FIAP

Sistemas de informação

enterprise analytics and data warehousing

Profº Fabiano J. Cury Marques

https://br.linkedin.com/in/fabianocury

FIAP

enterprise analytics and data warehousing

Volumetria

...

Introdução

Introdução a volumetria de dados

Volumetria de dados

Pontos a serem considerados:

- ✘ Um DataWarehouse deve receber cargas de dados periódicas ao longo de seu ciclo de vida;
- ✘ Uma etapa importante do processo elaboração de um DataWarehouse consiste em conhecer o dinâmica de adição de dados em cada um dos Data Marts;
- ✘ Essa informação é importante para estimar os limites de capacidade de armazenamento independentemente do modelo de infra adotado (on premisse ou cloud computing);

Volumetria de dados

Pontos a serem considerados:

- ✘ Conhecer a priori o volume de dados inserido em um Data Mart pode não ser uma tarefa fácil;
- ✘ Uma abordagem possível pode considerar um questionamento, ao time de negócio, acerca da produção periódica de dados na empresa. Note que, neste caso, as informações podem vir carregadas de subjetividade;
- ✘ Outra abordagem possível pode passar por uma investigação na base de dados operacional existente, afim de conhecer a quantidade de registros inseridos em algumas tabelas de interesse;

Volumetria de dados

Calcular o tamanho do registro de uma tabela:

- ✘ Cada banco de dados pode adotar uma estratégia diferente para realizar o armazenamento do dado em cada uma de suas tabelas em disco rígido.
- ✘ O SQL Server 2000, por exemplo, limita o tamanho da linha em 8060 bytes e utiliza um mecanismo de paginação por blocos, de modo que cada MB contem 128 páginas.;
- ✘ Realizar um calculo de volumetria de dados considerando um banco de dados especifico exigiria conhece-lo em detalhes, o quê parece ser um preciosismo para uma abordagem numérica de caráter aproximativo;

Volumetria de dados

Calcular o tamanho do registro de uma tabela:

- ✘ Considera-se neste estudo o calculo baseado no tamanho dos registros.
- ✘ Outros elementos, tais como índices, foram desprezados das contas, mas podem ser facilmente considerados;
- ✘ O primeiro passo é conhecermos o tamanho de cada registro de cada tabela do datawarehouse;
- ✘ Conhecer o tamanho de cada registro passa por conhecer o tamanho de cada campo;

