Arquitetura de SGBD Relacionais
— Métodos de Acesso Físico —

Caetano Traina Jr.

Grupo de Bases de Dados e Imagens
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo - São Carlos
caetano@icmc.usp.br

3 de abril de 2013
São Carlos, SP - Brasil

Esta apresentação trata dos conceitos de como Histogramas são usados em Bases de Dados, como uma maneira de prover estatísticas sobre a distribuição dos dados nas relações.

Outline

Motivação

Recordação

- Um SGBDR opera em um de três modelos de previsão de seletividade e de custo (Modos de Operação):
  1. Nenhuma informação conhecida dos dados: Distribuição Constante.
     Assume-se Distribuição constante (usualmente adotada como 10%)
  2. Faixa de valores conhecida: Distribuição Uniforme.
     Assume-se Distribuição Uniforme (*Uniform distribution assumption)*
  3. Distribuição de valores conhecida: **Histograma**.
     Assume-se uma aproximação de frequência de valores que pode ser diferente para cada valor.
- Histogramas provêm a informação mais detalhada sobre a Distribuição de Dados no domínio dos atributos.

Motivação

Recordação

- Informações sobre a distribuição dos dados armazenados são muito úteis, mas o espaço necessário para descrever a distribuição com precisão tem que ser pequeno;
- Histogramas são uma maneira de representar a distribuição de maneira compacta, e ajustável para atender o compromisso entre as necessidades e os recursos (de memória e tempo) disponíveis.
- As principais aplicações de histogramas em SGBDR incluem:
  - Estimar a seletividade de um ou mais Atributos em modelos de custo de execução;
  - Pré-execução de consultas (aproximadas) para realimentação de usuários (*pre-execution user-level query feedback*)
  - Estimar como particionar dados para o balanceamento de carga na execução de junções em processamento paralelo;
  - Estimar alternativas de uso de comunicação em processamento distribuído;
  - ...

Motivação

Introdução

- Histogramas são aproximações da Distribuição da Frequência dos Valores dos atributos nas relações da base de dados.
- A frequencia pode ser:

Absoluta – por exemplo, a contagem de ocorrências de cada valor; ou

Relativa – Por exemplo, a porcentagem de ocorrências de cada valor em relação à contagem total.

Motivação

Introdução

- Qualquer mecanismo de estimação de dados em um SGBDR deve:
  - Ser rápido para ser usado;
  - Ser rápido para coletar as estatísticas necessárias;
  - Ter baixo custo de memória para armazenar as estatísticas.

Motivação

Definição

- Um Histograma é uma aproximação da Frequência de Distribuição de Valores de um atributo ou de um conjunto de atributos X, que:
  - Registra os pares $\langle Valor, Frequencia \rangle$ em que o domínio de valores é particionado em β $(1 \leq \beta \leq Dom^*(X))$ "Faixas" (*Buckets*);
  - A frequência de cada valor em uma faixa é aproximada pela média das frequências de todos os valores daquela faixa.
- Particularmente, hos histogramas usados em Sistemas de Gerenciamento de Bases de dados:
  - As faixas de valores formam uma partição do domínio ativo dos atributos (não existe sobreposição);
  - Adota-se uma representação compacta para indicar os valores de uma faixa (usam-se valores em uma faixa contínua ou estruturas de índice);

Motivação

Definição

- Um histograma com uma única faixa aproxima todos os valores para a média de frequências do domínio ativo, portanto corresponde à *uniform distribution assumption* e ao modo 2 de previsão de seletividade.
- Portanto, o Modo de Operação 2: **Distribuição Uniforme** é um caso particular do Modo de Operação 3: **Histograma** em que $\beta = 1$;
- e um histograma assim é chamado Histograma Trivial.

Motivação

Exemplo

- Vamos considerar que exista uma relação de Professores.
- A Distribuição de Frequências dos valores de `Área` então é:

```
Professor = {Nome, Idade, Area} =
           {<Lucio,   34,  BD>,
            <Laura,   45,  IA>,
            <Lucia,   54,  CG>,
            <Lucas,   43,  RE>,
            <Luana,   38,  IA>,
            <Livia,   61,  MM>,
            <Lineu,   42,  IA>,
            <Lacir,   45,  BD>,
            <Luigi,   33,  BD>,
            <Lidia,   28,  ES>,
            <Ligia,   51,  RE>,
            <Licio,   43,  IA>,
            <Leila,   56,  IA>,
            <Luzia,   45,  CG>}
```

Area	Frequência
BD	3
CG	2
ES	1
IA	5
MM	1
RE	2

Classificação de Histogramas

- Existem várias classes de Histogramas:
  - Histogramas Equi-`Algo`
  - Histogramas Seriados
  - Histogramas de Extremos Detalhados (*End biased*)

Classificação de Histogramas

Histogramas Equi-Algo

- Histogramas registram pelo menos três dimensões dos dados:
  - Os Valores dos Dados;
  - A quantidade de valores em cada faixa; Equi-Faixa (*equi-width*)
  - A frequência dos valores em cada faixa. Equi-frequência (*equi-depth*)
- Histogramas em que a quantidade ou a frequência é mantida constante (ou o mais constante possível) são chamados Equi-Algo.

Motivação

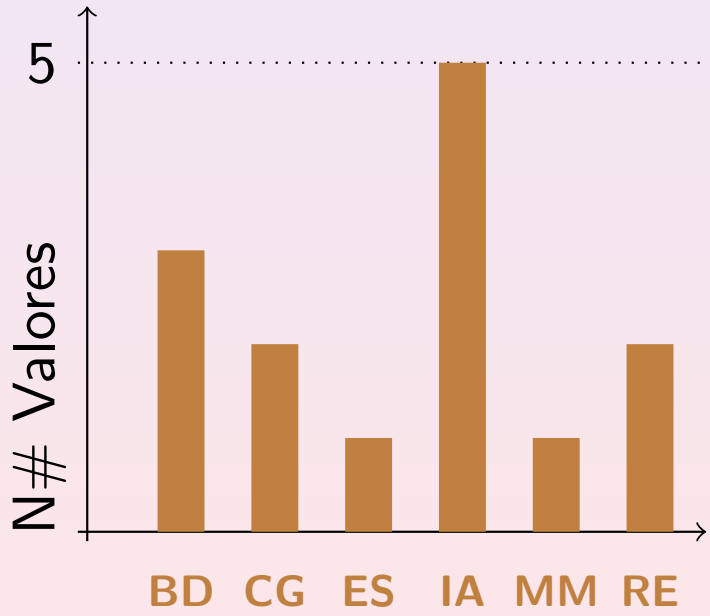
Praticamente todos os gerenciadores existentes, comerciais ou de domínio público, podem usar histogramas *equi-width* e *equi-depth* para a geração de estimativas, principalmente de seletividade.

