PROVA PRÁTICA
ESPECIALISTA EM ENGENHARIA DE DADOS - DEVOPS

OBSERVATÓRIO DA INDÚSTRIA

O que queremos receber?

Um repositório no seu Github com texto da pergunta 1 e 2ª, além dos scripts das demais questões. Além do repositório, queremos também receber nos e-mail´s: dgasilva@sfiec.org.br.

1) Auto avaliação

Auto-avalie suas habilidades nos requisitos de acordo com os níveis especificados usando o link abaixo.

Qual o seu nível de domínio nas técnicas/ferramentas listadas abaixo, onde:

- 0, 1, 2 - não tem conhecimento e experiência;
- 3, 4 ,5 - conhece a técnica e tem pouca experiência;
- 6 - domina a técnica e já desenvolveu vários projetos utilizando-a.

Tópicos de Conhecimento:

- Manipulação e tratamento de dados com Python: __
- Manipulação e tratamento de dados com Pyspark: __
- Desenvolvimento de data workflows em Ambiente Azure com databricks: __
- Desenvolvimento de data workflows com Airflow: __
- Manipulação de bases de dados NoSQL: __
- Web crawling e web scraping para mineração de dados: __
- Construção de APIs: REST, SOAP e Microservices: __

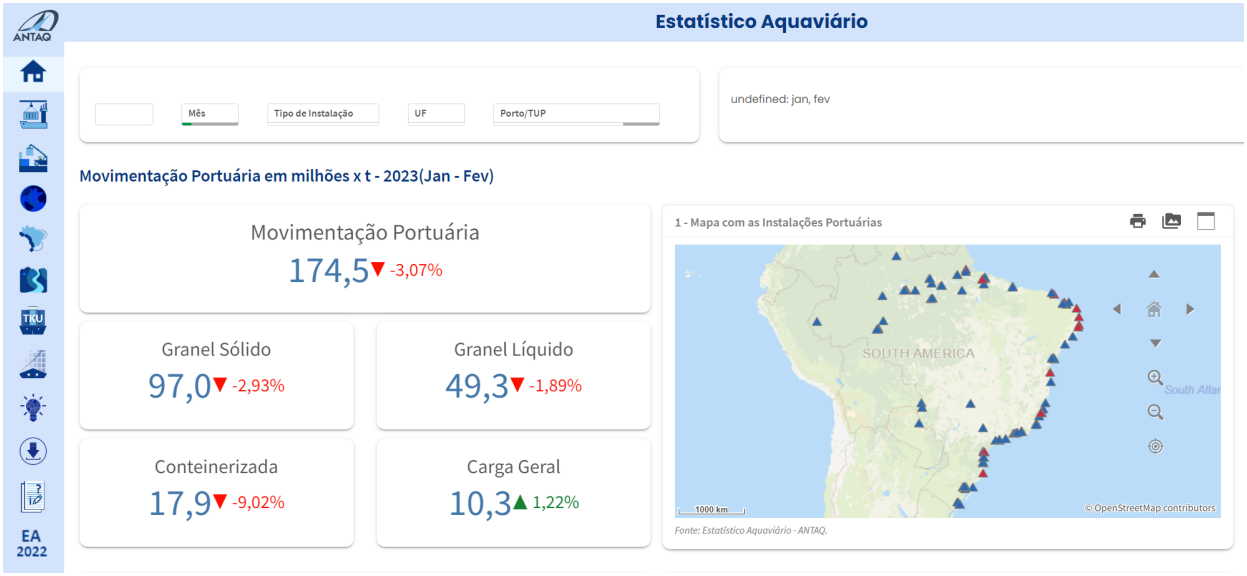
2) Desenvolvimento de pipelines de ETL de dados com Python, Apache Airflow, Hadoop e Spark.

Foi solicitado à equipe de AI+Analytics do Observatório da Indústria/FIEC, um projeto envolvendo os dados do Anuário Estatísticos da ANTAQ (Agência Nacional de Transportes Aquáticos).
O projeto consiste em uma análise pela equipe de cientistas de dados, bem como a disponibilização dos dados para o cliente que possui uma equipe de analistas própria que utiliza a ferramenta de BI (*business intelligence*) da Microsoft.

Para isto, o nosso *cientista de dados* tem que entender a forma de apresentação dos dados pela ANTAQ e assim, fazer o ETL dos dados e os disponibilizar no nosso data lake para ser consumido pelo time de cientistas de dados, como também, elaborar uma forma de entregar os dados tratados ao time de analistas do cliente da melhor forma possível.

