DataAnalytics_test_Respostas

July 19, 2021

1 ST IT Cloud - Data and Analytics Test LV.4

Esse teste deve avaliar alguns conceitos de big data e a qualidade técnica na manipulacão de dados, otimização de performance, trabalho com arquivos grandes e tratamento de qualidade.

1.1 Passo a passo

- *Parte teórica:* responda as questões abaixo preenchendo as células em branco.
- *Parte prática:* disponibilizamos aqui 2 cases para, leia os enunciados dos problemas, desenvolver os programas, utilizando a **stack definida durante o processo seletivo**, para entregar os dados de acordo com os requisitos descritos abaixo.

Faz parte dos critérios de avaliacão a pontualidade da entrega. Implemente até onde for possível dentro do prazo acordado.

Os dados de pessoas foram gerados de forma aleatória, utilizando a biblioteca FakerJS, FakerJS-BR e Faker

LEMBRE-SE: A entrega deve conter TODOS os passos para o avaliador executar o programa (keep it simple).

Questão 1 - Descreva de forma detalhada quais são as etapas na construção de um pipeline de dados, sem considerar ferramentas específicas, imagine que é seu primeiro contato com o cliente e você precisa entender a demanda dele e explicar quais são os passos que você terá que implementar para entregar a demanda.

1- Avaliar o Problema de Negocio 2- Avaliar as possiveis soluções e abordagens para resolver o problema 3- Escolher um caminho a seguir entre as abordagens 5- Desenhar a Solução 4- Definir quais Ferramentas serão utilizadas 5- Desenvolver um prototipo inicial 6 - Fazer as Validações internas e testes com o Usuario 7- Fazer Deploy em Produçao, é um processo que tem um inicio , meio e fim , ou do ponto de vista dos dados tem uma entrada , processamento e saida ou seja é a montagem de um fluxo operacional a ser seguido que tem um inicio e um fim em termos operacionais , e que pode ser aplicado a qualquer coisa , desde de a um gerenciamento de produção de uma fabrica que fabrica determinado produto , ou a uma Fabrica de Sottware que desenvolve aplicativos ou mesmo a uma arquitetura de fluxo operacional de dados, ou seja se trata de gerenciamento organizado de fluxo operacional de trabalho.

Questão 2 - Defina com suas palavras um processamento em streaming e processamento em batch. Qual sua experiência com cada uma delas.

Processamento em Streaming é o processamento em tempo real ou seja no momento da produção da informação (como por exemplo a coleta de uma informação de venda online que acabou de acontecer), processamento em batch é um processamento através de lotes ou seja é enviado lotes de dados com tamanhos pre-definidos para processamento em intervalos de tempo pre definidos e pode ser muito tempo depois do fato ter ocorrido ou pode ser proximo ao tempo real.

Questão 3 - Quais são as camadas de um Data Lake?

Ingestão , Armazenamento em Cache e Processamento

Questão 4 - Quais as diferenças de um Data Lake e um DW?

Basicamente o Data Lake suporta dados não estruturados e o Data Warehouse não suporta , mas ambos suportam dados estruturados e agregados por segmentos e tambem umm Data Lake por ser projetado para Big Data suporta armazenamento e processamento distribuido com volumes de dados gigantescos e alta integridade das informações ou seja atende ao requisito de grandes volumes de dados com alta integridade e tambem ao requisito de processamento de dados em alta velocidade, pois trabalha com armazenamento e processamento distribuido , operando através de um cluster de computadores como se fosse uma unica maquina , coisa que um DW não comporta pois não foi projetado para isso.

Questão 5 - O que é arquitetura Lambda e Kappa? Descreva com suas palavras.