Classificação de Histogramas

Histogramas Equi-faixa – *equi-width*

- Histogramas *equi-width* têm todas as faixas da mesma "largura" (mesma quantidade de valores por faixa), e registram a frequência dos valores em cada faixa.

Area	Frequência
BD	3
CG	2
ES	1
IA	5
MM	1
RE	2



Classificação de Histogramas

Histogramas Equi-faixa – *equi-width*

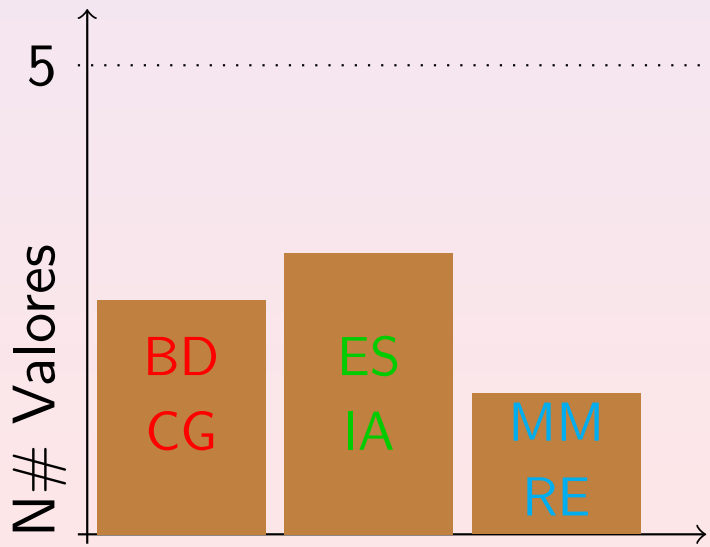
- O exemplo anterior é um Histograma *equi-width*, pois cada faixa representa exatamente um valor.
- Além disso, como o número de faixas é o mesmo da cardinalidade do domínio ativo do atributo ($Dom^*(\text{Area}) = 6, \beta = 6$), esse é um **Histograma Completo**.

Classificação de Histogramas

Histogramas Equi-faixa – *equi-width*

- Para reduzir o tamanho do histograma em domínios de cardinalidade grande, é comum que se coloque em cada faixa diversos valores.
- Por exemplo, agrupando as faixas como: {BD, CG}, {ES,IA}, {MM, RE} fica ($Dom^*(\text{Area}) = 6, \beta = 3$):
- Veja que a frequência registra a média da frequência dos valores em cada faixa.

Area	Frequência	
BD CG	3 2	2,5
ES IA	1 5	3,0
MM RE	1 2	1,5

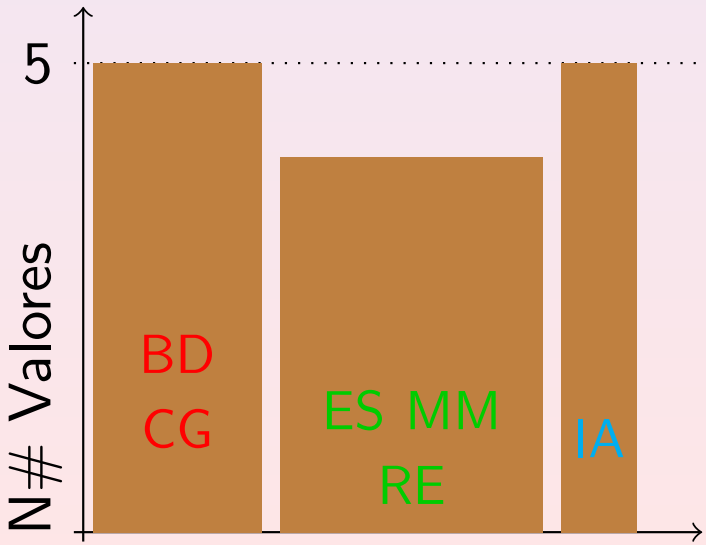


Classificação de Histogramas

Histogramas Equi-frequência – *equi-height*

- Histogramas equi-frequência (*equi-depth* ou *equi-height*) têm todas as faixas aproximadamente da mesma "altura" (frequência), e registram a quantidade total de valores em cada faixa necessários para atingir aquela frequência.
- Por exemplo, agrupando as faixas como: {BD, CG}, {IA}, {ES, MM, RE} fica

Area	Frequência
BD	3
CG	2
ES	1
IA	5
MM	1
RE	2

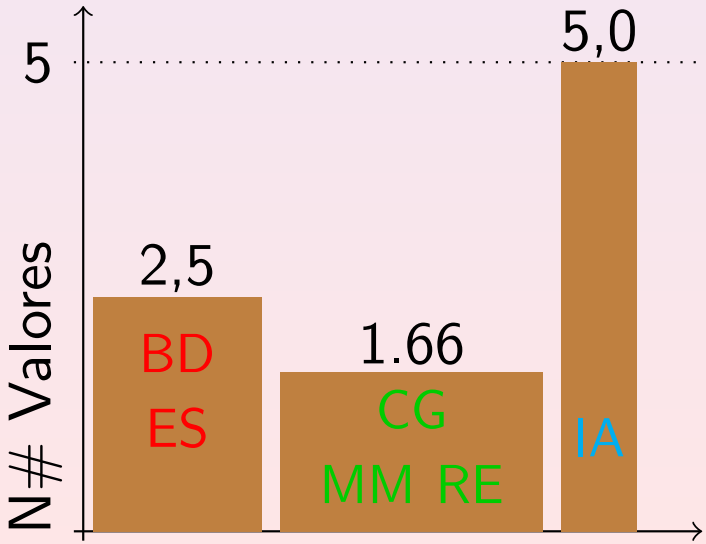


Classificação de Histogramas

Histogramas Equi-área – *equi-area*

- Histogramas *equi-area* são parecidos aos histogramas *equi-height*, mas ao invés de se procurar reduzir a variância da soma total de ocorrências dos valores do atributo em cada faixa, procura-se reduzir a variância da média de ocorrência deles.
- Por exemplo, agrupando as faixas como: {BD, ES}, {IA}, {CG, MM, RE} fica

