Teste teórico de analista de Business Intelligence

Nome do Candidato (a): Mateus Romero Verissimo da Silva
Data 06/08/2020

Crie um repositório no github e ponha os resultados e os código lá. Envie o link de acesso ao seu repositório criado.

Questão 4 – Uma tabela de clientes possui uma coluna sexo com dois valores possíveis (M – Masculino e F – Feminino). Grande parte das consultas considera o sexo como critério de pesquisa na cláusula WHERE juntamente com outros campos. Que tipo de índice que deve ser utilizado nessa coluna?

( ) Clustered Index.

( ) Nonclustered Index.

( ) Bitmap Index.

(X) Não deve ser utilizado um índice nessa coluna por sua alta densidade.

( ) Não deve ser utilizado um índice nessa coluna por sua alta seletividade.

Resposta Mateus.: Não faz muito sentido a inclusão de índice pela coluna de sexo uma vez que só há dois tipos de valores possíveis e, dependendo da base, a divisão de registros tende a ser equilibrada, sendo assim, a consulta continuaria a ser executada extraindo um grande volume de registros. Há a necessidade de procurar por colunas onde possa ser aberto o leque de opções. Em tabelas fatos por exemplo (que não é o caso), geralmente o ideal é utilizar a coluna de data do evento como indice.

Questão 5 – De acordo com o T-SQL, quais são as cláusulas obrigatórias em uma query de SELECT?

( ) As cláusulas FROM E SELECT.

( ) As cláusulas SELECT E WHERE.

(X) A cláusula SELECT.

( ) As cláusulas SELECT, FROM E WHERE.

Resposta Mateus.: Ambientes Oracle o “FROM” também se faz necessário. Em sqlserver não é necessário.

Questão 7 - O que acontece após a execução do comando: SELECT TRY_CAST('abc' AS INT).

( ) Um erro é gerado Um valor.
(X) null é retornado.
( ) Um valor inteiro é retornado.
( ) Uma string é retornada.

Resposta Mateus.: Quando não é possível converter para o datatype solicitado no try_cast, em sqlserver, ele retorna null. No ORACLE a função CAST daria erro, não havendo uma função similar pronta pelo fabricante.

Questão 8 - Em relação à clausula Where e Having podemos afirmar?

( ) Ambas tem a mesma função.
( ) São funções diferentes.
(X) Ambas tem a mesma função mas o filtro da clausula where linha por linha e o Having após o agrupamento.
( ) Ambas acontecem durante o agrupamento.

Questão 9– Você está criando um pacote SSIS na sua máquina que aponta para uma base SQL Server com uma conta SQL e é executado via Job agendado. Após concluir o pacote remete para produção e no outro dia quando verifica o JobHistory tem o seguinte erro

DTS_E_OLEDBERROR. An OLE DB error has occurred. Error code: 0x80040E4D. An OLE DB record is available. Source: "Microsoft SQL Native Client" Hresult: 0x80040E4D Description: "Login failed for user '<User_Name>'."

O que você deve fazer para que o pacote execute corretamente a noite?

( ) Mude todas as conexões para usar SQL Authentication.
(X) Mude todas as conexões para usar Windows Authentication
( ) Encriptar o pacote com "EncryptSensitiveWithPassword" ou "EncryptAllWithPassword" e forneça a senha cada vez que o usuário precisar executar.
( ) Crie um DTSConfig para fornecer informações de conexão para o pacote em tempo de execução.

Questão 11 – Quais componentes são do MS-SQL Server Integration Services:

( ) Designer SSIS, Cubos OLAP, Tarefas e Elementos de Fluxo de dados.

(X) Designer SSIS, Contêineres, Tarefas e Elementos de Fluxo de dados.

( ) Data Mart, Designer SSIS, Contêineres e Elementos de Fluxo de Dados.

( ) Data Mart, Designer SSIS, Tarefas e Elementos de Fluxo de Dados.

( ) Data Mart, Cubos OLAP, Contêineres e Tarefas.

Questão 12 - Em um comando SQL, o operador LIKE é usado em uma cláusula WHERE para buscar um determinado padrão em uma coluna.

(X) Certo.
( ) Errado.

Questão 14 - Muitos autores consideram a tecnologia de Data Warehousing (o processo de fazer Data Warehouse) como sendo uma evolução natural do ambiente de apoio à decisão. As empresas utilizam Data Warehouse com mais frequência, pois há a necessidade de domínios de informações estratégicas que podem garantir respostas rápidas, assegurando, dessa forma, a competitividade no mercado concorrente e em constantes mudanças. O DW possui diversas características. "A arquitetura do Data Warehouse inclui, além de estrutura de dados, mecanismos de comunicação, processamento da informação para o usuário. " Assinale, a seguir, a característica correspondente.

( ) Não volátil.
( ) Integração.
(X) Variação de tempo.
( ) Orientado por assunto.
( ) Arquitetura do ambiente.