Arquitetura Lambda Foi desenvlvida para contornar uma limitação do MapReduce que ao processar grandes volumes de dados demorava muito para dar um retorno do resultado. Isso foi resolvido pelo menos parcialmente ,através da arquiteura Kappa que dá dois caminhos para o resultado dos dados , o caminho quente que é mais rapido mas menos preciso e em uma janela pequena de tempo e o caminho frio que é mais lento e mais demorada e em uma janela de tempo maior mas é mais preciso em relação as informações, este é processado em Batch , tudo é unificado na camada de serviço que possui as informações completas unificadas entre os processamento em streeaming de dados e o processamento em batch , os dados dos eventos são imutaveis.Dessa forma acaba atendendo a necessidade de ver mais rapidamente os dados. A Arquitetura Kappa foi desenvolvida para suprir uma deficiencia da arquitetura Lambda pois havia a necessidade de dois codigos diferentes para tratar a camada em batch e a camada em streming , ja a Arquitetura Kappa pensa em termos de unificação dos dados na origem e para isso usa o Log para fazer essa unificação onde não importa de onde venha o dado , se de streaming de dados , de um csv , xml de um banco de dados , todas essas fontes de dados são tratados como eventos no arquivo de log e é um log imutavel todo o dado que entra entra como um simbolo de + e os updates e reduções de dados entram com um simbolo de - , e assim o seu processamento se torna muito mais rapido, e dessa forma tudo é processado em tempo real e armazenado em um repositório de Longo Termo e possui uma camada de serviço que serve para oferecer os dados para as ferramentas de visualização ou de extração de informações que foram processadas.

Questão 6 - O que é Data Quality para você e como você implementa isso nos seus processos?

É um processo de que visa validar os dados afim de manter a qualidade dos mesmos . Existe varias formas de fazer isso , como processos que validam os nomes das colunas, os tipos de dados e que comparam dados de periodos diferentes para detectar anomalizas ou falhas na produção da informação na origem dos dados

Questão 7 - Em uma escala de 0 a 10, qual seria seu nível de experiência com PySpark?

[ ]:
```
6, utilizei no curso de Python com Spark que fiz.
```

Questão 8 - Em uma escala de 0 a 10, qual seria seu nível de experiência com SQL?

[ ]:
```
8 , já trabalho com a linguagem SQL e banco SQL a muito tempo
```

Questão 9 - Descreva suas expeciências com banco de dados SQL e NoSQL.

Ja trabalhei com o Banco SQL e atualmente Trabalho com Impala banco NoSQL usando a ferramenta Hue para rodar as consultas,tendo tambem o Hive como opção adicional.

Questão 10 - Tem experiência com versionamento de código? Com quais ferramentas já trabalhou? Descreva.

Só conheço GitHub , utilizamos na empresa que trabalho , mas depende muito do projeto , tem projeto que usamos mais , outros usamos menos e outros não usamos.

Questão 11 - Tem experiência em desenvolvimento em cloud? Se sim, especifique a(s) plataforma(s) que já trabalhou e suas principais implementações e conhecimentos em cada serviço.

Não tenho , a unica ferramenta que conheço que trabalha com este Viés é o Azure Machine Learning que aprendi ela no curso que fiz de Cientista de Dados na DSA.

Questão 12 - Tem experiência com metodologia ágil? Qual?

Sim , A maioria dos projetos que trabalhei e que estou trabalhando na Cargill utiliza a metodologia Scrum , então temos o Scrum Master que conduz a reunião em conjunto com os lideres do projeto normalmente alguem de negócio e alguem da area de TI. São realizadas reuniões diarias denominadas Dailys e as Plenys a cada 15 dias e mais uma reunião Geral no Final da Sprint e tambem pode ser marcada uma reunião após o termino da sprint de retrospectiva para analise de pontos de dificuldades e o que pode ser melhorado para a proxima Sprint, cada Sprint são atividades de trabalhos vinculado a um prazo , são determinada metas de trabalhos e delegado as atividades aos responsaveis que terão como meta concluir as atividades até o final da Sprint, o prazo da Sprint pode varias em relação ao tempo , normalmente 15 ou 30 dias, é o que eu conheço. Mas sei que a Metodologia Scrum é bem maleavel então esta estrutura pode variar um pouco de empresa para empresa.