Area	Frequência
BD	3
CG	2
ES	1
IA	5
MM	1
RE	2



Classificação de Histogramas

Histogramas Seriados

- Histogramas equi-`algo` cuidam da ordem dos valores apenas dentro de cada faixa.
- **Histogramas seriados** são organizados de maneira que as faixas são ordenadas pelo valor da frequência dos valores dentro de cada faixa **E** entre as faixas.
- portanto histogramas seriados são usados principalmente para domínios não contínuos, em que a ordem dos valores não seja importante (tipicamente para valores discretos).
- Veja que um histograma seriado pode ser também equi-algo ou não.

Classificação de Histogramas

Histogramas Seriados

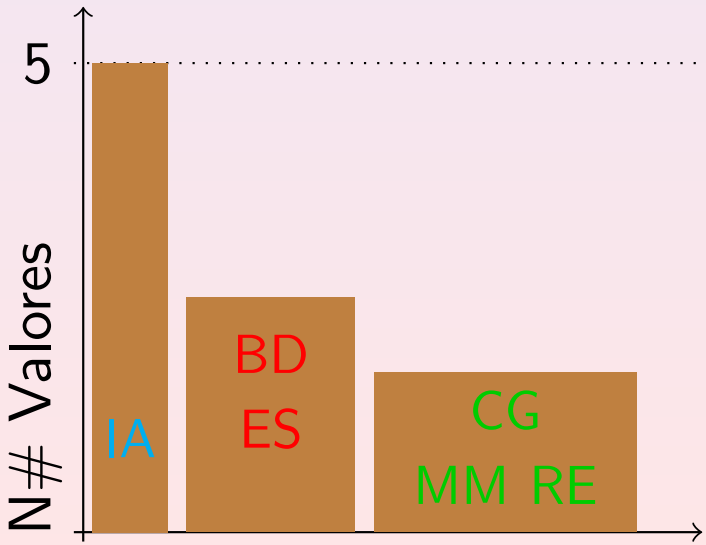
- Os histogramas seriados podem ser crescentes ou decrescentes.
- Em bases de dados os histogramas seriados são sempre decrescentes, assim sempre que for usado o termo "histograma seriado" nesta apresentação, sub-entende-se um Histograma Seriado Decrescente.
- Um histograma seriado é construído ordenando todas as faixas pela frequência, de maneira que as frequências dos atributos associados a uma faixa são todas estritamente maiores (seriado decrescente) do que as frequências dos atributos das faixas seguintes.
- Portanto, as faixas de um histograma seriado **agrupa** as frequências que são próximas, sem intercalação.

Classificação de Histogramas

Histogramas Seriados

- Nenhum dos histogramas mostrados como exemplo até agora são seriais.
- Por exemplo, o último exemplo do Histograma Equi-área, não é seriado mesmo se as faixas forem re-ordenadas:
- porque as faixas dos atributos que têm os atributos {BD, ES} e {CG, MM, RE} têm frequências iguais.

Area	Frequência
BD	3
CG	2
ES	1
IA	5
MM	1
RE	2

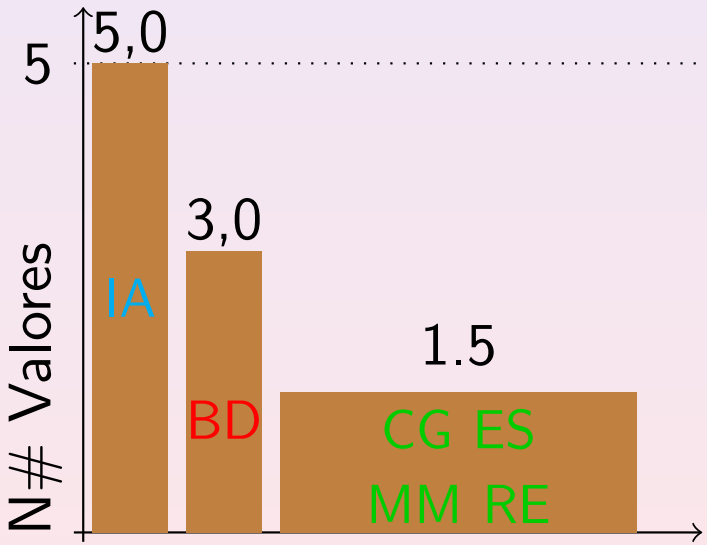


Classificação de Histogramas

Histogramas Seriados

- O seguinte é um exemplo de um histograma seriado com três faixas:

Area	Frequência
BD	3
CG	2
ES	1
IA	5
MM	1
RE	2



Classificação de Histogramas

Histogramas de Extremos Detalhados (*End biased*)

- Um histograma de Extremos Detalhados (*End biased*) é um caso especial de histograma seriado:
- Em um histograma *End biased*, uma determinada quantidade de atributos dentre os de mais alta frequência e/ou os de mais baixa frequência são representados explicitamente em faixas individuais (extremos detalhados),

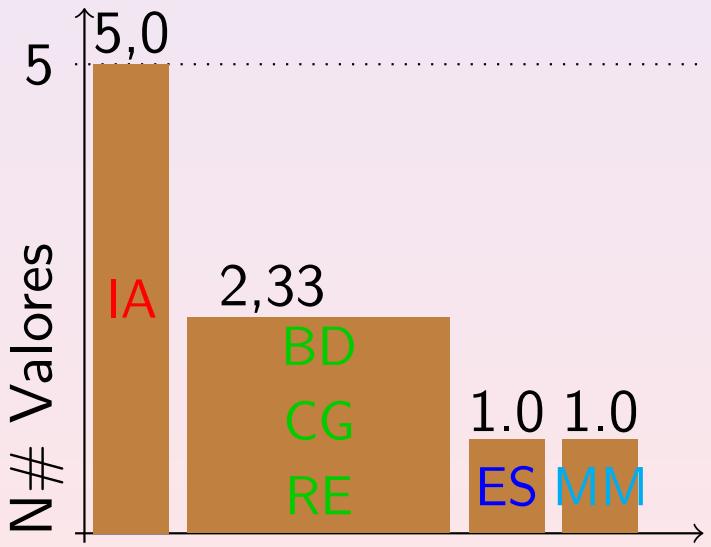
  e os atributos com frequências médias são todos representados em uma única faixa central.
- Note-se que poderia ser pensado em criar mais de uma faixa para as frequências médias, mas isso não traz benefícios: ele seria um histograma seriado teria todos os problemas inerentes aos histogramas seriados sem nenhum benefício adicional.

