DADOS DO ALUNO(A):

Nome: Pedro Penido Gama

Matrícula: 149820

**AVISO**

1. Entregue essa atividade até o prazo estipulado no sistema CANVAS.
2. Essa atividade é individual. Responda da forma como assimilou o que foi ensinado.

EXERCÍCIO 1 - Conceitos

1. Analise os seguintes casos de uso:
   1. Empresa deseja criar uma solução para integrar e analisar dados de seu CRM. Ela possui dados desde 2002, com formato estruturado e já conhecido pelos desenvolvedores. A necessidade de atualização das informações é diária e o volume de crescimento do CRM é algo em torno de 1GB/mês. Apenas usuários de negócio analisarão as informações para identificar as vendas realizadas no tempo e conseguir gerar relatórios de acompanhamento da saúde financeira da organização.
   2. Empresa deseja criar uma solução para integrar dados de seus usuários, cruzando as informações com dados de redes sociais e outras fontes de dados que podem ser adquiridas no mercado. Seu propósito é gerar modelos que consigam identificar usuários com potencial de compras para determinados produtos em um intervalo curto de tempo e com possibilidade de novas fontes de dados serem incluídas.

Quais abordagens você utilizaria para cada um dos casos de uso? Explique-as.

1. Com as suas palavras, explique a diferença entre ETL e ELT.
2. O Big Data traz a necessidade de novas abordagens para o processamento de dados. Discuta a aplicação de, pelo menos, 4 dos 7 “V’s” apresentados sobre o tema.
3. Cite 5 vantagens da utilização de Data Lakes.
4. Cite 5 diferenças entre um Data Warehouse e um Data Lake.
5. Qual a diferença entre processos *Batch* e Processos *Streaming*?
6. Na modelagem de domínio, qual a diferença entre a visão MACRO e MICRO?

**RESPOSTAS:**

1 - a) - Utilizaria a abordagem ETL devido a capacidade de armazenar dados tratados com baixa necessidade de atualização das informações e porque não e preciso visualizar o dado em tempo real.

b) - Utilizaria a abordagem ELT devido a necessidade de gerar modelos que consigam identificar usuários com potencial de comprar para determinado produto em um CURTO tempo.

2 - A diferença entre ETL e ELT vai além da ordem em que as etapas são executadas. E em ambos os casos não existem pior nem melhor, existe aquele que melhor se adequa a necessidade da empresa. No ETL (extration, transformation e load), os dados são tratados e armazenados, com fácil implantação e não há necessidade de acesso ao dado em tempo real. Já no ELT (extration, load e transformation) os dados são transformados diretamente no sistema destino, se dá bem com uma grande quantidade de dados, podendo trata-los em tempo real.

3 -

1. Volume: O Big Data agrupa uma enorme quantidade de dados que são gerados a cada segundo. É só imaginar todos os e-mails, vídeos, fotos e mensagens que circulam nas redes diariamente. Assim, o BD atua de maneira a lidar com esse volume de dados com eficiência, tornando possível o seu agrupamento através de softwares.

2. Velocidade: É a agilidade com a qual os dados são produzidos e manipulados. O Big Data vai analisar os dados no instante em que são criados sem precisar armazená-los. Isso acontece com as transações de cartão de crédito, viralização de mensagens em redes sociais, publicações em sites e blogs, entre outras.

3. Variedade: Os dados podem ser gerados em vários formatos estruturados (numéricos) ou não estruturados. Nessa última categoria, estão incluídos os arquivos de áudio, vídeo, e-mail, textos e cotações e transações financeiras.

4. Valor: De nada adianta ter acesso a uma grande quantidade de informação se ela não puder agregar valor, certo? Pode-se dizer que o valor do Big Data está na análise precisa dos dados e nas informações e insights fornecidos para as empresas a partir do seu conteúdo.

4 - Cinco vantagens Data Lakes: velocidade de adição de dados. Ter menor custo para implementação. Não exige a estruturação e organização de dados e, por isso mesmo, permite análises em tempo real e por fim maior escalabilidade.

5 - Cinco diferenças entre DW e DL:

|  |  |  |
| --- | --- | --- |
| **Característica** | **Data Warehouse** | **Data Lake** |
| **Dados** | Estruturado, processado, analisado | Processado, não analisado |
| **Propósito** | Motivo pré-definido para armazenar o dado | Motivo pré-definido para armazenar o dado |
| **Schema** | Schema on Write (desenvolvido antes da implementação do DWH) | Schema on Read (Gerado no momento da análise) |
| **Casos de uso** | BI, Reporting | ML, Exploração de Dados, Preditivo, Inteligência Artificial |
| **Usuários** | Analistas de Negócio | Cientistas de dados, alguns analistas de negócio |
| **Preço/Tempo** | Maior custo, maior tempo de implementação, manutenção simples | Menor custo, implementação rápida, manutenção difícil |

6 – Batch e o processo de carga de dados em lotes, onde não há necessidade de analise em tempo real, subdividido em duas categorias: carga incremental onde a carga não sofre alteração ao longo do tempo e carga completa onde os dados são carregados sempre que o processo é executado, sobrescrevendo dados já existentes. Já o Streaming tem o processo de carga em tempo real, subdividido em duas categorias: Logs onde são executados grande carga de dados com baixa latência, e os pub/sub onde cria tópicos que recebem e “escutam” chamadas de serviços e envia para seus “assinantes”.

7 – O objetivo de uma visão MACRO é entender a relevância estratégica da entidade em análises exploratórias, sem a necessidade de ler seu conteúdo e analisar seus valores e atributos. O objetivo de uma visão MICRO é entender a estrutura da entidade, como os campos são gerados e se relacionam com outras entidades, como são calculados, quais são as regras de negócio inerentes ao item analisado e qual os níveis de agregação / granularidade pode-se avaliar no contexto da entidade.