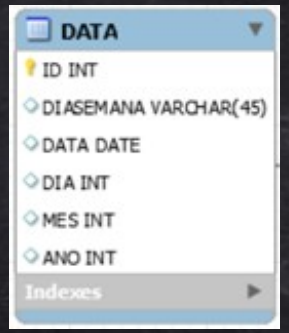
Volumetria de dados

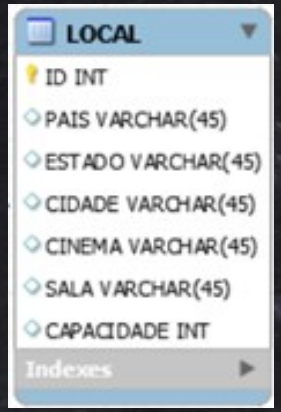
Datatype	Description	Column Length and Default
CHAR (*size*)	Fixed-length character data of length *size* bytes.	Fixed for every row in the table (with trailing blanks); maximum size is 2000 bytes per row, default size is 1 byte per row. Consider the character set (one-byte or multibyte) before setting *size*.
VARCHAR2 (*size*)	Variable-length character data.	Variable for each row, up to 4000 bytes per row. Consider the character set (one-byte or multibyte) before setting *size*. A maximum *size* must be specified.
NCHAR(*size*)	Fixed-length character data of length *size* characters or bytes, depending on the national character set.	Fixed for every row in the table (with trailing blanks). Column *size* is the number of characters for a fixed-width national character set or the number of bytes for a varying-width national character set. Maximum *size* is determined by the number of bytes required to store one character, with an upper limit of 2000 bytes per row. Default is 1 character or 1 byte, depending on the character set.
NVARCHAR2 (*size*)	Variable-length character data of length *size* characters or bytes, depending on national character set. A maximum *size* must be specified.	Variable for each row. Column *size* is the number of characters for a fixed-width national character set or the number of bytes for a varying-width national character set. Maximum *size* is determined by the number of bytes required to store one character, with an upper limit of 4000 bytes per row. Default is 1 character or 1 byte, depending on the character set.
CLOB	Single-byte character data.	Up to 2^32 - 1 bytes, or 4 gigabytes.
`NCLOB`	Single-byte or fixed-length multibyte national character set (NCHAR) data.	Up to 2^32 - 1 bytes, or 4 gigabytes.
`LONG`	Variable-length character data.	Variable for each row in the table, up to 2^31 - 1 bytes, or 2 gigabytes, per row. Provided for backward compatibility.
`NUMBER (p, s)`	Variable-length numeric data. Maximum precision *p* and/or scale *s* is 38.	Variable for each row. The maximum space required for a given column is 21 bytes per row.
`DATE`	Fixed-length date and time data, ranging from Jan. 1, 4712 B.C.E. to Dec. 31, 4712 C.E.	Fixed at 7 bytes for each row in the table. Default format is a string (such as DD-MON-YY) specified by NLS_DATE_FORMAT parameter.
`BLOB`	Unstructured binary data.	Up to 2^32 - 1 bytes, or 4 gigabytes.
`BFILE`	Binary data stored in an external file.	Up to 2^32 - 1 bytes, or 4 gigabytes.
`RAW (size)`	Variable-length raw binary data.	Variable for each row in the table, up to 2000 bytes per row. A maximum *size* must be specified. Provided for backward compatibility.
`LONG RAW`	Variable-length raw binary data.	Variable for each row in the table, up to 2^31 - 1 bytes, or 2 gigabytes, per row. Provided for backward compatibility.
`ROWID`	Binary data representing row addresses.	Fixed at 10 bytes (extended ROWID) or 6 bytes (restricted ROWID) for each row in the table.
`MLSLABEL`	Trusted Oracle datatype.	See the *Trusted Oracle* documentation.

Oracle data types:

Volumetria de dados

Calcular o tamanho do registro de uma tabela:





- Considerando os campos:
  - ID = 4 bytes, DIASEMANA= 45 bytes, DATA = 7 bytes, DIA = 4bytes, MÊS = 4 bytes e ANO = 4 bytes, temos um total de 68 bytes por registro

- Considerando os campos:
  - ID = 4 bytes, PAIS = 45 bytes, ESTADO = 45 bytes, CIDADE = 45 bytes, CINEMA = 45 bytes, SALA = 45 bytes e CAPACIDADE = 4 bytes, temos um total de 233 bytes por registro


2.
Carga Inical
Carga de dados Inicial

Volumetria de dados

Carga de dados inicial:

- É comum que um projeto de Data Warehouse receba uma expressiva carga de dados inicial;
- A carga de dados inicial pode, inclusive, ser dividida em algumas fases, permitindo realizar testes no modelo recém desenhado;
- É comum que esse processo seja feito de forma manual, com auxilio do time de TI, para depois ser automatizado utilizando ferramenta apropriada;

Volumetria de dados

Podemos calcular o volume de uma carga de dados inicial da seguinte forma:

$$C_0 = \sum_{j=1}^{m} T_j * Q0_j$$

Onde,

– índice da tabela

– quantidade total de tabelas no data warehouse;

- Tamanho (em bytes) do registro da tabela j;

- Quantidade inicial aproximada de registros inseridos na tabela j;

- Carga inicial em bytes (período 0);

Volumetria de dados

Considerando o tamanho do registro de cada tabela abaixo, calcule a carga de dados inicial:

Nome		
Tabela1	123	200
Tabela2	132	300
Tabela3	45	150
Tabela4	88	1000
Tabela5	99	280
Tabela6	125	3650
Tabela7	110	50000