Classificação de Histogramas

Histogramas de Extremos Detalhados (*End biased*)

- Para um exemplo de histograma *End biased*, pode-se agrupar os atributos nas seguintes faixas: {IA}, {BD, CG, RE}, {ES}, {MM}

Area	Frequência
BD	3
CG	2
ES	1
IA	5
MM	1
RE	2



Espaço de Histogramas

Conceitos

- Seguindo o conceito de que todos os elementos que compõem uma base de dados devem poder ser representados em uma base de dados (na meta-base), os vários histogramas, de diferentes tipos, são concebidos como elementos de um "Espaço de Histogramas".
- Para fins de seu uso em SGBDR, os histogramas são representados num espaço multidimensional multivariado que tem como dimensões:
  - Regra de Partição,
  - Aproximação da estimativa da faixa,
  - Aproximação da frequência,
  - Algoritmo de construção,
  - Garantia de limite do erro.

Espaço de Histogramas

Regras de partição – Classe de partição

Regras de partição

- A regra de partição indica como os dados do domínio a ser representado no histograma devem ser separados em faixas.
- Esta dimensão, por sua vez é indicada por quatro características mutuamente independentes:
  - Parâmetro para Ordenação,
  - Parâmetro Fonte,
  - Classe de partição, e
  - Particionamento.

Espaço de Histogramas

Regras de partição – Parâmetro para Ordenação

Regras de partição

- **Parâmetro para Ordenação** – esta característica indica qual é o parâmetro dos dados (ou extraído dos dados), que deve ser usado para ordenar as faixas do histograma.
- Usualmente são usados:
  - O Valor do atributo: V
  - A frequência calculada: F ou
  - A área da faixa (quantidade de valores representados na faixa vezes a frequência média da faixa): A.

Espaço de Histogramas

Regras de partição – Parâmetro Fonte

Regras de partição

- **Parâmetro Fonte** – esta característica indica qual é a propriedade da distribuição de dados que deve ser representada no histograma.
- Usualmente são usados:
  - a quantidade de valores distintos na faixa (*spread*): S
  - A frequência calculada: F ou
  - A área da faixa: A.

Espaço de Histogramas

Regras de partição – Classe de partição

Regras de partição

- **Classe de partição** – esta característica indica quais restrições existem (quando existirem) quanto a como formar cada faixa.
- O caso mais comum é a regra de agrupar e ordenar algum parâmetro de importância para o histograma, o qual é indicado pelo parâmetro de ordenação.
- Um outro caso é o dos histogramas de extremos detalhados, os quais são um caso particular de histograma ordenado.
- Aqui consideramos que um histograma é sempre ordenado, a menos que seja explicitamente indicado ser diferente.

Espaço de Histogramas

Regras de partição – Restrição de partição

Regras de partição

- **Particionamento** – esta característica indica qual é a restrição colocada sobre o parâmetro Fonte.
- Diversas restrições têm sido tentadas: as mais comuns são:
  - *equi-algo*,
  - *v-optimal*,
  - *maxdiff* e
  - *compressed*.
- O Particionamento *Equi-algo* (em inglês *equi-sum*) indica que deve se tentar manter aproximadamente igual o parâmetro fonte.

Espaço de Histogramas

Regras de partição – Exemplos

- A regra de partição de um histograma é expressa como: *regra*(*Ordenação*,*fonte*).
- **Histogramas equi-width**: a partição dos valores em faixas procura manter a quantidade de valores igual em cada faixa;
- portanto eles procuram manter "aproximadamente igual o total" (*equi-sum*) da "quantidade" dos "valores", ou seja ele será $\boxed{\textit{equi-sum}(V, S)}$.

- **Histogramas equi-depth:** a partição dos valores em faixas procura manter a frequência igual em cada faixa;
- portanto eles procuram manter "aproximadamente igual" (*equi-sum*) a "frequência" dos "valores", ou seja ele será $\boxed{\textit{equi-sum}(V, F)}$.

Espaço de Histogramas

Regras de partição – Particionamento *v-optimal*

- O Particionamento *v-optimal* procura separar as faixas de maneira que **a variância** dos valores do parâmetro fonte é minimizada.
- Procura-se minimizar:

$$W = \sum_{b=1}^{\beta} n_b V_b,$$

onde

n_b é o número de tuplas que têm valores representados na faixa b, e V_b é a variância dos valores representados na faixa b.

- Veja que a minimização dessa equação é de complexidade exponencial no domínio ativo do atributo e do número de faixas do histograma.
- No entanto, este é o histograma que possui o melhor desempenho teórico.

Espaço de Histogramas

Regras de partição – Particionamento *maxdiff*

- O Particionamento *maxdiff* procura separar as faixas de maneira que os pontos de quebra dos valores do parâmetro fonte ocorre nos pontos de maior diferença entre o final de uma faixa e o início da próxima.
- A complexidade de construção deste histograma é quadrática no número de faixas, mas ele possui um desempenho bem próximo do particionamento *v-optimal*.

Espaço de Histogramas

Regras de partição – Particionamento *compressed*

- O Particionamento *compressed* procura separar as faixas colocando os maiores valores do parâmetro fonte em faixas isoladas, e o restante é colocado em faixas usando particionamento equi-algo.
- Esse tipo de histograma tem desempenho que depende de como os valores são distribuídos nas faixas.
- Este tipo de histograma pode aproximar o desempenho de um histograma serial de extremos detalhados, com desempenho aproximadamente igual (as vezes até melhor) ao de um *maxdiff*.
- Ele pode ser construído facilmente, desde que exista um algoritmo para escolher o número de faixas (ou usando o DBA para isso!).

