## Teste teórico de analista de Business Intelligence

Nome do Candidato (a): Luciano Correa dos Santos

Data 08 / 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?

(**X**) Clustered Index.

( ) Nonclustered Index. ( ) Bitmap Index.

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

( ) 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. ( **X** ) null é retornado.

( ) 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.

( ) 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.

( ) 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.

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

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

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

( ) Variação de tempo.

(**X**) 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 rea. ( ) 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).

Start Schema (Ralph Kimball) – é um dos modelos dimencionais mais utilzados em um processo de BI. O mesmo composto por uma tabela central chamada de fato, ligada a demais tabelas chamadas dimensões. Por isso o nome de Esquema de Estrela ou Star Shema. Embora tenho uma questão de ocupar muito espaço, esse modelo é mais simples de implementar.

SnowFlake (Bill Inmon) – diferente do modelo Start Schema, esse modelo ocupa menos espaço por ser nomalizado, embora atenda ao porpósito de facilitar a tomada de decisão, o mesmo é mais complexo de implementar.

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

É o menor nível da informação desejado, ou seja, o maior detalhamento possível dos dados. Quanto maior for o detalhamento, maior será a flexibilidade para se obter respostas. O nivel de granularidade por ser compreendido nas operações de Drill Down e Drill Up.

# 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 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:

**Proposta de solução**

Coleta de dados da pulseira

Criar uma API (Paython) para consultar os dados da Pulseira no respectivo repositório

Incluir a APIS em uma fila na AWS para disponibilizar a consulta dos dados em tempo real

Simple Queue Service (SQS)

Tratamento e armazenamento dos dados

Criar uma API em Python para ler/tratar dados da fila da AWS

Realizar o tramento de dados e para amazenar em Datamarts do Elastich Search

Gerar os respectivos Json e inclui-los nos respctivos Datamarts

Criar Datamarts no Elastich para amazenamento

Visualização de dados

Utilizar uma ferramenta robusta para criação de Dashboards

Kibana (seria a pricipal opção pelo fato de utilizarmos o Elastich como repositário)

Tableau Desktop (ótima ferramenta para criação de Dashboard e com curva de aprendisado baixo)

Power BI (ótima ferramenta para criação de Dashboard e com grande quantidade de profissionais qualificados no mercado)

Utilizar AWS (SQS) para falicitar a escalabilidade e obter dados em tempo real

Utilizar o Elastich como repositório (formato Json), para agilizar o acesso e possibilitar rapidez na visualização dos mesmos

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

Amazon Athena é um serviço de consultas de dados no Amazon S3 usando SQL padrão, e não necessita de infraestrutura para gerenciar, isso facilita o seu uso, basta apontar para o Amazon S3 e definir o esquema para iniciar as consultas.

A vantagem desse serviço é a facilidade do uso da linguagem SLQ padrão.

Amazon Redshift Spectrum recuperar dados estruturados e semiestruturados dos arquivos no Amazon S3 através de consultas, porém sem a necessidade de carregar no Amazon Redshift.

A grande vantagem desse serviço é recuperar grandes quantidades de dados trabalhando com paralelismo amssivo, em servidores dediados do Amazon Redshift.

1. 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??

Criar um código em paython utilizando PIL (Python-Imaging), para converter as cores das imgens em lote.

1. 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.

1. Which technologies below are related to Big Data on Cloud?
   1. Kubernetes, Jenkins, Terraform
   2. Azure SQL Server, AWS Lambda, AWS EC2
   3. **Google BigQuery, Apache Spark, Amazon Redshift**
   4. Digital Ocean, Packet, Javascript
   5. AWS, Google, Facebook
2. Which file type is the best to read/write tabular data on big scales?
   1. **CSV**
   2. Protobuf
   3. Gzip
   4. Parquet
   5. JSON
   6. Avro
3. Choose all correct answers To real-time data processing which technology is best for the streaming layer?
   1. **Apache Kafka**
   2. MySQL
   3. MongoDB
   4. Python
   5. Apache Spark
4. Explain the main points that define the concepts of ELT and ETL.

ETL é o pipeline tradicional para movimentar dados em fontes para um ambinde de Data Warehouses onde,

Extrair: recuperar dados de fontes

Transformar: transformar e formatar dados para integração

Carregar: transfira dados para o armazém de dados de Data Warehouses

ELT é uma varição do pipeline tradicional de movimentação de dados, alterando a ordem dessa movimentação, onde a extreção e a carga sõa realizadas,

antes da tranformação, devido o aumento no volume de dados, atualmente os pipeline de movimentação de dados, estão adotando esse processo, deixando a tranformação para um processo

de tranformeção de dados hospedado em nuvem (por exemplo), melhorando e agilizando acesso a infomação.

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

* Relational:

Banco de Dados relacional, é o mais utilizado na grande maioria de sistemas de diverssas empresas. Esse modelo é baseado em armazenamendo de dados em tabelas que estão relacionados entre si por chaves de identificação.  
Exemplos: MySQL, SQLServer.  
Sistemas: ERP, Cadastro de Clientes.

* Key Value:

Banco de Dados NoSQL do tipo Chave Valor e o mais utilizado, armazenamento é realizado direto em sistema de arquivos, e com grande vatagem de trabalhar rapidamento com grande volume de dados.

Exemplo: MongoDB, Hive  
Sistemas: Facebook, BigData (Haddop)

* Documents:

Banco de Dados NoSQL do tipo Documents é utilizando para armazenar documentos tipo JSON, facilitando o armazenamento e consulta por sistemas que utilizam o mesmo formato de dados. (JSON).  
Exemplo: Amazon DocumentDB, Azure Cosmos DB

* Graphs:

Banco de Dados NoSQL são formados por vértices (node) que são ligadas por arestas (relationships). Muito utilizado em aplicções moveis e jogo.

Exemplo: Amazon DynamoDB, Neo4j  
Sistemas: Netflix

* Timeseries:

Banco de Dados Timeseries são utilizados para armazenar Séries Temporais por meio de pares de tempos e valores associados.  
Exemplo: eXtremeDB

* In-Memory:

Banco de Dado In-Memory é um sistema que gerecia os dados em menmoria principal difrente da maiorio dos demais que utiliza o HD, porém o grande problema desse tipo é a volatilidade da memória RAM.  
Exemplos: SAP HANA, Oracle Database In-Memory

Sistemas: SAP

# 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.