Informações Importantes:

Painel de BI: https://web3.antaq.gov.br/ea/sense/index.html#pt



Documentação: https://web3.antaq.gov.br/ea/sense/download.html#pt

Banco SQL da FIEC: SQL Server
Banco NoSQL da FIEC: Mongo DB
Ferramenta dos analistas de BI do cliente: Power BI

Supondo que você seja nosso Cientista de dados:

a) Olhando para todos os dados disponíveis na fonte citada acima, em qual estrutura de banco de dados você orienta guardá-los no nosso Data Lake? SQL ou NoSQL? Discorra sobre sua orientação. (1 pts)

b) Nosso cliente estipulou que necessita de informações apenas sobre as atracações e cargas contidas nessas atracações dos últimos 3 anos (2017-2019). Logo, o time de cientistas de dados, em conjunto com você, analisaram e decidiram que duas tabelas, uma para atracação e outra para carga, seriam suficientes tanto para o trabalho do Observatório como para trabalho do time externo.
Assim, desenvolva um script em python que extraia os dados do anuário, e transforme-os em duas tabelas fato, atracacao_fato e carga_fato, com as respectivas colunas abaixo.
Lembrando que os dados têm periodicidade mensal, então *script's* automatizados e robustos ganham pontos extras. (2 pontos + 1 ponto para solução automatizada e elegante)

Colunas da tabela atracacao_fato:

IDAtracacao	Tipo de Navegação da Atracação
CDTUP	Nacionalidade do Armador
IDBerco	FlagMCOperacaoAtracacao
Berço	Terminal
Porto Atracação	Município
Apelido Instalação Portuária	UF
Complexo Portuário	SGUF
Tipo da Autoridade Portuária	Região Geográfica
Data Atracação	Nº da Capitania
Data Chegada	Nº do IMO
Data Desatracação	TEsperaAtracacao
Data Início Operação	TesperaInicioOp
Data Término Operação	TOperacao
Ano da data de início da operação	TEsperaDesatracacao
Mês da data de início da operação	TAtracado
Tipo de Operação	TEstadia

Colunas da tabela carga_fato: (atente-se que para o tipo de carga conteinerizada, pois cada contêiner pode ter mais de uma mercadoria)

IDCarga	FlagTransporteViaInterioir
IDAtracacao	Percurso Transporte em vias Interiores
Origem	Percurso Transporte Interiores
Destino	STNaturezaCarga
CDMercadoria (Para carga conteinerizada informar código das mercadorias dentro do contêiner.)	STSH2
Tipo Operação da Carga	STSH4
Carga Geral Acondicionamento	Natureza da Carga
ConteinerEstado	Sentido
Tipo Navegação	TEU
FlagAutorizacao	QTCarga
FlagCabotagem	VLPesoCargaBruta
FlagCabotagemMovimentacao	Ano da data de início da operação da atracação
FlagConteinerTamanho	Mês da data de início da operação da atracação

FlagLongoCurso	Porto Atracação
FlagMCOperacaoCarga	SGUF
FlagOffshore	Peso líquido da carga (Carga não conteinerizada = Peso bruto; Carga conteinerizada = Peso sem contêiner)

c) Essas duas tabelas ficaram guardadas no nosso Banco SQL SERVER. Nossos economistas gostaram tanto dos dados novos que querem escrever uma publicação sobre eles. Mais especificamente sobre o tempo de espera dos navios para atracar. Mas eles não sabem consultar o nosso banco e apenas usam o Excel. Nesse caso, pediram a você para criar uma consulta (query) otimizada em sql em que eles vão rodar no excel e por isso precisa ter o menor número de linhas possível para não travar o programa.
Eles querem uma tabela com dados do Ceará, Nordeste e Brasil contendo número de atracações, para cada localidade, bem como tempo de espera para atracar e tempo atracado por meses nos anos de 2018 e 2019. Segundo tabela abaixo: (2pts)

Localidade	Número de Atracações	Variação do número de atracação em relação ao mesmo mês do ano anterior - Bônus	Tempo de espera médio	Tempo atracado médio	Mês	Ano

3) **Criação de ambiente de desenvolvimento com Linux e Docker.**

Finalmente, este processo deverá ser automatizado usando a ferramenta de orquestração de *workflow* Apache Airflow + Docker. Escreva uma DAG para a base ANTAQ levando em conta as características e etapas de ETL para esta base de dados. Esta também deve conter operadores para enviar avisos pore-mail quando necessário (e.g.: caso os dados não sejam encontrados, quando o processo for finalizado, etc). Todos os passos do processo ETL devem ser listados como *tasks* e orquestrados de forma otimizada, porém não é necessário migrar o código criado anteriormente para dentro das *tasks do Airflow* (foque em mostrar o fluxo de *tasks* e as estruturas básicas de uma DAG). Caso isso seja feito, será considerado um extra. (2 pontos)

4) Configuração de pipelines de CI/CD com Gitlab ou Github.