Espaço de Histogramas

Aproximação da estimativa da faixa

Aproximação da estimativa da faixa

- A menos do caso de um histograma completo, sempre é necessário indicar o valor que deve ser assumido para representar cada faixa.

☞ Note que isso é independente da regra para particionar os dados.

- Existem três alternativas principais para estimar o valor aproximado por uma faixa:
  - Para dados que cobrem a faixa de maneira continua e sem intercalação, assume-se que os valores em cada faixa são constantes, e que
    - **Continuous value assumption** – existe um valor fixo independente da quantidade de valores por faixa. É equivamente a assumir uma distribuição constante por faixa;
    - **Uniform spread assumption** – armazena-se a quantidade de valores da faixa. É equivalente a assumir uma distribuição uniforme por faixa;
  - Para dados discretos, em que a ordem pode não ser importante (ou não existe), é dada uma função de mapeamento que indica um valor de referência para o valor consultado. Usualmente essa função indica o próprio valor consultado.

Espaço de Histogramas

Aproximação da frequência

Aproximação da frequência

- A dimensão "Aproximação da frequência" indica como a frequência deve ser aproximada para cada faixa.
- A estratégia dominante é a **Uniform distribution assumption**, na qual todos os valores correspondentes àquela faixa são assumidos terem a mesma frequência, e calculada como a média das frequências reais dos valores representados naquela faixa.

Espaço de Histogramas

Algoritmo de construção

Algoritmo de construção

- Dada uma regra de partição, é necessário haver um algoritmo que execute a construção de um histograma que atenda aquela regra.
- Note que pode haver diferentes algoritmos de construção de um mesmo tipo de histograma, cada um com desempenho melhor ou pior dependendo do tipo de dados sobre o qual será feita a análise de distribuição e de como os dados estão armazenados.

Espaço de Histogramas

Garantia de limite do erro

Garantia de limite do erro

- Todo histograma deve permitir estimar o limite máximo de erro que o uso do histograma causa.
- Esses limites dependem dos dados mantidos no histograma, e existem potencialmente mais de uma maneira de estimar o erro, e portanto cada histograma deve ter indicado como ele pode ser usado para o fim a que se destina.

Considerações sobre a implementação de Histogramas

- Como foi visto, por ser uma das estruturas para armazenar as "Estatísticas" de uma base de dados, os histogramas devem atender às restrições de:
  - Usar pouca memória;
  - Poder ser calculados rapidamente;
  - Apresentarem estimativas precisas obteníveis rapidamente.

Considerações sobre a implementação de Histogramas

Espaço em memória

- O espaço necessário para armazenagem de um histograma depende
  - Do número de faixas escolhidas
    ☞Tipicamente em torno de 20;
    ☞Usualmente ≤ 256;
  - Da quantidade de parâmetros armazenados;
    ☞Podem ser armazenados um valor fixo por faixa, ou parâmetros para estimar uma distribuição (usualmente linear) dentro da faixa;
  - Da estrutura de representação de valores em cada faixa. É comum:
    - Usar o parâmetro fonte definindo o Início e o Final de cada faixa se o valor for numérico;
    - Ou usar estruturas de indexação.
      ☞Tipicamente Arquivos Invertidos.

Considerações sobre a implementação de Histogramas

Construção

- A Construção de um Histograma pode ser:
  - **Estática** – o histograma é construído de uma vez sobre os dados disponíveis;
    Se os dados forem atualizados, o histograma deve ser reconstruído inteiro.
  - **Dinâmica** – o histograma vai sendo atualizado conforme os dados vão sendo atualizados.

- Os dados usados na construção podem ser:
  - **Domínio Ativo** – todos os dados armazenados são usados;
  - **Amostragem** – é feita uma amostragem sobre os dados. Essa amostragem pode ser
    ☞ Específica – executa-se uma operação de leitura específica para coletar os dados necessários;
    ☞ Baseada em *buffers* – aproveitam-se *buffers* em que resultados de operações normais de consulta são utilizados
    Isso pode levar a um histograma tendencioso, o que pode ser ruim ou ser intencional.

Considerações sobre a implementação de Histogramas

Geração de Estimativas

- A geração de estimativas deve:
  - Ser rápida de calcular;
  - Ser precisa;
  - Permitir estimar o erro máximo admissível.

Considerações sobre a implementação de Histogramas

Implementação de Histogramas em SGBDRs

- Para atender às três restrições, nem todos os tipos de histogramas são usados em SGBDRs "genéricos" atualmente;
- A maioria (todos os mais comuns) dos SGBDRs implementam o seguinte algoritmo:
- Dado um conjunto de atributos A de uma relação e um número B de faixas:
  1. Ordena a relação em A;
  2. Calcula $N = \|Dom^*(A)\|$;
  3. Se $N \leq B$ cria Equi-faixa;
  4. senão, se foi solicitada a reserva de $R < B$ faixas, cria Equi-Frequência com extremos detalhados;
  5. senão cria Equi-Frequência.

Implementação de Histogramas em SGBDRs

Exemplo

- Dada a relação de Professores, seja $A = \{\texttt{Area}\}$:
  1. Ordena a relação em $A = \{\texttt{Area}\}$:

```
Professor = {Nome, Idade, Area} =
          {<Lucio,   34,  BD>,
           <Laura,   45,  IA>,
           <Lucia,   54,  CG>,
           <Lucas,   43,  RE>,
           <Luana,   38,  IA>,
           <Livia,   61,  MM>,
           <Lineu,   42,  IA>,
           <Lacir,   45,  BD>,
           <Luigi,   33,  BD>,
           <Lidia,   28,  ES>,
           <Ligia,   51,  RE>,
           <Licio,   43,  IA>,
           <Leila,   56,  IA>,
           <Luzia,   45,  CG>}
```

```
Professor = {Nome, Idade, Area} =
                    {<Lucio,   34,  BD>,
                     <Lacir,   45,  BD>,
                     <Luigi,   33,  BD>,
                     <Lucia,   54,  CG>,
                     <Luzia,   45,  CG>,
                     <Lidia,   28,  ES>,
                     <Laura,   45,  IA>,
                     <Luana,   38,  IA>,
                     <Lineu,   42,  IA>,
                     <Licio,   43,  IA>,
                     <Leila,   56,  IA>,
                     <Livia,   61,  MM>,
                     <Lucas,   43,  RE>,
                     <Ligia,   51,  RE>}
```

