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

→Start Schema – (Ralph Kimball)

O esquema estrela é composto no centro por uma tabela fato, rodeada por tabelas de dimensão, ficando parecido com a forma de uma estrela.

A ideia é propor uma visão para modelagem de base de dados para sistemas de apoio à decisão, que é o caso do Data Warehouse.

O processo começa com a extração, a transformação e a integração dos dados para um ou mais Data Marts. Estes são modelados através de um modelo dimensional.

Os projetos serão menores, independentes, focando áreas ou assuntos específicos (departamento, por exemplo). Este tipo de implementação permite que o planejamento e desenho dos Data Marts possam ser realizados sem esperar que seja definida uma infraestrutura corporativa para o Data Warehouse.

A intenção é que os Data Marts sejam implementados de maneira incremental.

→SnowFlake – (Bill Inmon)

O Snowflake também é projetado para suportar tomada de decisão, mas economizando espaço em disco. Aqui, várias dimensões se relacionam com uma tabela de fato, de modo que os dados ficam em uma cascata (hierarquia) e mais distantes da tabela de fato. O dado passa por todos os outros objetos até atingir o destino, a análise. Este esquema tem uma ótima governança de dados e não apresenta redundâncias. Entretanto ele apresenta mais lento e mais complexo, pois quando necessitamos analisar o nível menos granular é necessário relacionar todas as tabelas durante a análise até atingir a análise desejada

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

O mais importante aspecto do projeto de um DW é a questão de sua granularidade. A granularidade diz respeito ao nível de detalhe ou de resumo contido nas unidades de dados existentes no DW. Quanto mais detalhe, mais baixo será o nível de granularidade e quanto menos detalhe, maior será o nível de granularidade.

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:

Nesse processo poderia utilizar o Kafka(streaming) ➔ NIFI, usar o processor jolt para transformar o arquivo de JSON e fazer o tratamento. Após utilizar o conversor para avro → e fazer o envio para HDFS, e através de uma external ver a tabela no Hive.

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

Resposta:

O Amazon Athena usa o Presto com suporte completo a SQL padrão, e funciona com diversos formatos de dados padrão, como CSV, JSON, ORC, Avro e Parquet. O Athena consegue lidar com análises complexas, inclusive grandes associações, funções de janela e matrizes. O Amazon Athena é a forma mais fácil de rodar queries ad hoc para dados no S3 sem a necessidade de configurar nem gerenciar nenhum servidor.

O Amazon Redshift tem a maior rapidez no desempenho de query para relatórios corporativos e cargas de trabalho de business intelligence, especialmente naqueles que envolvem SQL extremamente complexo com múltiplas junções e subqueries.Um data warehouse como o Amazon Redshift é a sua melhor opção quando você precisar reunir em um formato comum dados de várias fontes diferentes – como sistemas de inventário, sistemas financeiros e sistemas de vendas a varejo – e armazená-lo por longos períodos, de forma a criar relatórios de negócios sofisticados com base em dados históricos; nesse caso, um data warehouse como o Amazon Redshift é a melhor escolha.

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??

Com as operações em lotes do S3, você pode copiar objetos entre buckets, substituir conjuntos de tags de objetos, modificar controles de acesso e restaurar objetos arquivados no Amazon S3 Glacier, com uma única solicitação à API do S3 ou com alguns cliques no Console de gerenciamento do Amazon S3. Também é possível usar as operações em lotes do S3 para executar funções do AWS Lambda nos objetos para executar lógica de negócios personalizada, como processamento de dados ou transcodificação de arquivos de imagem.

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

→ETL (extrair, transformar, carregar)

Nesse processo, os dados são retirados (extraídos) de um sistema-fonte, convertidos (transformados) em um formato que possa ser analisado, e armazenados (carregados) em um armazém ou outro sistema

→ELT (extrair, carregar, transformar)

Extrai dados de um sistema-fonte, os carrega em um sistema de destino e, então, usa o poder de processamento do sistema-fonte para conduzir as transformações. Isso acelera o processamento de dados porque acontece onde os dados estão.

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

→Relational

Um **banco de dados relacional** armazena **dados** em tabelas. Tabelas são organizadas em colunas, e cada coluna armazena um tipo de **dados** (inteiro, números reais, strings de caracteres, data, etc.). Os **dados** de uma simples "instância" de uma tabela são armazenados como uma linha.

→Chave e Valor

Armazenamento de informações de Sessão, perfis de usuários e preferencias, Dados de carrinhos de compras, mas não é bom para relacionamento entre dados, transações com múltiplas operações, consulta por dados e atributos e operações por conjuntos. Exemplos (Azure Table Storage, Riak, Redis, Voldemont, DynamoDB)

→Documents

Acesso fácil ao atributos internos dos documentos, uso de visões materializadas para agregar informações ou estabelecer consultas especificas, e possibilita realizar consulta dos dados dentro do documento a nível de atributo. Limitação para armazenamento, pois os documentos devem ser de uma mesma coleção.

Exemplos (MongoDB, CouchDB, Terrastore)

→Graphs

Elementos principais, nós e relacionamentos, baseado na teoria de grafos, permite armazenar relacionamento entre entidades, permite encontra padrões interessantes entre nós.

Exemplos(Neo4j, Infinite Graph, Trinity)

→Timeseries

Os dados de serie temporal são um conjunto de valores organizados por tempo. Exemplos comuns de dados de serie temporal incluem dados de sensor, preços de compra de ações, dados de fluxo de cliques e telemetria do aplicativo. Os dados de

serie temporal podem ser analisados quanto a tendências históricas, alertas em tempo real ou modelagem preditiva. Dessa forma, os dados de serie temporal são mais bem visualizados com gráficos de dispersão ou de linha

→In-Memory

Arquitetura In-**Memory**. O Oracle **Database In-Memory** fornece uma arquitetura de formato duplo exclusiva que permite que as tabelas sejam representadas simultaneamente na memória usando o formato de linha tradicional e um novo formato de coluna in-**memory**.