A equipe de desenvolvimento deseja implantar a pipeline/DAG de ETL da ANTAQ nos ambientes de homologação e produção. Contudo, não quer se dar ao trabalho de implantar

esta pipeline de forma manual cada vez que houver uma alteração no seu código. Assim, querem utilizar os conceitos de CI/CD para realizar a implantação (deploy) de forma automatizada para os ambientes de execução.

Dessa forma, ainda com base no repositório criado para a pipeline/DAG da ANTAQ, crie um arquivo de configuração para pipelines de CI/CD que seja capaz de "testar", "empacotar" e "distribuir" o código fonte da pipeline/DAG, para os ambientes de homologação e produção. Este arquivo deverá residir no repositório junto com os demais arquivos. (1pts)

A execução da pipeline de CI/CD será analisada e testada pelos avaliadores em ambiente real. Por tanto, para as etapas de "cópia" dos arquivos para os ambientes, você pode usar um comando "fake" para simular a cópia. Caso queira copiar os artefatos/arquivos da pipeline de CI/CD para um ambiente real, isso será considerado como um extra. Ao final dessa etapa, deve existir uma pipeline de CI/CD em execução, empacotando e distribuindo o código fonte do repositório da ANTAQ.

5) Implantação de aplicações com Kubernetes.

A equipe de desenvolvimento possui várias pipelines/DAG´s de ETL que necessitam ser executadas em ambiente de produção. A quantidade de pipelines/DAG´s que a equipe implanta diariamente vem crescendo nos últimos tempos e a infraestrutura atual não está comportando tamanha demanda. Para tornar o ambiente de execução das pipelines/DAG´s de ETL mais robusto, decidiu-se utilizar uma solução de gerenciamento de containers. Assim, fica muito mais fácil escalar as aplicações para atender as demandas de ingestão de novas pipelines/DAG´s.

Dessa forma, crie script´s de configuração, que seja capaz de implantar um servidor Kuberentes em um Ubuntu Server 20.04. Os scripts devem ser capazes de implantar um nó master e nós filhos do Kubernetes, bem como realizar o deploy de um serviço de Airflow para um namespace específico. (4pts)
O serviço do Airflow criado deve conter no mínimo os seguintes componentes (WebServer, Scheduler, Worker).
O serviço de Airflow deve se capaz de ler a pipeline/DAG da ANTAQ. Esta DAG deve ser armazenada e disponibilizada para o serviço do Airflow através de volumes Persistentes do Kubernetes mapeados para um Storage.

Os scripts serão avaliados e testados pelos avaliadores em ambiente real. Por tanto, crie os scripts com comentários explicativos e em uma sequência lógica que facilite a implantação do servidor. Ao final dessa etapa, deve existir um servidor Kubernetes em execução, com a aplicação Airflow Webserver acessível via HTTP e com a DAG criada anteriormente registrada na interface.

6) **Implantação de Data Lake com Hadoop**

A equipe de desenvolvimento necessita disponibilizar os dados capturados pelas pipelines em um ambiente rápido, robusto e tolerante a falhas. Diante disso, a equipe decidiu utilizar o Apache Hadoop File System, como ferramenta para implantar o seu data lake.

O Observatório possui uma massa de dados públicos que ultrapassa 14 Bilhões de registros. Atualmente, essa massa de dados consume perto de 2 TeraBytes de dados. O Data Lake deverá ser utilizado tanto por usuários e analista de negócio, cientistas e engenheiros de dados, além de ferramentas na camada de visualização como Power BI,

QLikView, API, dentro outros. O data lake também é utilizado pelo time de desenvolvimento para fins de testes, onde leituras e escritas de dados massivas são realizadas. Por tanto, o data lake deve ser capaz de suportar grandes operações de escrita e leitura dos dados, em dois ambiente distintos sendo, desenvolvimento e produção.

Assim, como um membro da equipe de infraestrutura, sugira uma arquitetura para configuração de um cluster de Hadoop levando em consideração o cenário descrito.

Ao final de dessa etapa, deve existir um diagrama informando a arquitetura sugerida para suportar as operações de data lake da instituição. O diagrama deve conter além da descrição dos componentes, bem como as especificações de hardware necessárias (processamento, memória, disco).