Implementação de Histogramas em SGBDRs

Exemplo

2. Calcula $N = |Dom^*(\texttt{Area})| = 5$;

- Vamos assumir que $B >= 5$. 3. Como $N \leq B \Rightarrow$ **Equi-faixa**;

```
Professor = {Nome, Idade, Area} =
          {<Lucio,   34,  BD>,
           <Lacir,   45,  BD>,
           <Luigi,   33,  BD>,
           <Lucia,   54,  CG>,
           <Luzia,   45,  CG>,
           <Lidia,   28,  ES>,
           <Laura,   45,  IA>,
           <Luana,   38,  IA>,
           <Lineu,   42,  IA>,
           <Licio,   43,  IA>,
           <Leila,   56,  IA>,
           <Livia,   61,  MM>,
           <Lucas,   43,  RE>,
           <Ligia,   51,  RE>}
```

Area	Frequência
BD	3
CG	2
ES	1
IA	5
MM	1
RE	2

- Esse histograma é mantido em uma estrutura de índice ou em vetor que pode ser pesquisado por busca binária.
- O passo de ordenação prepara/aproveita esse índice.

Implementação de Histogramas em SGBDRs

Exemplo

- Vamos assumir que $B = 3$ e não há reserva de faixa.
- 5 Como $N > B$ SEM reserva de faixa $\Rightarrow$ **Equi-frequência**;

```
Professor = {Nome, Idade, Area} =
          {<Lucio,   34,  BD>,
           <Lacir,   45,  BD>,
           <Luigi,   33,  BD>,
           <Lucia,   54,  CG>,
           <Luzia,   45,  CG>,
           <Lidia,   28,  ES>,
           <Laura,   45,  IA>,
           <Luana,   38,  IA>,
           <Lineu,   42,  IA>,
           <Licio,   43,  IA>,
           <Leila,   56,  IA>,
           <Livia,   61,  MM>,
           <Lucas,   43,  RE>,
           <Ligia,   51,  RE>}
```

- A situação ideal é que cada grupo de B/N tuplas seja representado em cada faixa...
- ... mas cada valor deve estar totalmente contido numa faixa.

Area	Frequência
$\leq$CG	5
$\leq$IA	6
$\leq$RE	3

Implementação de Histogramas em SGBDRs

Exemplo

- Vamos assumir que $B = 3$ e há reserva de faixa $R = 1$.
- ❹ Como $N > B$ COM reserva de faixa $\Rightarrow$ **Equi-frequência com extremos detalhados**;

```
Professor = {Nome, Idade, Area} =
          {<Lucio,   34,  BD>,
           <Lacir,   45,  BD>,
           <Luigi,   33,  BD>,
           <Lucia,   54,  CG>,
           <Luzia,   45,  CG>,
           <Lidia,   28,  ES>,
           <Laura,   45,  IA>,
           <Luana,   38,  IA>,
           <Lineu,   42,  IA>,
           <Licio,   43,  IA>,
           <Leila,   56,  IA>,
           <Livia,   61,  MM>,
           <Lucas,   43,  RE>,
           <Ligia,   51,  RE>}
```

- O valor com maior cardinalidade de tuplas é `IA`, com 5 tuplas.
- Portanto, restam 14-5=9 tuplas que devem ser divididas em 2 faixas.

Implementação de Histogramas em SGBDRs

Exemplo

- Vamos assumir que $B = 3$ e há reserva de faixa $R = 1$.
- 4 Como $N > B$ COM reserva de faixa $\Rightarrow$ **Equi-frequência com extremos detalhados**;

```
Professor = {Nome, Idade, Area} =
          {<Lucio,   34,  BD>,
           <Lacir,   45,  BD>,
           <Luigi,   33,  BD>,
           <Lucia,   54,  CG>,
           <Luzia,   45,  CG>,
           <Lidia,   28,  ES>,
           <Laura,   45,  IA>,
           <Luana,   38,  IA>,
           <Lineu,   42,  IA>,
           <Licio,   43,  IA>,
           <Leila,   56,  IA>,
           <Livia,   61,  MM>,
           <Lucas,   43,  RE>,
           <Ligia,   51,  RE>}
```

Area ($\leq$)	Frequência
=IA	5
$\leq$CG	5
$\leq$RE	4

- Veja que a análise da tabela deve primeiro procurar por um valor igual ao da condição no detalhe,
- não havendo procura-se no restante da tabela.

Considerações sobre a implementação de Histogramas

Implementação de Histogramas em SGBDRs

- Complexidade do algoritmo:
  1. Ordena a relação em A; Complexidade $N \log(N)$
  2. Calcula $N = \|Dom^*(A)\|$; Complexidade 1
  3. Se $N \leq B$ cria Equi-faixa; Complexidade N
  4. senão, se foi dada reserva de faixas, cria Equi-Frequência com extremos detalhados; Complexidade $B + R$
  5. senão cria Equi-Frequência. Complexidade B
- Como N é o numero de valores distintos nos atributos usados no histograma (usualmente pequeno), $R < B$ e B e R somente são usados se $B < N$, a criação desse tipo de histograma é rápida.

Considerações sobre a implementação de Histogramas

Implementação de Histogramas em SGBDRs

- Esse algoritmo não implementa corretamente nenhum dos tipos teóricos a menos dos histogramas equi-faixa,
- mas é muito rápido, ocupa pouca memória e permite obter rapidamente as estimativas para predicados **por identidade**, **por faixa** e **por lista**.
- Além disso, ele tem desempenho muito semelhante aos histogramas equi-frequência ou equi-frequência com extremos detalhados, a um custo muito menor do que um histograma que siga a definição teórica à risca.
- Note também que no caso de extremos detalhados, ele é usado para detalhar apenas o extremo da alta frequência;
- Ele poderia ser usado para detalhar a baixa frequência da mesma maneira, mas o aumento de custo que isso sempre traz para a geração de estimativas não compensa o ganho de precisão geralmente muito pequeno que isso traz.

Uso de Histogramas em SGBDR

- O comando básico para geração de histogramas em qualquer SGBDR é o `ANALYZE STATISTICS`.
- No entanto, esse comando não existe na especificação ISO, portanto cada fabricante implementa esses recursos de maneira própria.
- Existe um conjunto básico de conceitos compartilhado pela maioria das implementações, mas aspectos mais específicos são próprios de cada fabricante/SGBDR/versão.