2 TESTE PRÁTICO

Problema 1: Você está recebendo o arquivo 'dados_cadastrais_fake.csv' que contem dados cadastrais de clientes, mas para que análises ou relatórios sejam feitos é necessário limpar e normalizar os dados. Além disso, existe uma coluna com o número de cpf e outra com cnpj, você precisará padronizar deixando apenas dígitos em formato string (sem caracteres especiais), implementar uma forma de verificar se tais documentos são válidos sendo que a informação deve se adicionada ao dataframe em outras duas novas colunas.

Após a normalização, gere reports que respondam as seguintes perguntas: - Quantos clientes temos nessa base? - Qual a média de idade dos clientes? - Quantos clientes nessa base pertencem a cada estado? - Quantos CPFs válidos e inválidos foram encontrados? - Quantos CNPJs válidos e inválidos foram encontrados?

Ao final gere um arquivo no formato csv e um outro arquivo no formato parquet chamado (problema1_normalizado), eles serão destinados para pessoas distintas.

EXTRA: executar as mesmas validações no *1E8.csv.gz

[3]:
```
# Instalando e Importando Bibliotecas e Modulos
# !pip install --user pandas -U
import pandas as pd
import sys
import os.path

#Importa Modulo Adicional
import validacpfcnpj

#Desativa os avisos
import warnings
warnings.filterwarnings("ignore")
```

[4]:
```
# Cria funcao secundaria

def roda_validacao_cpf_cnpj(a):

    cpf_cnpj = validacpfcnpj.ValidaCpfCnpj(str(a))


    return cpf_cnpj.valida()
```

[520]:
```
#Testa Função com CPF - True
roda_validacao_cpf_cnpj(97566536800)
```

[520]: True

[521]:
```
#Testa Função com CPF - False
roda_validacao_cpf_cnpj(97566536801)
```

[521]: False

[522]:
```
#Testa Função com CNPJ - True
roda_validacao_cpf_cnpj("06589184909526")
```

[522]: True

[523]:
```
#Testa Função com CNPJ - false
roda_validacao_cpf_cnpj("06589184909521")
```

[523]: False

[16]:
```
# Usando o método read_csv
df =pd.read_table('dados_cadastrais_fake.csv', sep = ';',encoding="utf8")
df
```

[16]:

	nomes	idade	cidade	estado	cpf	cnpj
0	Dennis Daniels	31	ACRELÂNDIA	AC	97566536800	06589184909526
1	Leah Becker	42	ÁGUA BRANCA	AL	425.263.807-07	25.673.336/2350-20
2	Sally Ford	18	ALVARÃES	AM	34647754103	26543101702989
3	Colleen Duncan	21	SERRA DO NAVIO	AP	252.531.560-03	19.062.080/5100-98
4	Jeff Stephenson	73	ABAÍRA	BA	49668886542	97794530015384
...	...	...	...	...	...	...
9995	Rebekah Mitchell PhD	55	ABAIARA	CE	744.822.622-34	16.740.076/9329-75
9996	Lisa Parrish Jr.	73	Brasília	DF	10683395190	32246978843482
9997	Michael Young MD	87	AFONSO CLÁUDIO	ES	538.223.638-04	86.601.303/7580-88
9998	Kevin Watson DDS	82	ABADIA DE GOIÁS	GO	11632512408	08651414023648
9999	Mr. Joseph Wilson MD	50	AÇAILÂNDIA	MA	192.134.492-08	08.908.871/5161-91

[10000 rows x 6 columns]

[525]:
```
# Quantos clientes temos nessa base? R: 33 clientes distintos
max(df.nomes.value_counts())
```

