Teste teórico de analista de Business Intelligence Nome do Candidato (a): Liz Alexandrita de Souza Barreto

Data 07/07/2020

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. *Usado em PK, melhora busca, preserva ordem*

( ) Nonclustered Index. *Melhora update, pode ser usado em qq Key, bom para alta seletividade*

( ) Bitmap Index. *Melhora busca em baixa cardinalidade porém entre 100:1 até cerca do 10.000:1*

( X) Não deve ser utilizado um índice nessa coluna por sua alta densidade. *Densidade relativamente alta de ½, 0.5. É mais performático fazer um full scan na tabela base.*

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

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

(X ) As cláusulas FROM E SELECT.

( ) As cláusulas SELECT E WHERE.

( ) A cláusula SELECT.

( ) As cláusulas SELECT, FROM E WHERE.

Questão 7 - O que acontece após a execução do comando: SELECT

TRY\_CAST(‘abc’ AS INT).

( ) Um erro é gerado Um valor. *Só gera erro qdo a conversão é explicitamente proibida, tipo datetime as xml*

(X ) null é retornado. *Inclusive bastante usando com case when para lidar com exceções.*

( ) Um valor inteiro é retornado.

( ) Uma string é retornada.

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. *Having usa só depois do Group by*

( ) 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.

(X ) Crie um DTSConfig para fornecer informações de conexão para o pacote em tempo de execução.

*Ele gerou o erro pq não conseguiu ler a senha encriptada durante tempo de execução do pacote. A solução indicada pela Microsoft foi colocar Windows autentication, encriptar o pacote com uma das opções acima e criar um arquivo de config para fornecer info de conexão durante tempo de execução do pacote.*

*Referência:* <https://techcommunity.microsoft.com/t5/sql-server-support/ssis-error-code-dts-e-oledberror-an-ole-db-error-has-occurred/ba-p/315868>

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. *Eu usava para pegar nomenclaturas parecidas, tipo:*

*SELECT DISTINCT NomeEmpresa, CNPJ\_Raiz, Cidade, Estado FROM TabelaFato WHERE NomeEmpresa LIKE ‘PAPELARIA FOFONILDA %’.*

*Assim eu pegava a Papelaria Fofonilda LTDA, Papelaria Fofonilda LTD, Papelaria Fofonilda LTDA., Papelaria Fofonilda LTDA EM RECUPERAÇÃO JUDICIAL e etc, e catalogava as diferentes nomenclaturas para futuramente usar de Base de treinamento para um algoritmo de classificação (machine learning supervisionado) no qual eu tenho controle dos rótulos (NomeEmpresa\_Padronizado).*

( ) ​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. *Não apaga informação depois de gravada.*

( ) ​Integração. *Mesmas unidades de medida para o mesmo atributo entre todas as fontes*

( ) ​Variação de tempo. *Acumula dados no tempo*

( ) ​Orientado por assunto. *Informação por assunto ao invés de transação/operação*

(X ) ​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).

*Star Schema: É um modelo dimensional ou multidimensional de dados relacionais. Contém uma tabela Fato central com as tabelas das Dimensões (caracterizações dos atributos como produto, tempo, geolocalização, nomes e nomenclaturas, pessoas e etc) ligadas diretamente à Tabela Fato através de relações P/FK (Chaves primária e estrangeira). As dimensões não são normalizadas. Mais usado em Data Marts para necessidades específicas dos departamentos de uma organização (visão bottom-up).*

*SnowFlake Schema: No modelo dimensional do Inmon, ele normaliza as tabelas de dimensão do Star Schema criando outras ligações de P/FK a partir das dimensões, ocupando menos espaço de armazenamento dessas tabelas e diminuindo a redundância de dados, porém aumentando a complexidade e tempo das consultas (quantidade de joins e tempo de processamento da consulta dependendo da quantidade de níveis das tabelas dimensão). Usado em arquiteturas centralizadas de DWs (visão corporativa ou top-down).*

Questão 23 ​– O que podemos entender por “Granularidade do dado”?