Como Gerar Histogramas em *ORACLE*

- Existem duas maneiras de solicitar a geração de Histogramas em *Oracle*:
  - O comando `ANALYZE STATISTICS`
  - Usar o pacote DBMS_STATS. Este é o método recomendado a partir do *Oracle* 10g:

☞ Os dois métodos são incompatíveis entre si!

- A técnica de usar pacotes como o DBMS_STATS permite que terceiros possam implementar (e vender) soluções próprias

☞ no caso do *Oracle*, os pacotes `SAS` são um bom exemplo.

Como Gerar Histogramas em *ORACLE*

Comando `ANALYZE STATISTICS`

ANALYZE STATISTICS – Oracle

```
ANALYZE TABLE <tabela>
    {COMPUTE [SYSTEM] STATISTICS |
     DELETE STATISTICS |
     ESTIMATE [SYSTEM] STATISTICS
              [SAMPLE <numero> {ROWS | PERCENT}]}
        [FOR TABLE | ALL INDEXES |
            ALL [INDEXED] COLUMNS [SIZE <faixas>]
            COLUMNS (<atributo> [SIZE <faixas>], )]
        [INTO <tabela>]
```

- O valor `<faixas>` ≥ 1 corresponde ao número de faixas do histograma. Se omitido, ele é assumido ser 75.
- Não existe um comando para apagar histogramas: uma vez criados, eles existem até que as estatísticas de uma tabela sejam apagadas e recriadas sem a indicação do histogramas.

Como Gerar Histogramas em *ORACLE*

- Os histogramas são criados sobre Indices ou sobre Atributos:

☞ Eles são preferencialmente criados sobre **Índices**, e nesse caso correspondem a qualquer número de atributos que compõem o índice;

☞ Histogramas criados sobre **Atributos** são criados sobre um único atributo;

- Independente de ser sobre índices ou sobre atributos, sempre são usados no máximo os 32 primeiros *bytes* do valor de cada tupla (excessão feita para o pacote DBMS_STATS).

Como Gerar Histogramas em *ORACLE*

- Existem dois tipos de histogramas usados em *Oracle*:
  - **FREQUENCY** – Histogramas Equi-faixa.
    Usado quando o número de faixas indicado é maior ou igual à cardinalidade do domínio ativo do atributo.
    ☞`DBA_TAB_COLUMNS.HISTOGRAM='FREQUENCY'`
  - **HEIGHT BALANCED** – Histogramas Equi-altura.
    Usado quando o número de faixas indicado é menor do que a cardinalidade do domínio ativo do atributo.
    ☞`DBA_TAB_COLUMNS.HISTOGRAM='HEIGHT BALANCED'`

  - Quando não existe um histograma
    para uma coluna, `DBA_TAB_COLUMNS.HISTOGRAM='NONE'` ou
    para um índice `DBA_INDEXES.HISTOGRAM='NONE'`.

Como obter dados dos Histogramas em *ORACLE*

As estatisticas são armazenadas no *Oracle Data Dictionary*.

- Estatísticas de Tabelas
  ☞ Usa a Visão `DBA_TABLES`.

  Inclui:
  - Número de tuplas (`CARD` ☞ `NUM_ROWS`),
  - Número de páginas usadas (`NPAGS` ☞ `NUM_BLOCKS`),
  - Precisão (`SAMPLE_SIZE`), e
  - Data da coleta (`LAST_ANALYZED`).
- Estatísticas de Atributos
  ☞ Usa a Visão `DBA_TAB_COLUMNS`.

  Inclui:
  - Número de valores distintos (`COLCARD` ☞ `NUM_DISTINCT`),
  - Menor valor (`LOW2KEY` ☞ `LOW_VALUE`),
  - Maior valor (`HIGH2KEY` ☞ `HIGH_VALUE`),
  - Tipo de Histograma (`HISTOGRAM`),
  - Número de Faixas (`NUM_BUCKETS`),
  - Precisão (`SAMPLE_SIZE`), e
  - Data de coleta (`LAST_ANALYZED`).
- Estatísticas de Índices
  ☞ Usa a Visão `DBA_INDEXES`.

  Inclui:
  - Altura da árvore (`H` ☞ `BLEVEL`),
  - Número de Folhas (`NLEAFS` ☞ `LEAF_BLOCKS`),
  - Número de chaves distintas (`KEYFULLCARD` ☞ `DISTINCT_KEYS`),
  - Índice de *clustering* (`CLUSTERRATIO` ☞ `CLUSTERING_FACTOR`),
  - Tipo de Histograma (`HISTOGRAM`),
  - Precisão (`SAMPLE_SIZE`), e
  - Data de coleta (`LAST_ANALYZED`).
- Estatísticas de Tabelas
  ☞ Usa a Visão `DBA_TAB_HISTOGRAMS`.

  Inclui um *array* com informações de faixa, contendo:
  - Número de elementos da faixa (`ENDPOINT_NUMBER`),
  - Valor final da faixa (`ENDPOINT_ACTUAL_VALUE`).

Como obter dados dos Histogramas em *ORACLE*

Exemplo

- A consulta seguinte mostra quais histogramas foram criados e quantas faixas cada um tem, para a tabela `Alunos`:

```
SELECT COLUMN_NAME, HISTOGRAM, NUM_BUCKETS
    FROM DBA_TAB_COLUMNS
    WHERE TABLE_NAME = 'ALUNOS' AND
        HISTOGRAM != 'NONE';
```

Como obter dados dos Histogramas em *ORACLE*

Exemplo

- A consulta seguinte mostra quais tabelas com mais de 10 mil tuplas foram amostradas com pelo menos 10%:

```
SELECT TABLE_NAME, NUM_ROWS, SAMPLE_SIZE,
       TO_CHAR(LAST_ANALYZED, 'dd.mm.yyyy hh24:mi:ss')
  FROM DBA_TABLES
  WHERE NUM_ROWS > 10000 AND
  SAMPLE_SIZE >= 0.1*NUM_ROWS;
```

Uso de Histogramas em *Oracle*

- Oracle permite usar somente histogramas do tipo **Equi-faixa** e **Equi-frequência**.
- Oracle permite usar o **Domínio Ativo** ou realizar **Amostragem** para criar os histogramas.
- Oracle permite adotar apenas a **Construção Estática** dos histogramas.