Questão 15 - O objetivo dessa área é criar um ambiente intermediário de armazenamento e processamento dos dados oriundos de aplicações OLTP (Online Transaction Processing) e outras fontes, para o processo ETL (Extract Tranform Load), possibilitando seu tratamento, e permitindo sua posterior integração em formato e no tempo, evitando problemas após a criação do Data Warehouse e a concorrência com o ambiente transacional no consumo de recursos. A área citada é conhecida como:

( ) Transaction area.
( ) Warehouse.
( ) Backup area.
(X) Staging area.
( ) Cube area.

Questão 19 - VIEW é uma tabela virtual cujo conteúdo está definido por uma instrução SELECT
(X) Certo.
( ) Errado.

Questão 20 - No MS SQL Server, as tabelas criadas por meio do comando CREATE TABLE são temporárias se:
( ) A opção TEMP é especificada logo após o termo CREATE.
( ) O comando é executado dentro de uma stored procedure.
( ) O usuário não possui privilégio para criação de tabelas.
(X) O nome da tabela é iniciado por #.
( ) A opção ON refere-se ao filegroup TEMP.

Questão 21 – Descreva os modelos Start Schema (Ralph Kimball) e SnowFlake (Bill Inmon).

Resposta Mateus.: Os modelos Star Schema e Snow Flake são bastante utilizados em qualquer que seja a empresa que tenha algum DW para utilização. A crucial diferença entre os dois, se da no relacionamento entre as dimensões e as fatos. No modelo Star Schema, dimensões se relacionam apenas com fatos, ou seja, tende a ser formada uma estrela com N dimensões ligadas para uma fato centralizada. Fatos não se relacionam nesse modelo e nem dimensões se relacionam entre si, apenas a ligação de DIMENSÃO -> FATO. Já no modelo Snow Flake, as dimensões podem se relacionar entre si para depois serem associadas a fato. Geralmente existem alguns problemas com esse modelo ao ser utilizado para a visualização de relatório, sempre tendendo a utilizar modelo Star Schema para a visualização.

Questão 23 – O que podemos entender por "Granularidade do dado"?

Resposta Mateus.: Pode ser entedido como em que nível o dado se encontra. Se está no nível de transação, se está em um nível agregado, etc. Quando falamos na extração de sistemas transacionais, a extração tende a estar no menor nível possível, ou seja, no nível de granularidade da transação que foi feita diretamente no sistema de origem. Quando vamos para as transfromações e staging, há a possiblidade da agregação dos dados extraídos para serem carregados agregados para produzirem uma melhor performnace e os números serem exibidos de forma rápida, segura e coerente.

Exemplo: 10 transações de venda para o estado de são paulo são 10 linhas em um banco transacional. Se o dado for agregado, podemos trazer uma coluna com quantidade de transações e outra com o estado, na primeira coluna teria o número 10 de quantidade de linhas e no segundo o código do estado de SP. Nesse exemplo, 10 linhas viraram apenas 1 com granularidades diferentes.

Teste Big Data (Daqui para baixo está em inglês)

**1) You work on a start-up that developed a bracelet to track down data about the health
of inpatients. Each bracelet sends the data in JSON every 6 seconds to be analyzed and stored.These data will be used to generate a daily report on the Health Portal andyou need to come up with a real-time solution for analytics that is durable, scalable and
parallel to support the whole operation.**

Describe and justify the possible choices for the following architecture components:

Resposta Mateus: Podemos utilizar o Kafka para ficar "escutando" o recebimento dos arquivos JSONs, após isso, utilizar o FLUME para capturar os mesmos e jogar para um ambiente HDFS hadoop os arquivos gerados em real time. Com os arquivos já no hadoop e sendo acumulados, é possível manter uma rotina de tratamento de dados via Spark para adaptar o consumo dos mesmo para carregar em um banco NoSQL como o HBASE e ser consumido porteriormente por alguma ferramenta de visualização de dados para acompanhamento , como Tableau.

No final do dia, é possível carregar os arquivos em algum banco de dados para não se perder o histórico dos dados e que seja possível as análises retroativas em dashboards.

As ferramentas acima escolhidas foram escolhidas por serem de mercado, amplantamente conhecidas por serem eficazes na sua área de atuação e que se integram bem entre si. Há a possibilidade da substituição das ferramentas por outras similaraes com a mesma função, de acordo com as licenças possíveis adquiridas pelo cliente.

2) Explain the difference between Amazon Athena and Redshift Spectrum as well as
the main use cases for each of them.

Resposta Mateus: Ambos são utilizados para realizar queries em linguagem SQL em buckets de S3. A diferença é que o Athena é um serviço e possui recursos alocados próprios dele, embora os dois sejam serveless. Já o Redshift Spectrum, é necessário possuir um cluster proprio para processamento Red Shift e não utilizando recursos proprios como o Athena. A performance do Spectrum se da a partir do cluster que foi adquirido, ou seja, se tiverem mais nodes a performance é mais alta e mais baixa caso menos. O Athena é um serviço e o Red Shift Spectrum é uma feramenta embarcada do cluster do Red Shift Spectrum. O Red Shift Spectrum é mais aconselhável quando já existe um ambiente como DW ou Datalake para ser consumido.