Nome	
Tabela1	24600
Tabela2	39600
Tabela3	6750
Tabela4	88000
Tabela5	27720
Tabela6	456250
Tabela7	5500000
	6142920

- ✘ Note que temos dois vetore: T e Q0;

Volumetria de dados

Note que se fossemos escrever um programa para fazer essas contas poderíamos escreve-lo da forma abaixo :

1 – Dados :

Nome		
Tabela1	123	200
Tabela2	132	300
Tabela3	45	150
Tabela4	88	1000
Tabela5	99	280
Tabela6	125	3650
Tabela7	110	50000

2 – Código em linguagem C :

```c
#include <stdio.h>
int main()
{
  int m = 7;
  int  T[]={123, 132, 45, 88, 99, 125, 110};
  int Q0[]={200,300,150,1000,280,3650,50000};

  int C = 0;
  int soma = 0;
  for(int j=0; j< m; j++)
  {
    soma += T[j]*Q0[j];
    printf("T%d: %8d \n", j+1, T[j]*Q0[j]);
  }
  C = soma;

  printf("C0: %8d \n", C);
  return 0;
}
```

$$C_0 = \sum_{j=1}^{m} T_j * Q0_j$$

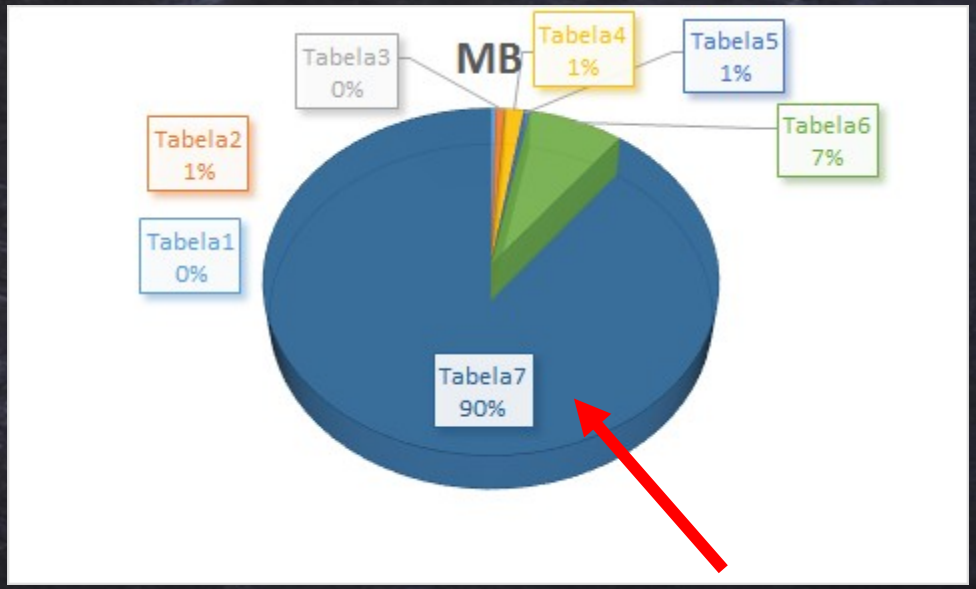
3 - Resultado:

```
> ./main
T1:    24600
T2:    39600
T3:     6750
T4:    88000
T5:    27720
T6:   456250
T7:  5500000
C0:  6142920
>
```

Volumetria de dados

Considerando o tamanho do registro de cada tabela abaixo, calcule a carga de dados inicial:

Nome		MB
Tabela1	24600	0.02346
Tabela2	39600	0.037766
Tabela3	6750	0.006437
Tabela4	88000	0.083923
Tabela5	27720	0.026436
Tabela6	456250	0.435114
Tabela7	5500000	5.245209
	6142920	5.858345



✘ Geralmente a fato!

✘ Carga inicial = 6142920 bytes, ~5.8 MB!