Como Gerar Histogramas em *Postgres*

- Em *Postgres*, os histogramas são gerados usando uma combinação de dois comandos:
  - O comando `ANALYZE STATISTICS`
  - O comando `ALTER TABLE <table> SET STATISTICS`
- Pacotes de terceiros também podem ser usados.

Como Gerar Histogramas em *Postgres*

Comando `ANALYZE STATISTICS`

ANALYZE STATISTICS – *Postgres*

```
ANALYZE [VERBOSE] [tabela [(atributo [, ...])]]
```

- *Postgres* escreve as estatísticas na meta-tabela `pg_statistic`, mas essa é uma tabela somente acessível ao DBA_Admin (ela pode conter dados sigilosos).
- A *meta-view* `pg_stats` mostra alguns valores de `pg_statistic` e fica disponível para leitura em geral.
- Não existe um comando para apagar histogramas: uma vez criados, eles existem até que as estatísticas de uma tabela sejam apagadas e recriadas sem a indicação do histogramas.

Como Gerar Histogramas em *Postgres*

Comando `ANALYZE STATISTICS`

ANALYZE STATISTICS – Postgres

```
ALTER TABLE <nome>
   ALTER [COLUMN] <atrib> SET STATISTICS [<num_faixas>];
ou
   ALTER [COLUMN] <atrib> SET (n_distinct=<num_distinct>);
```

- `SET STATISTICS <num_faixas>` indica o número de faixas que um histograma terá, usando a mesma lógica do *Oracle*:
  ☞Se o domínio ativo do histograma for maior do que `num_faixas`, ele será um histograma equi-frequência, senão será equi-faixa.
- Se não especificado, `num_faixas` vale 100.
- `SET (n_distinct=<num_distinct>)` indica o número de valores distintos não nulos que terão faixas exclusivas.
- um valor `num_distinct` entre $[-1, 0]$ indica um fator de multiplicação de quantidade de valores em cada faixa.
  Por exemplo: `num_distinct`=0.25 indica que cada faixa trata de 4 valores.

Como obter dados dos Histogramas em *Postgres*

- As estatísticas são armazenadas no *Postgres System Catalog*.
- Existe uma tabela específica para Estatísticas, que armazena essencialmente dados de Historamas: `pg_statistic` e sua visão "pública" `pg_stats`;
- e uma tabela específica para objetos que incluem atributos (colunas): `pg_class`.

Como obter dados dos Histogramas em *Postgres*

Tabela do catálogo: `pg_class`

Tabela do catálogo: `pg_class`

- Nome do Objeto: `relname`,
- Tipo do Objeto: `relkind` (`r`=tabela, `i`=index, `v`=view, ...),
- Número de tuplas (CARD ☞ `reltuples`),
- Número de páginas usadas (NPAGS ☞ `relpages`)

Como obter dados dos Histogramas em *Postgres*

Tabela do catálogo: `pg_stats`

Tabela do catálogo: `pg_stats`

- Nome da tabela: `tablename` (`references pg_class.relname`),
- Nome do atributo: `attname`,
- Fração de tuplas com valor nulo: `null_frac`,
- Número médio de bytes nas tuplas: `n_distinct` (p.ex. `VARCHAR`),
- Lista de valores mais comuns: `most_common_vals` (`NULL` se ˜ uniforme),
- Frequências dos valores mais comuns: `most_common_freqs`,
- Limites superiores de cada faixa: `histogram_bounds`,
- Contagem de tuplas em cada faixa: `elem_count_histogram`

Como obter dados dos Histogramas em *Postgres*

Exemplo

- A consulta seguinte mostra quais histogramas foram criados e quantas faixas cada um tem, para a tabela `Alunos`:

```
SELECT attname, n_distinct,
       array_length(most_common_vals,1)as NFaixas
  FROM pg_stats
  WHERE tablename = 'ALUNOS';
```

Como obter dados dos Histogramas em *Postgres*

Exemplo

- A consulta seguinte mostra quais tabelas com mais de 10 mil tuplas têm histogramas:

```
SELECT relname,attname, n_distinct,
       array_length(most_common_vals,1) AS NFaixas
  FROM (SELECT relname
        FROM pg_class
        WHERE relname NOT LIKE 'pg%' AND
              relname NOT LIKE 'sql%' AND
              relkind='r' AND
              reltuples>10000) AS taboid
       JOIN pg_stats ON relname=tablename;
```

Uso de Histogramas em *Postgres*

- *Postgres* permite usa naturalmente histogramas do tipo **Equi-faixa** e **Equi-frequência**, e permite criar uma combinação de histogramas **Seriados com extremos detalhados** com histogramas **Compressed**, dependendo da atuação do DBA.
- *Postgres* permite usar apenas o **Domínio Ativo** para criar os histogramas.
- *Postgres* permite adotar apenas a **Construção Estática** dos histogramas.

Ajuste Fino de SGBDR usando Histogramas

- Acrescentar Índices pode reduzir o desempenho de comandos de atualização;
- Histogramas são coletados apenas quando solicitado por comandos `ANALYZE`!
  ☞ Portanto causam *overhead* muito pequeno em comandos `INSERT`, `UPDATE` e `DELETE`.
- Mas a atualização de dados não atualiza os histogramas.

Ajuste Fino de SGBDR usando Histogramas

Quando usar histogramas ?

- É importante usar Histogramas em Atributos que tenham grande variação de frequência, e que:
  - Sejam indexados e participam de junções com alta seletividade na junção;
  - São usados em condições na cláusula `WHERE`, sejam eles indexados ou não.

Ajuste Fino de SGBDR usando Histogramas

Quando NÃO usar histogramas ?

- Não se deve criar histogramas para atributos que:
  - Tenham uma distribuição naturalmente uniforme;
  - Atributos que não sejam usados em condições de consulta (nem σ nem $\bowtie$);
  - Atributos que são chave ou tenham taxa de repetição muito pequena.

Arquitetura de SGBD Relacionais
— Métodos de Acesso Físico —

Caetano Traina Jr.

Grupo de Bases de Dados e Imagens
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo - São Carlos
caetano@icmc.usp.br

3 de abril de 2013
São Carlos, SP - Brasil

FIM