[525]: 33

[526]:
```
# Qual a média de idade dos clientes? R: 53 anos
df.idade.mean()
```

[526]: 53.7831

2.0.1 Quantos clientes nessa base pertencem a cada estado?

R: Segue a lista abaixo

[17]:
```

```

```
df['estado'].replace({' ':''},inplace=True, regex=True)
df.replace({'MINASGERAI':'MG','MINASGERAIs':'MG','distritofederal':
 ↪'DF','riodejaneiro':'RJ','saopaulo':'SP',
            'sãopaulo':'SP'},inplace=True)
df[['estado','nomes']].groupby(['estado']).nunique()
```

[17]:

estado	nomes
AC	365
AL	362
AM	360
AP	365
BA	364
CE	364
DF	361
ES	359
GO	365
MA	364
MG	360
MS	361
MT	365
PA	361
PB	366
PE	362
PI	363
PR	362
RJ	362
RN	359
RO	363
RR	362
RS	360
SC	361
SE	363
SP	363
TO	363

[ ]:

2.0.2 Quantos CPFs válidos e inválidos foram encontrados?

R: Foram encontrado 10 mil cpfs validos ou seja todos os cpfs da base csv fornecida são validos.Não foram encontrados nenhum cpf invalido.

--> Segue detalhes do código abaixo

[18]:
```
#df[['cpf_limpo']]= df[['cpf']].replace({'.':'','-':'','/':''},inplace=True,␣
 ↪regex=True)
```

```
df['cpf'] = df['cpf'].str.replace('.', '')
df['cpf'] = df['cpf'].str.replace('-', '')


#df[['valida_cpf']]=df[['cpf']].apply(roda_validacao_cpf_cnpj)
df['valida_cpf']=df['cpf'].apply(roda_validacao_cpf_cnpj)

df
```

[18]:

	nomes	idade	cidade	estado	cpf	cnpj	valida_cpf
0	Dennis Daniels	31	ACRELÂNDIA	AC	97566536800	06589184909526	True
1	Leah Becker	42	ÁGUA BRANCA	AL	42526380707	25.673.336/2350-20	True
2	Sally Ford	18	ALVARÃES	AM	34647754103	26543101702989	True
3	Colleen Duncan	21	SERRA DO NAVIO	AP	25253156003	19.062.080/5100-98	True
4	Jeff Stephenson	73	ABAÍRA	BA	49668886542	97794530015384	True
...	...	...	...	...	...	...	...
9995	Rebekah Mitchell PhD	55	ABAIARA	CE	74482262234	16.740.076/9329-75	True
9996	Lisa Parrish Jr.	73	Brasília	DF	10683395190	32246978843482	True
9997	Michael Young MD	87	AFONSO CLÁUDIO	ES	53822363804	86.601.303/7580-88	True
9998	Kevin Watson DDS	82	ABADIA DE GOIÁS	GO	11632512408	08651414023648	True
9999	Mr. Joseph Wilson MD	50	AÇAILÂNDIA	MA	19213449208	08.908.871/5161-91	True

[10000 rows x 7 columns]

[529]:
```
# Todos os 10 mil registros são True ou seja todos os Cnpjs são validos
df[['cpf','valida_cpf']].groupby(['valida_cpf']).nunique()
```

[529]:

valida_cpf	cpf
True	10000

[502]:
```
# Os 10 mil registros possuem no campo CPF 11 carateres validos , ou seja não␣
 ↪existe nenhum registro que foge ao padrão de
# Numeros de carateres validos
```

```
dfx=df[df['cpf'].str.len() == 11]
max(dfx.valida_cpf.value_counts())
```

[502]: 10000

[ ]:

2.0.3 Quantos CNPJs válidos e inválidos foram encontrados?

R: Todos os 10 mil registros de CNPJ existentes no arquivo CSV analisado , são validos , não fo

[19]:
```
#df[['cpf_limpo']]= df[['cpf']].replace({'.':'','-':'','/':''},inplace=True,␣
 ↪regex=True)

df['cnpj'] = df['cnpj'].str.replace('.', '')
df['cnpj'] = df['cnpj'].str.replace('/', '')
df['cnpj'] = df['cnpj'].str.replace('-', '')


#df[['valida_cpf']]=df[['cpf']].apply(roda_validacao_cpf_cnpj)
df['valida_cnpj']=df['cnpj'].apply(roda_validacao_cpf_cnpj)

df
```

