Case 1

Cenário Atual

Estamos passando por um processo de transformação digital, onde o seu papel será definir a arquitetura de referência para plataforma de dados do Grupo Boticário e ser uma referência técnica para engenheiros e analistas de dados. No cenário atual, utilizamos SAP Hana como nosso repositório principal de data warehouse. Existem processos de ETL que fazem ingestão de dados de 50 transacionais. Mais de 90% das bases são de origem transacionais de diferentes DBMS's (DB2, MS SQL etc) e estão alocados em ambiente on-premises. Além do SAP Hana, a empresa possui algumas aplicações hospedadas em nuvens públicas como Microsoft Azure e Amazon Web Services. Dentro da empresa, o tratamento e o consumo dos dados são tratados em silos, onde diferentes unidades de negócios acabam utilizando diferentes ferramentas para processar, analisar dados e apresentar dados. Algumas ferramentas que podemos citar como exemplo são Jupyter Notebook, Qlick, Qlick Sense. Outro aspecto importante está ligado a governança de dados, onde aspectos como acesso a dados sensíveis, catalogação e permisionamento carecem de melhorias.

O que esperamos?

1. Que você defina uma arquitetura de referência com tecnologias de alguma nuvem pública, preferencialmente AWS ou GCP. Você deve considerar os seguintes requisitos:
   • Permear as camadas de ingestão, processamento, armazenamento, consumo, análise, segurança e governança;
   • Substituição gradativa do cenário on-premises atual;
   • Incorporação de componentes e tecnologias que permitam a analisarmos dados em tempo real;
   • Que a arquitetura considere componentes que a habilitem a empresa organizar e fornecer dados para diferentes fins, tais como: Analytics, Data Science, API's e serviços para integrações com aplicações. Ressaltando que necessariamente precisaremos manter a comunicação on-premises x cloud para diversas finalidades.

2. Que você prepare uma apresentação para discutir a arquitetura definida, detalhando a motivação da escolha dos componentes, especificando as vantagens, desvantagens e riscos do modelo definido.

3. Que você apresente de forma presencial ou videoconferência para nosso time de arquitetura. Você terá 30 minutos para apresentação.

4. Que você traga exemplos práticos, onde tenha aplicado anteriormente a utilização dos componentes selecionados para esta arquitetura. Observação: Ao definir a arquitetura, considere que somos uma das maiores empresas de varejo do Brasil.

O que não esperamos?

Para este case, avaliaremos seu conhecimento de arquitetura e capacidade de atuar de forma consultiva junto a engenheiros, cientistas e analistas de dados. Não é esperado que você implemente a arquitetura ou que realize qualquer tipo de codificação.

Case 2

Junto com este descritivo, você está recebendo 3 arquivos com dados aleatórios de vendas de 2017 a 2019. Para a execução deste teste, você pode utilizar as ferramentas que estiver mais familiarizado, seguindo apenas as seguintes premissas:

1. Os dados necessariamente devem ser armazenados em tabelas de banco de dados (MySQL, PostgreSQL, BigQuery, MS SQL, Oracle etc) e não em arquivos ou planilhas;

2. Você deve necessariamente utilizar as linguagens SQL e Python nos processos de carga, consulta e transformação dos dados;

3. Utilizar uma ferramenta que lhe permita criar os processos de ETL ou DAG's para ingestão e transformação de dados;

4. Você deve implementar um controle de versionamento para seus códigos.

Obs. Embora você possa utilizar ferramentas de sua escolha, estamos estruturando nossa plataforma de dados na nuvem do Google, se você realizar as implementações utilizando ferramentas do GCP, ficaremos mais felizes ainda! #ficaadica ;-)

Você precisará:

1. Realizar a importação dos dados dos 3 arquivos em uma tabela criada por você no banco de dados de sua escolha;

2. Com os dados importados, modelar 4 novas tabelas e implementar processos que façam as transformações necessárias e insiram as seguintes visões nas tabelas:

a. Tabela 1: Consolidado de vendas por ano e mês;

b. Tabela 2: Consolidado de vendas por marca e linha;

c. Tabela 3: Consolidado de vendas por marca, ano e mês;

d. Tabela 4: Consolidado de vendas por linha, ano e mês;

3. Você deverá criar um método para realizar uma pesquisa no Spotify via requests e trazer os primeiros 50 resultados referente a podcasts procurando pelo termo "data hackers" e criar uma tabela apenas com os campos abaixo:

a. Tabela 5: name = Nome do poscast.

description = Descrição sobre o programa de poscast.

id = Identificador único do programa. total_episodes = Total de episódios lançados até o momento.

4. Realizar a extração de dados de todos os episódios lançados pelos Data Hackers via requests e ingerir esse resultado em duas tabelas seguindo os critérios abaixo:

a. Tabela 6: Resultado de todos os episódios.

b. Tabela 7: Apenas com os resultados dos episódios com participação do Grupo Boticário.

Levar apenas os campos abaixo para as tabelas 6 e 7:

id - Identificação do episódio.

name - Nome do episódio.

description - Descrição do episódio.

release_date - Data de lançamento do episódio.

duration_ms - Duração em milissegundos do episódio.

language - Idioma do episódio.

explicit - Flag booleano se o episódio possui conteúdo explícito.

type - O tipo de faixa de áudio (Ex: música / programa)

Para realizar o credenciamento na API do Spotify e obter as credenciais necessárias, você precisa seguir os seguintes passos:

1. Criar uma conta no Spotify Developer Dashboard: Acesse o site https://developer.spotify.com/dashboard/ e crie uma conta gratuita.

2. Criar um novo aplicativo: Na sua conta no Spotify Developer Dashboard, clique em “Criar um novo aplicativo” e siga as instruções para criar seu aplicativo.

3. Obter as credenciais de acesso: Na tela principal do seu aplicativo você conseguirá identificar o Client ID e logo abaixo basta clicar em Show Client Secret para mostrar o Client Secret.

Agora que você tem as credenciais, você já pode iniciar o processo de extração. O que esperamos de você?

1. Que consiga realizar todas as etapas acima;

2. Que gere um repositório com todos os scripts e nos disponibilize para consulta;

3. Que nos demonstre todos os processos rodando e as tabelas sendo carregadas. Marcaremos uma reunião virtual para que possa demonstrar;

4. Que nos apresente como se organizou para realizar as tarefas.