Exercício 1 – Volumetria

- Considerando os dados para a carga inicial em um data warehouse calcule o volume inserido em cada tabela e o valor total da carga:

Nome		
Tabela1	100	500
Tabela2	200	2000
Tabela3	210	3650
Tabela4	350	45
Tabela5	400	30
Tabela6	100	200
Tabela7	125	30000

Resposta:

```
> ./main
T1:     50000
T2:    400000
T3:    766500
T4:     15750
T5:     12000
T6:     20000
T7:   3750000
C0:   5014250
>
```


3.
Carga Periódica
Carga de dados periódica

Volumetria de dados

Podemos calcular o volume de uma carga de dados em um período da seguinte forma (para i > 0) :

$$C_i = \sum_{j=1}^{m} T_j * Q_j$$

Onde,

– índice associado ao período definido (dia, mês, ano,..)

– índice da tabela

– quantidade total de tabelas no data warehouse;

- Tamanho (em bytes) do registro da tabela j;

- Quantidade média aproximada de registros inseridos na tabela j;

- Carga em bytes no período i;

Volumetria de dados

1. Qual é o tamanho do registro (), em bytes, para cada tabela ?

   R: T1 = 35, T2 = 42, T3 = 18

2. Qual é o intervalo() de cargas ()? (por exemplo)

   R: mensal (diária, semanal, ,..)

3. Qual é a quantidade média de registros () por tabela ? (por exemplo)

   R: T1 = 200, T2 = 300, Tn = 1200

4. Qual a quantidade de períodos () analisados?

   R: 10 meses

5. Qual é o volume total () armazenado em cada intervalo ()?

   R: 0, (35*200 + 42*300 + 18*1200)*1, (35*200 + 42*300 + 18*1200)*2, (35*200 + 42*300 + 18*1200)*3, ... , (35*200 + 42*300 + 18*1200)*10

Volumetria de dados

O volume total de dados inseridos no Data warehouse no -ézimo período pode ser estimado:

$$VT_n = \sum_{i=1}^{n} C_i = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} T_j * Q_j \right)$$

Onde,

– índice associado ao período definido (dia, mês, ano,..)

– índice da tabela

– quantidade total de tabelas no data warehouse;

- Tamanho (em bytes) do registro da tabela j;

- Quantidade média aproximada de registros inseridos na tabela j;

- Carga em bytes no período i;

- Volume total, em bytes, no período i;

Exercício 2 – Volumetria

- Considerando o tamanho do registro de cada tabela abaixo, calcule a carga de dados periódica ao longo de n=10 períodos (mensais):

Nome		
Tabela1	100	0
Tabela2	200	20
Tabela3	210	30
Tabela4	350	5
Tabela5	400	1
Tabela6	100	0
Tabela7	125	300

Volumetria de dados

Considerando o tamanho do registro de cada tabela abaixo, calcule a carga de dados periódica ao longo de n=10 períodos (mensais):

Nome		
Tabela1	100	0
Tabela2	200	20
Tabela3	210	30
Tabela4	350	5
Tabela5	400	1
Tabela6	100	0
Tabela7	125	300

2 – Código em linguagem C :

```c
#include <stdio.h>
int main()
{
  int m = 7, n = 10;
  int  T[]={100,200,210,350,400,100,125};
  int  Q[]={0, 20, 30, 5, 1, 0, 300};
  int  C[]={0,0,0,0,0,0,0,0,0,0};

  int soma, VT = 0;
  for(int i=0; i< n; i++)
  {
    printf("  Periodo %d:\n", i+1);
    for(int j=0; j< m; j++)
    {
      C[i] += T[j]*Q[j];
      printf("  T%d: %8d \n", j+1, T[j]*Q[j]);
    }
    VT += C[i];
    printf("VT%2d: %8d \n\n", i+1, VT);
  }
  return 0;
}
```