[19]:

	nomes	idade	cidade	estado	cpf	\
0	Dennis Daniels	31	ACRELÂNDIA	AC	97566536800	
1	Leah Becker	42	ÁGUA BRANCA	AL	42526380707	
2	Sally Ford	18	ALVARÃES	AM	34647754103	
3	Colleen Duncan	21	SERRA DO NAVIO	AP	25253156003	
4	Jeff Stephenson	73	ABAÍRA	BA	49668886542	
...	...	...	...	...	...	
9995	Rebekah Mitchell PhD	55	ABAIARA	CE	74482262234	
9996	Lisa Parrish Jr.	73	Brasília	DF	10683395190	
9997	Michael Young MD	87	AFONSO CLÁUDIO	ES	53822363804	
9998	Kevin Watson DDS	82	ABADIA DE GOIÁS	GO	11632512408	
9999	Mr. Joseph Wilson MD	50	AÇAILÂNDIA	MA	19213449208	

	cnpj	valida_cpf	valida_cnpj
0	06589184909526	True	True
1	25673336235020	True	True
2	26543101702989	True	True
3	19062080510098	True	True
4	97794530015384	True	True
...	...	...	...
9995	16740076932975	True	True
9996	32246978843482	True	True

```
9997  86601303758088        True         True
9998  08651414023648        True         True
9999  08908871516191        True         True

[10000 rows x 8 columns]
```

[531]:
```
df[['cnpj','valida_cnpj']].groupby(['valida_cnpj']).nunique()
```

[531]:
```
              cnpj
valida_cnpj
True         10000
```

[532]:
```
# Os 10 mil registros possuem no campo CNPJ 14 carateres validos , ou seja não existe nenhum registro que foge ao padrão de
# Numeros de carateres validos
dfx=df[df['cnpj'].str.len() == 14]
max(dfx.valida_cnpj.value_counts())
```

[532]:
```
10000
```

Gere um arquivo no formato csv chamado problema1_normalizado –> Gerado Segue o código abaixo

[25]:
```
# Exportando Para CSV
df.to_csv('problema1_normalizado.csv', index = False, sep=';', encoding='utf-8'')
```

[67]:
```
# Exportando Para Parquet
df.to_parquet("./problema1_normalizado.pq")
```

2.0.4 Não houve tempo habil para desenvolver as demais atividades

[ ]:
```

```

Problema 2: Você deverá implementar um programa, para ler, tratar e particionar os dados.

O arquivo fonte está disponível em `https://st-it-cloud-public.s3.amazonaws.com/people-v2_1E6.csv.gz`

2.0.5 Data Quality

- Higienizar e homogenizar o formato da coluna `document`
- Detectar através da coluna `document` se o registro é de uma Pessoa Física ou Pessoa Jurídica, adicionando uma coluna com essa informação
- Higienizar e homogenizar o formato da coluna `birthDate`
- Existem duas colunas nesse dataset que em alguns registros estão trocadas. Quais são essas colunas?
- Corrigir os dados com as colunas trocadas
- Além desses pontos, existem outras tratamentos para homogenizar esse dataset. Aplique todos que conseguir.

2.0.6 Agregação dos dados

- Quais são as 5 PF que mais gastaram (`totalSpent`)?
- Qual é o valor de gasto médio por estado (`state`)?
- Qual é o valor de gasto médio por `jobArea`?
- Qual é a PF que gastou menos (`totalSpent`)?
- Quantos nomes e documentos repetidos existem nesse dataset?
- Quantas linhas existem nesse dataset?

2.0.7 Particionamento de dados tratados com as regras descritas em `DATA QUALITY`

- Particionar em arquivos PARQUET por estado (`state`)
- Particionar em arquivos CSV por ano/mes/dia de nascimento (`birthDate`)