3) You work for a start-up of photos processing and you need to swap the colors to black and white after loading them into Amazon S3. How can you do this on AWS??

Resposta Mateus: Uma função lambda serveless criada a partir de uma trigger de toda imagem que é feita upload para o banco, onde toda iamgem que entra no bucket é disparado o gatilho para transformar a mesma em preto e branco e já salva a mesma no bucket ou em outro bucket;

4) An organization implemented a streaming solution, on which a data goes through a

**Kinesis Data Stream and a Kinesis Data Stream until it is stored on Redshift and is made available to analysis. A new product requirement specifies some events which
should be processed with a minimum delay and could trigger some actions afterward.**

Resposta Mateus: Acho que a pergunta veio pela metade, mas se eu entendi, talvez seria para diminuir o número de etapas de um streaming e se tornar mais rápido com menos delay. Para essa situação, eu tentaria ramificar os dados para partes menores para que o delay seja o menor possível dentro do ecossistema desejado.

5) Which technologies below are related to Big Data on Cloud?

a. Kubernetes, Jenkins, Terraform
b. Azure SQL Server, AWS Lambda, AWS EC2
c. Google BigQuery, Apache Spark, Amazon Redshift
d. Digital Ocean, Packet, Javascript
e. AWS, Google, Facebook

6) Which file type is the best to read/write tabular data on big scales?

a. CSV
b. Protobuf
c. Gzip
d. Parquet
e. JSON
f. Avro

**7) Choose all correct answers To real-time data processing which technology is best for
the streaming layer?**

a. Apache Kafka
b. MySQL
c. MongoDB
d. Python
e. Apache Spark

8) Explain the main points that define the concepts of ELT and ETL.

Resposta Mateus: A principal diferença está entre carregar os dados antes de fazer as transformações ou não. O ELT, carrega os dados para depois realizar algum tipo de tratamento. Exemplo: Datalakes, primeiro os dados são recebidos/extraidos e são carregados sem muitos tratamentos para depois realizarem os tratamentos necessários e se necessário. Isso encurta a necessidade final de alguns projetos, onde não é necessário entregar a

fase de transformação para o dado ser consumido. Já o ETL, os dados já são tratados antes de serem carregados em ambientes de Data Warehouse. O dado já é carregado tratado e pronto para ser consumido por aplicações ou usuários.

9) Define in some lines the characteristics, 2 examples, and 2 use cases each for the
following types of Databases:

• Relational:

Resposta Mateus: Bancos estruturados e com restrições para não haver registros duplicados, geralmente utilizados por sistemas transacionais e servem como base de extração para a maioria dos DWs do mercado. Ex ferramentas.: Oracle, SqlServer, etc. Ex cases.: Sistemas de venda, sistemas de CRM, etc.

• Key Value:

Resposta Mateus: São os famosos bancos NoSQL, bastante utilizado em ecossitemas de Big Data, onde não é considerado um banco relacional e sim um não-relacional. Tende a ser mais escalável do o que o relacional pois envolve menos custo par aisso. É utilizado por grandes empresas para grandes volumes de dados por conta da referência de sua chave. Exemplos: Mongo DB, Dynamo DB Amazon. Ex cases.: Facebook, maps, etc.

• Documents:

Resposta Mateus: Outro tipo de banco não relacional projetado para armazenar e consumir dados do tipo JSON. Exemplos de ferramenta: Amazon, Mongo Db, etc. Exemplos de uso: blogs, catalogos, etc.

• Graphs:

Resposta Mateus: Para os bancos de dados de grafos, a ideia é criar um modelo menos generico que o relacional, proporcionando algo mais simples e focado buscar obeter mais performance para operações com joins. Um exemplo bem comum de uso seria os “amigos de amigos” do Facebook. Exemplo: NEO4J, Caso de uso: Amigos do facebook.

• Timeseries:
Resposta Mateus: Banco de dados temporais, geralmente utilizado para uso de tendencias ou séries ou baseado em ações futuras. Alguns exemplos que podemos utilizar seria na área médica para diagnósticos, sistemas de reserva de hoteis, situações geográficas, etc.

• In-Memory:

Resposta Mateus: principal característica que o nome já diz, é a utilização da memória para resolver alguns assuntos mais rápidamente, pois uma vez que processado na memória o acesso fica mais rápido quando comparado ao disco. Existem ferramentas específicas que podem trazer o melhor do mundo do DW com comparações a origem, ou seja, OLAP e OLTP em uma única visualização, como é o exemplo do SAP Hana.

Teste Python
Baixe o arquivo e resposta as perguntas abaixo: (use pandas e numpy para lhe ajudar)
1.What is the average distance traveled by trips with a maximum of 2 passengers;
2 - What is the average trip time on Saturdays and Sundays;
3- To be able to provision your entire environment in a public cloud, preferably AWS.

https://github.com/mromerovsv/teste_trip_data