Projeto Final de Spark

Campanha Nacional de Vacinação contra Covid-19

O projeto foi dividido em dois níveis, básico e avançado. Recomendo fortemente fazer

primeiro o básico e se sobrar tempo, pode aventurar no avançado.

Os exercícios podem ser feitos em qualquer linguagem e todas as questões são bem

abertas, tendo várias formas de serem realizadas e interpretadas, pois a idéia é não

termos projetos iguais.

O projeto deve estar no github.com, a forma de organizar o conteúdo é por sua conta,

caso nunca tenha usado, este já é seu primeiro desafio.

Ao final do projeto você precisa preencher o formulário com o seu nome completo, e-

mail utilizado no treinamento e o link do github do seu projeto.

Link do formulário para envio:

https://forms.office.com/Pages/ResponsePage.aspx?id=2H\_OZbilA0GZftoGjNhf1Y4a9b

NsmMNEil2MBcFKJolUMFlTQVBNUVhRTVlSNVJUUDBWM0ZIRDVKQS4u

O formulário também estará na página do treinamento.

OBS: Todas as imagens de exemplo (Visualizações) são apenas para referencias, o

projeto irá ter valores diferentes e as formas de se criar tabelas com

dataframe/dataset das visualizações, pode ser realizado da maneira que preferir.

Nível Básico:

Dados: https://mobileapps.saude.gov.br/esus-

vepi/files/unAFkcaNDeXajurGB7LChj8SgQYS2ptm/04bd3419b22b9cc5c6efac2c652810

0d\_HIST\_PAINEL\_COVIDBR\_06jul2021.rar

Referência das Visualizações:

• Site: https://covid.saude.gov.br/

• Guia do Site: Painel Geral

1. Enviar os dados para o hdfs

2. Otimizar todos os dados do hdfs para uma tabela Hive particionada por

município.

3. Criar as 3 vizualizações pelo Spark com os dados enviados para o HDFS:

4. Salvar a primeira visualização como tabela Hive

5. Salvar a segunda visualização com formato parquet e compressão snappy

6. Salvar a terceira visualização em um tópico no Kafka

7. Criar a visualização pelo Spark com os dados enviados para o HDFS:

8. Salvar a visualização do exercício 6 em um tópico no Elastic

9. Criar um dashboard no Elastic para visualização dos novos dados enviados

Nível Avançado:

Replicar as visualizações do site “https://covid.saude.gov.br/”, porém acessando

diretamente a API de Elastic.

Link oficial para todas as informações:

https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao

Informações para se conectar ao cluster:

• URL https://imunizacao-es.saude.gov.br/desc-imunizacao

• Nome do índice: desc-imunizacao

• Credenciais de acesso

o Usuário: imunizacao\_public

o Senha: qlto5t&7r\_@+#Tlstigi

Links utéis para a resolução do problema:

• Consumo do API:

https://opendatasus.saude.gov.br/dataset/b772ee55-07cd-44d8-958f-

b12edd004e0b/resource/5916b3a4-81e7-4ad5-adb6-

b884ff198dc1/download/manual\_api\_vacina\_covid-19.pdf

• Conexão do Spark com Elastic:

https://www.elastic.co/guide/en/elasticsearch/hadoop/current/spark.html

https://docs.databricks.com/data/data-sources/elasticsearch.html#elasticsearch-

notebook

https://github.com/elastic/elasticsearch-hadoop

https://www.elastic.co/guide/en/elasticsearch/hadoop/current/configuration.html

• Instalar Dependências:

https://www.elastic.co/guide/en/elasticsearch/hadoop/current/install.html