$$VT_n = \sum_{i=1}^{n} C_i = \sum_{i=1}^{n} \sum_{j=1}^{m} T_j * Q_j$$

Exercício 2 – Volumetria

- Resposta:

```
./main
  Periodo 1:          Periodo 3:          Periodo 5:          Periodo 7:          Periodo 9:
  T1:         0       T1:         0       T1:         0       T1:         0       T1:         0
  T2:      4000       T2:      4000       T2:      4000       T2:      4000       T2:      4000
  T3:      6300       T3:      6300       T3:      6300       T3:      6300       T3:      6300
  T4:      1750       T4:      1750       T4:      1750       T4:      1750       T4:      1750
  T5:       400       T5:       400       T5:       400       T5:       400       T5:       400
  T6:         0       T6:         0       T6:         0       T6:         0       T6:         0
  T7:     37500       T7:     37500       T7:     37500       T7:     37500       T7:     37500
VT 1:     49950     VT 3:    149850     VT 5:    249750     VT 7:    349650     VT 9:    449550

  Periodo 2:          Periodo 4:          Periodo 6:          Periodo 8:          Periodo 10:
  T1:         0       T1:         0       T1:         0       T1:         0       T1:         0
  T2:      4000       T2:      4000       T2:      4000       T2:      4000       T2:      4000
  T3:      6300       T3:      6300       T3:      6300       T3:      6300       T3:      6300
  T4:      1750       T4:      1750       T4:      1750       T4:      1750       T4:      1750
  T5:       400       T5:       400       T5:       400       T5:       400       T5:       400
  T6:         0       T6:         0       T6:         0       T6:         0       T6:         0
  T7:     37500       T7:     37500       T7:     37500       T7:     37500       T7:     37500
VT 2:     99900     VT 4:    199800     VT 6:    299700     VT 8:    399600     VT10:    499500
```

Volumetria de dados

$$VT_n = \sum_{i=1}^{n} C_i = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} T_j * Q_j \right)$$

- Note que no modelo apresentado trabalhamos com tamanhos médios aproximados de registros fixos em cada período
- Uma abordagem mais realista deveria considerar valores médios aproximados variáveis em cada período i,

Volumetria de dados

$$VT_n = \sum_{i=1}^{n} C_i = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} T_j * Q_{ij} \right)$$

- Nesta abordagem os valores médios aproximados podem variar em cada período i

- A expressão acima permite utilizar diversos modelos de previsão para mensurar variações (crescimento/decrescimento) do volume de dados inseridos nas tabelas do data warehouse, aumenta assim a qualidade da estimativa.

Exemplo – Volumetria

Nome					
Tabela1	10	510	520	530	540
Tabela2	15	71	72	73	74
Tabela3	20	810	820	830	840

- Considerando o tamanho do registro de cada tabela abaixo, calcule a carga de dados periódica ao longo de n=4 períodos (mensais):

Para a carga no período 1, temos:

$C1 = T1 * Q11 + T2 * Q12 + T3 * Q13 =$

$C1 = 10 * 510 + 15 * 71 + 20 * 810 = 22365$

Para a carga no período 2, temos:

$C2 = T1 * Q21 + T2 * Q22 + T3 * Q23 =$

$C2 = 10 * 520 + 15 * 72 + 20 * 820 = 22680$

Para a carga no período 3, temos:

$C3 = T1 * Q31 + T2 * Q32 + T3 * Q33 =$

$C3 = 10 * 530 + 15 * 73 + 20 * 830 = 22995$

Para a carga no período 4, temos:

$C4 = T1 * Q41 + T2 * Q42 + T3 * Q43 =$

$C4 = 10 * 540 + 15 * 74 + 20 * 840 = 23310$

Para a carga no período 1, temos:

$$VT_n = \sum_{i=1}^{n} C_i = C_1 + C_2 + C_3 + C_4 =$$

$$VT_n = \sum_{i=1}^{n} C_i = 22365 + 22680 + 22995 + 23310 = 91350$$

Lembrando que:

