Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
MBA em Inteligência Artificial e Big Data
– Curso 3: Administração de Dados Complexos em Larga Escala –

Questões sobre Preparação e Compressão de Dados
Prof. Dr. Caetano Traina Júnior

Exercício 1) Usar Sistemas Gerenciadores de Bases de Dados é importante para analisar "*big data*" porque eles permitem:

1. controlar quais dados cada aplicação pode acessar e/ou atualizar, garantindo a confidencialidade dos dados armazenados e dos processos de análise.
2. escalar os dados em quantidades crescentes que esgotam a capacidade de memória principal de aplicativos de análise e tratar os processos de preparo dos dados diretamente em sua fonte.
3. armazenar rapidamente grandes volumes de dados.
4. trabalhar com os dados garantindo o controle da concorrência e da consistência em todos os níveis de granularidade.
5. gerenciar muitos tipos distintos de dados e/ou muitos servidores para processamento distribuído.

Exercício 2) Uma `Windowm function` é equivalente a uma operação de agrupamento:

1. e calcula um valor para cada tupla, mas não preserva os atributos originais.
2. e calcula um valor para todo o grupo, além de preservar as tuplas originais, com todos os seus atributos.
3. onde o atributo de agrupamento pode ser ordenado na ordem inversa daquela indicada na cláusula `GROUP BY`.
4. mas o atributo de agrupamento pode ser preservado na cláusula `PARTITION BY`.
5. onde o atributo usado como argumento da função pode ser particionado pelo *frame* indicado pela especificação `GROUP | RANGE | ROW`.

Exercício 3) Funções de agregação podem ser usadas como `Window function`, mas o inverso nem sempre é verdade porque:

1. Uma `Window function` sempre requer que seja especificado um atributo como argumento, mas a função de agregação não (embora possa ser usado um '*', como na função `Count(*)`).
2. A função de agregação requer que exista um só valor para todo o grupo (como média, min, e max), e uma `Window function` pode retornar valores diferentes.
3. Existem `Windowm functions` que só podem ser usadas quando a ordem entre as tuplas é indicada. Por exemplo, a `moda` de um atributo só tem sentido quando ele é usado de maneira ordenada.
4. Existem `Windowm functions` que só podem ser usadas quando a ordem entre as tuplas é indicada.
5. Essa afirmação é falsa, pois todas as `Window functions` também pode ser usadas como funções de agregação

6. As `Windowm functions` requerem a que cada *frame* seja explicitado, o que não ocorre se a função não for usada como `Window function`.

Exercício 4) A redução da **cardinalidade** de um conjunto de dados pode ser feita:

1. com amostragem dirigida, para preservar sem distorções as propriedades de cada classe, ou com amostragem aleatória, para garantir que todas as classes do conjunto estejam representadas.
2. com amostragem aleatória, para agilizar a geração das amostras, ou com amostragem por densidade, para explorar determinadas regiões de interesse no espaço de dados original.
3. usando funções de geração de valores aleatórios para escolher as tupla nas tabelas armazenadas no SGBD, ou usando a cláusula `TABLESAMPLE` para escolher as tabelas armazenadas.
4. com amostragem aleatória, para preservar sem distorções as propriedades da distribuição dos dados no conjunto original, ou com amostragem dirigida, para enfatizar alguma tendência que ser queira ressaltar no conjunto de dados original.
5. definindo uma taxa de amostragem para cada *bin* do histograma, ou definindo uma quantidade aleatória de amostras para cada *bin*.

Exercício 5) Para executar uma amostragem baseada em Histogramas:

1. Primeiro separam-se os atributos segundo as classes originais, em seguida efetuam-se os processos de discretização e particionamento, e finalmente escolhem-se as amostras mantendo as classes com as proporções ordenadas pelo tipo de histograma usado.
2. Primeiro discretizam-se os atributos usados para construir o histograma na sequencia dos *bins*, e finalmente efetua-se a amostragem segundo as quantidades de tuplas em cada faixa discretizada.
3. Primeiro se calcula o histograma dos atributos originais que devem ser medidos para a amostragem, em seguida efetuam-se os processos de discretização, e finalmente escolhem-se as amostras equilibrando a propriedade segundo o tipo de histograma usado.
4. Primeiro se separam as tuplas em faixas contínuas dos valores dos atributos de classificação, a seguir efetua-se a amostragem em cada faixa, e finalmente constrói-se o histograma de acordo com o tipo indicado.
5. Nenhuma das anteriores, pois a sequência de construção depende do tipo de histograma.