de dados.

Pergunta 2

1,67 / 1,67 pts

Qual a principal diferença entre as transformações narrow e wide?

- ( ) As transformações narrow não ativam o histórico do lazy evaluation, enquanto as transformações wide ativam.

Correto!

- (●) As transformações narrow não realizam shuffle de dados, enquanto as transformações wide realizam.
- ( ) As transformações wide não ativam o histórico do lazy evaluation, enquanto as transformações narrow ativam.
- ( ) As transformações wide não realizam shuffle de dados, enquanto as transformações narrow realizam.

Pergunta 3

1,67 / 1,67 pts

Qual dos blocos de código abaixo realiza o empilhamento de dois DataFrames **df1** e **df2?**

Correto!

- (●) df1.union(df2).
- ( ) df1.concat(df2).
- ( ) df1.append(df2).
- ( ) df1.add(df2).

Pergunta 4

1,67 / 1,67 pts

Em que momento do Catalyst Optimizer são realizadas otimizações nas operações de processamento?

- ( ) Geração de Código.
- ( ) Plano Físico.

Correto!

- (●) Plano Lógico.
- ( ) Análise.

Pergunta 5

1,67 / 1,67 pts

Sobre colunas e expressões, pode se dizer que:

Colunas são objetos manipuláveis de um DataFrame ou Dataset que servem para armazenamento de dados e realização de operações, enquanto as expressões são usadas para definir um schema.

Correto!

Colunas e expressões proporcionam duas formas diferentes de se realizar operações com colunas em um DataFrame ou Dataset, mas são equivalentes.

Colunas e expressões proporcionam duas formas diferentes de se realizar operações com colunas em um DataFrame ou Dataset, mas não são equivalentes.

Colunas são objetos não manipuláveis de um DataFrame ou Dataset que servem somente para armazenamento de dados, enquanto as expressões são usadas para realizar operações sobre os dados nas colunas.

Pergunta 6

1,67 / 1,67 pts

Quantos **filmes** (incluindo os da televisão) foram lançados no ano de 2015?

356877.

3558.

Correto!

19987.

16429.

Pergunta 7

1,67 / 1,67 pts

Qual o gênero de títulos mais frequente?

Dica: Utilize as funções split e explode.

- War.
- Drama. (Correto!)
- Comedy.
- Documentary.

Pergunta 8

1,67 / 1,67 pts

Qual o gênero com a melhor nota média de títulos?

- Documentary.
- Drama.
- History. (Correto!)
- Crime.

Pergunta 9

1,67 / 1,67 pts

Qual o vídeo game do gênero aventura mais bem avaliado em 2020?

- Half-Life: Alyx. (Correto!)
- Omori.
- Ghost of Tsushima.

- Final Fantasy VII Remake.

Pergunta 10

1,67 / 1,67 pts

Qual seria a forma mais adequada de preencher dados nulos da coluna “col1” com o valor da coluna “col2”?

Correto!

- df.withColumn(“col1”, coalesce(“col1” , “col2”)).
- df.fillna(col(“col2”), subset = [“col1”]).
- df.na.fill(col(“col2”), subset = [“col1”]).
- df.replace(None, col(“col2”), subset = [“col1”]).

Pergunta 11

1,67 / 1,67 pts

Dos títulos lançados em 2018, qual o **percentual** daqueles pertencentes ao gênero comédia?

Dica: Utilize as funções Split, explode e uma window function.

- 15,4%.
- 22,3%.

Correto!

- 19,6%.
- 18.7%.

Pergunta 12

1,67 / 1,67 pts

Qual das opções a seguir NÃO é uma vantagem de se utilizar o formato parquet para salvar dados?

- Preservação de metadados, incluindo tipos complexos.

Correto!

- (selected) Preservação de metadados, com exceção de tipos complexos.
- Compressão de dados na escrita.
- Armazenamento colunar.

Pergunta 13

1,67 / 1,67 pts

Deseja-se utilizar um join para retornar somente as linhas referentes a títulos que estão **sem nota,** isto é, **não aparecem no df_ratings**. Escolha o extrato de código que permite que isso seja feito:

Correto!

- (selected) df_titles.join(df_ratings, “tconst”, “anti”).
- df_titles.join(df_ratings, “semi”, “tconst”).
- df_titles.join(df_ratings, “anti”, “tconst”).
- df_titles.join(df_ratings, “tconst”, “semi”).

Pergunta 14

1,67 / 1,67 pts

Considere a definição de uma udf abaixo:

```
def sqr_divide(value):
    return (value**2)/2
sqr_divide_udf = udf(sqr_divide, IntegerType())
```

A definição de sqr_divide_udf possui um problema. Depois de solucionar o problema, ao executar

```
(
    df_ratings
    .withColumn('averageRating', f.col('averageRating').cast('double'))
    .select(sqr_divide_udf('averageRating').alias('averageRating'))
    .agg(f.mean('averageRating').alias('averageRating'))
    .show()
)
```

o que retorna é:

Obs.: Considere 3 casas decimais.

- Valor nulo ou erro.
- 20.744.
- **Correto!** 24.882.
- 6.915.

Pergunta 15

1,62 / 1,62 pts

Qual das seguintes operações NÃO é uma ação?

- save().
- **Correto!** agg().
- count().
- toPandas().

Pontuação do teste: **25** de 25