$VT1 = C1 = 22365$

$VT2 = C1 + C2 = 22365 + 22680 = 45045$

$VT3 = C1 + C2 + C3 = 22365 + 22680 + 22995 = 68040$

$VT4 = C1 + C2 + C3 + C4 = 22365 + 22680 + 22995 + 23310 = 91350$

Exercício 3 – Volumetria

- Considerando o tamanho do registro de cada tabela abaixo, calcule a carga de dados periódica ao longo de n=10 períodos (mensais):

Nome											
Tabela1	100	0	0	0	1	0	0	2	0	0	0
Tabela2	200	20	21	22	24	25	27	28	27	27	28
Tabela3	210	30	30	30	30	30	30	30	30	30	30
Tabela4	350	5	5	5	6	5	5	7	5	8	5
Tabela5	400	1	1	0	1	1	0	0	1	1	1
Tabela6	100	0	0	0	0	0	0	0	0	0	0
Tabela7	125	300	310	315	320	315	340	345	340	350	355

Volumetria de dados

2 – Código em linguagem C :

Considerando o tamanho do registro de cada tabela abaixo, calcule a carga de dados periódica ao longo de n=10 períodos (mensais):

Nome											
Tabela 1	100	0	0	0	1	0	0	2	0	0	0
Tabela 2	200	20	21	22	24	25	27	28	27	27	28
Tabela 3	210	30	30	30	30	30	30	30	30	30	30
Tabela 4	350	5	5	5	6	5	5	7	5	8	5
Tabela 5	400	1	1	0	1	1	0	0	1	1	1
Tabela 6	100	0	0	0	0	0	0	0	0	0	0
Tabela 7	125	300	310	315	320	315	340	345	340	350	355

```c
#include <stdio.h>
int main()
{
  int m = 7, n = 10;
  int  T[]={100,200,210,350,400,100,125};
  int  C[]={0, 0, 0, 0, 0, 0, 0, 0, 0, 0};
  int  Q[10][7]={  0,  20,  30,   5,   1,   0, 300,
                   0,  21,  30,   5,   1,   0, 310,
                   0,  22,  30,   5,   0,   0, 315,
                   1,  24,  30,   6,   1,   0, 320,
                   0,  25,  30,   5,   1,   0, 315,
                   0,  27,  30,   5,   0,   0, 340,
                   2,  28,  30,   7,   0,   0, 345,
                   0,  27,  30,   5,   1,   0, 340,
                   0,  27,  30,   8,   1,   0, 350,
                   0,  28,  30,   5,   1,   0, 355};
  int VT = 0;
  for(int i=0; i< n; i++)
  {
      printf("Periodo %d \n", i+1);
      for(int j=0; j< m; j++)
      {
        C[i] += T[j]*Q[i][j];
        printf(" T%d: %8d \n", j+1, T[j]*Q[i][j]);
      }
      VT += C[i];
      printf("VT%d: %8d \n\n", i+1, VT);
  }
  return 0;
}
```

$$VT_n = \sum_{i=1}^{n} C_i = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} T_j * Q_{ij} \right)$$

Exercício 3 – Volumetria

- ✘ Resposta:

```
> ./main
Periodo 1
 T1:        0
 T2:     4000
 T3:     6300
 T4:     1750
 T5:      400
 T6:        0
 T7:    37500
VT1:    49950

Periodo 2
 T1:        0
 T2:     4200
 T3:     6300
 T4:     1750
 T5:      400
 T6:        0
 T7:    38750
VT2:   101350

Periodo 3
 T1:        0
 T2:     4400
 T3:     6300
 T4:     1750
 T5:        0
 T6:        0
 T7:    39375
VT3:   153175

Periodo 4
 T1:      100
 T2:     4800
 T3:     6300
 T4:     2100
 T5:      400
 T6:        0
 T7:    40000
VT4:   206875

Periodo 5
 T1:        0
 T2:     5000
 T3:     6300
 T4:     1750
 T5:      400
 T6:        0
 T7:    39375
VT5:   259700

Periodo 6
 T1:        0
 T2:     5400
 T3:     6300
 T4:     1750
 T5:        0
 T6:        0
 T7:    42500
VT6:   315650

Periodo 7
 T1:      200
 T2:     5600
 T3:     6300
 T4:     2450
 T5:        0
 T6:        0
 T7:    43125
VT7:   373325

Periodo 8
 T1:        0
 T2:     5400
 T3:     6300
 T4:     1750
 T5:      400
 T6:        0
 T7:    42500
VT8:   429675

Periodo 9
 T1:        0
 T2:     5400
 T3:     6300
 T4:     2800
 T5:      400
 T6:        0
 T7:    43750
VT9:   488325

Periodo 10
 T1:        0
 T2:     5600
 T3:     6300
 T4:     1750
 T5:      400
 T6:        0
 T7:    44375
VT10:  546750
```