*Gosto de explicar que a granularidade dos dados é o menor conjunto de atributos e características que fazem de uma linha de uma tabela ser única. Por ex., se eu tiver um registro de uma tabela que tem nome do cliente, id do cliente, mês e ano de compra, e o valor de compra, o menor "grão" desses dados é o valor mensal de compra do cliente, não especificado por produto. Se alguém quiser info de produto, eu precisaria diminuir a granularidade dos dados (diminuir o grãozinho da informação), ou seja, incluir novos atributos para caracterizar um registro único na tabela, nesse caso, produto, o que aumentará a quantidade de registros e deixará mais específica a informação. Mais uma vez, se eu quiser a info de venda diária, precisarei especificar mais uma informação, adicionando o atributo de dia à tabela, aumentando mais uma vez a quantidade de registros, dando maior especificidade à informação.*

*A definição formal é que refere-se ao menor nível de informação de um modelo de dados (não necessariamente fica claro sem exemplos, pra quem não é de tecnologia).*

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 *(hehehe hehe).* 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 and you 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:

*Apache Kafka would be a great choice to provide online real time and easily scalable solution, as it deals with integration (ETL), streaming and flexible architectures (not just Kafka -> Kafka). As Kafka -> Data Base is a possible architecture, the storage and analytics can come in different ways, for example, the analytics (like ML, DL etc.) can be incorporated as part of a Kafka -> Spark architecture, as Spark deals with processing and structured streaming to visualization report tools and store on databases for further historical analysis.*

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

*Athena is a serverless analytics service, where you don’t have to manage infrastructure and it’s used to deploy code that uses a wide variety of data types and sources that can be either structured or non-structured, while in Redshift Spectrum you have to manage infrastructure like the mandatory Redshift Clusters as Spectrum doesn’t run without Redshift, so all the managed services of Redshift will have to be, humm... well, managed. Spectrum is a service to run SQL queries directly against all your data.*

*That being said, Athena is a good choice for initial exploratory analysis while Spectrum is a good choice for Redshift users, with bigger workloads on multiple Redshift tables that need to be combined.*

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??

*I didn’t know before your question, (like many Big Data specific questions, don’t tell my future boss, rsrs), but it seems that AWS has a Serverless Image Handler called* ***AWS Lambda*** *that retrieves images from your S3 bucket and uses Sharp to return a modified version of the image to the Amazon API Gateway.*

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.

*What, exactly, is the question? It’s not clear to me.*

*Kinesis Data Stream is used to collect and process streams in real time. If something needs to be delivered to Redshift or S3, it might be more helpful to use Kinesis Data Firehose, and if something needs to be analyzed, it might be more helpful to use Kinesis Data Analytics that analyses streaming data using SQL.*

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.

*ELT – I like to describe it as Load first, ask questions later, rsrs. I used it (a weaker form of it) on my HP business case: we* ***had*** *to have all tables, record by record, loaded on our infrastructure before knowing which were the Transformations rules, no matter the format or lack of it.*

*So, theoretically, the concept here is that there’s no Data Staging step once the Data Warehousing process performs basic transformations. It’s faster but less reliable, less secure and less compliant to some standards like LGPD. As you perform Transformations on a*

*ETL – Well, old school, but well developed, secure and compliant process used on data integration. It Extracts, Transforms and Loads into a data storage system, and it needs a very structured dataset and a well-defined workflow. Once on DW it’s faster to perform data analysis tasks.*

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

•Relational: *it’s a Structured Database (SQL) where data gets fit into a pre-defined format. Use cases are transactional data of an e-commerce web-site or a hospital database with patients’ registration. Examples are MSSQL and PostgreSQL.*

•Key Value: *it’s a NoSQL database that stores pairs of key and values. Dynamo and Voldemort are examples of it. They can be used to manage streams of non-transactional data like application logs and real-time page view counters.*

•Documents: *it’s a NoSQL database that saves a document, or a file (JSON and JSON-like for instance), as a whole or a single entity. MongoDB and CouchDB are examples of it. Uses vary as they are the most common NoSQL databases, but newspapers store articles as a whole document, or a facebook post can considered a single entity.*

•Graphs: *it’s a NoSQL database that uses graph theory, like nodes and edges to represent entities and relationships, to store, map and query. It’s used when the relationships between data points matter more than the individual points themselves. Fraud detection case uses Neo4j to detect fraud rings comprised of stolen and synthetic identities. And there’s an example of recommendation engine using InfiniteGraph.*

*By the way, I also found a paper proposing data management in large scale healthcare system using graph database. Haven’t read it yet.*

•Time series: *it’s a database optimized to store time series data as pairs of time and values. For example, a good use is to store measures from a geochemical sensor, or a device feeding the geolocation of cars (insurance companies will love it, rsrs). Examples of it is TimescaleDB (postgresql extension) and InfluxDB.*

•In-Memory: *it’s a database that relies on main memory for faster querying. They lack the durability property of a DBMS. Examples of it are Redis and TimesTen (Oracle’s). It’s said it’s used on applications requering real-time response like financial (like tranding) and defense (like flying objects detection) industries.*

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; *It’s not clear if you want a value for each day or one value for both days, so I did both cases. Saturday\_mean and sunday\_mean for the first case and satur\_sun\_days\_mean for both.*

3- To be able to provision your entire environment in a public cloud, preferably AWS. *(I did not understand what the question is.)*