Volumetria de dados

O volume total de dados inseridos no Data warehouse no -ézimo, juntamente com a carga inicial, pode ser estimado pela expressão:

Carga de dados inicial

$$VT_n = C_0 + \sum_{i=1}^{n} C_i$$

Carga de dados periódica

$$VT_n = \sum_{j=1}^{n} \left(T_j * Q0_j \right) + \sum_{i=1}^{n} \left(\sum_{j=1}^{m} T_j * Q_j \right)$$

- ✘ Nº de registros fixos por tabela

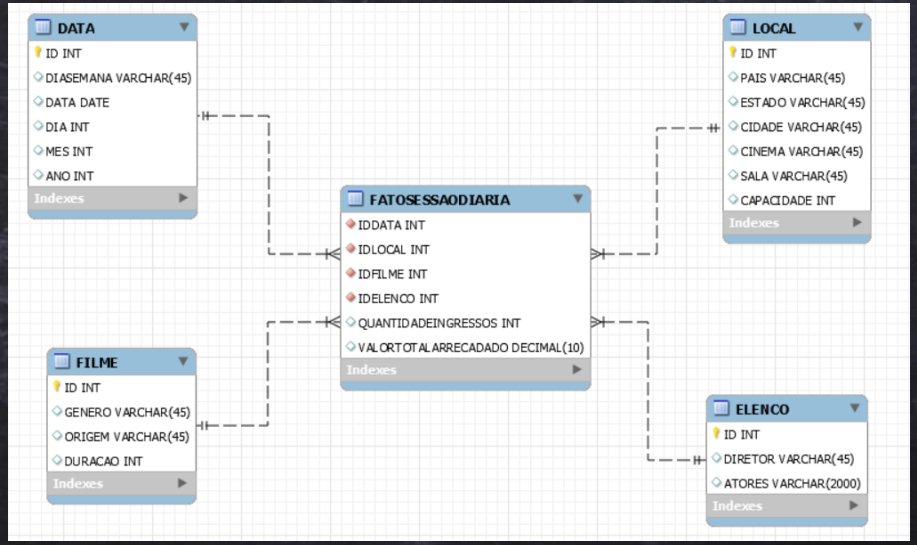
$$VT_n = \sum_{j=1}^{n} \left(T_j * Q0_j \right) + \sum_{i=1}^{n} \left(\sum_{j=1}^{m} T_j * Q_{ij} \right)$$

- ✘ Nº de registros variável por tabela

Exercício – Modelagem Dimensional

- Utilizando o Data Mart de distribuidora de filmes estudado anteriormente, apresente um estudo para estimar a volumetria de dados, considerando:

  - Utilizar nº fixo de registros;
  - Definir juntamente com o seu grupo os valores de e ;
  - Calcular a carga de dados inicial;
  - Calcular a carga de dados periódica (mensal), com n=24;



Referências

- KIMBALL, R., ROSS, M. The Data Warehouse Toolkit. 2ª ed., John Wiley Professional, 2002.
- MACHADO, F. N. R. Tecnologia e Projeto de Data Warehouse. 1ª ed., São Paulo: Ed. Érica, 2004.

Obrigado!

https://br.linkedin.com/in/wakimsaba