Data Mining
Dados

Prof. Dr. Joaquim Assunção

DEPARTAMENTO DE COMPUTAÇÃO APLICADA
CENTRO DE TECNOLOGIA
UFSM
2024

www.inf.ufsm.br/~joaquim

Fair user agreement

Este material foi criado para a disciplina de Mineração de Dados - Centro de Tecnologia da UFSM.

Você pode usar este material livremente*; porém, caso seja usado em outra instituição, **me envie um e-mail** avisando o nome da instituição e a disciplina.

*A maior parte deste material foi retirado do livro: “Joaquim V. C. Assunção. **Uma Breve Introdução à Mineração de Dados: Bases Para a Ciência de Dados, com Exemplos em R**. 192 páginas. Novatec. 2021. ISBN-10 : 6586057507.”

Prof. Dr. Joaquim Assunção.

joaquim@inf.ufsm.br

Dado - informação - Conhecimento

- Dado: unidade de informação mínima, geralmente um fato sobre algo.
- Informação: dado com semântica.
- Conhecimento: saber de informações e de seu contexto.

Dado - informação - Conhecimento

- Dado: unidade de informação mínima, geralmente um fato sobre algo.
- Informação: dado com semântica.
- Conhecimento: saber de informações e de seu contexto.
- Exemplo →

Dado - informação - Conhecimento

- Dado: unidade de informação mínima, geralmente um fato sobre algo.
- Informação: dado com semântica.
- Conhecimento: saber de informações e de seu contexto.

- Exemplo →

Dado	Verde 60 km/h 17 horas
Informação	Sinal verde, motorista pode seguir. Limite de velocidade da via é 60 km/h. Horário de maior fluxo na via é em torno das 17 horas.
Conhecimento	Pedestre deve esperar o sinal vermelho para os veículos. Se a sinaleira está piscando em amarelo, há problema. Atravessar no amarelo piscante, às 17h, pode ser perigoso.

Decisão com base em dados

- Dilema entre conhecimento empírico vs quantidade de dados.
- Dado é, geralmente, um fato.
- Decisões devem ser tomadas com bases em fatos.
- Dados são o material bruto para mineração de dados.

Tipos de dados

- Estruturados, Não-estruturados e Semiestruturados

Tipos de dados

- **Estruturados**, Não-estruturados e Semiestruturados
- Tipicamente tabulares (Banco de dados*, planilhas, etc.)
- A maioria dos sistemas de pequeno e médio porte geram dados tabulares.

* Relacionais e tabelas extraídas por SQL.

Tipos de dados

- **Estruturados**, Não-estruturados e Semiestruturados
- Tipicamente, dados em registro
  - Dados de transação ou cesta de mercado
  - Dados baseados em grafos
  - Matriz de dados dispersos ou matriz binária

Tipos de dados

- **Estruturados**, Não-estruturados e Semiestruturados

Exemplo de estrutura típica..

CPF	Nome	Idade	Profissão	Renda	Escolaridade	Filhos	Bom pagador
9990009 9900	Clotilde	67	Aposentada	2100	Médio	0	S
8884445 5500	Madruga	56	Desempregado	NA	Médio	1	N
1112223 3344	Ector	45	Ator	15000	Médio	0	S
8889990 0011	Girafales	48	Professor E.Básico	2500	Superior	0	S

Tipos

- Estruturados e Semiestruturados

Identificador
Útil apenas
para
indexação

CPF	Nome	Idade	Profissão	Renda	Escolaridade	Filhos	Bom pagador
9990009 9900	Clotilde	67	Aposentada	2100	Médio	0	S
8884445 5500	Madruga	56	Desempregado	NA	Médio	1	N
1112223 3344	Ector	45	Ator	15000	Médio	0	S
8889990 0011	Girafales	48	Professor E.Básico	2500	Superior	0	S

Tipos de dados

- **Estruturados**, Não-es truturados

Dados descritivos. Nome, via de regra, não é útil.

CPF	Nome	Idade	Profissão	Renda	Escolaridade	Filhos	Bom pagador
9990009 9900	Clotilde	67	Aposentada	2100	Médio	0	S
8884445 5500	Madruga	56	Desempregado	NA	Médio	1	N
1112223 3344	Ector	45	Ator	15000	Médio	0	S
8889990 0011	Girafales	48	Professor E.Básico	2500	Superior	0	S

Tipos de dados

- **Estruturados**, Não-estrut… …dos

Atributo **classe**, ou **rótulo**.

CPF	Nome	Idade	Profissão	Renda	Escolaridade	Filhos	Bom pagador
9990009 9900	Clotilde	67	Aposentada	2100	Médio	0	S
8884445 5500	Madruga	56	Desempregado	NA	Médio	1	N
1112223 3344	Ector	45	Ator	15000	Médio	0	S
8889990 0011	Girafales	48	Professor E.Básico	2500	Superior	0	S

Tipos de dados

- **Estruturados**, Não-est

Atributo **classe**, ou **rótulo** (neste caso, contínuo).

	cyl	disp	hp	drat	wt	qsec	am	gear	carb	mpg
Mazda RX4	6	160	110	3.9	2.62	16.46	1	4	4	21
Mazda RX4 Wag	6	160	110	3.9	2.875	17.02	1	4	4	21
Datsun 710	4	108	93	3.85	2.32	18.61	1	4	1	22.8
Hornet 4 Drive	6	258	110	3.08	3.215	19.44	0	3	1	21.4
Hornet Sportabout	8	360	175	3.15	3.44	17.02	0	3	2	18.7
Valiant	6	225	105	2.76	3.46	20.22	0	3	1	18.1
Duster 360	8	360	245	3.21	3.57	15.84	0	3	4	14.3
Merc 240D	4	146.7	62	3.69	3.19	20	0	4	2	

Tipos de dados

- **Estruturados**, Não-est… …ruturados

Dados descritivos. Nome, via de regra, não é útil.

	cyl	disp	hp	drat	wt	qsec	am	gear	carb	mpg
Mazda RX4	6	160	110	3.9	2.62	16.46	1	4	4	21
Mazda RX4 Wag	6	160	110	3.9	2.875	17.02	1	4	4	21
Datsun 710	4	108	93	3.85	2.32	18.61	1	4	1	22.8
Hornet 4 Drive	6	258	110	3.08	3.215	19.44	0	3	1	21.4
Hornet Sportabout	8	360	175	3.15	3.44	17.02	0	3	2	18.7
Valiant	6	225	105	2.76	3.46	20.22	0	3	1	18.1
Duster 360	8	360	245	3.21	3.57	15.84	0	3	4	14.3
Merc 240D	…	…	…	…	…	…	…	…	…	24.4

Tipos de dados

- Estruturados, **Não-estruturados** e Semiestruturados
- Textos, imagens, vídeos, etc.
- Há uma grande parte da literatura destinada a mineração de textos.
- Mineração de imagens e vídeos é um tópico crescente.
- Todos requerem pré-processamento.

Tipos de dados

- Estruturados, Não-estruturados e **Semiestruturados**
- Estrutura não tabular, mas há marcações que auxiliam na extração das informações.
- XML, HTML etc.

Tipos de Atributos - **Categóricos e Numéricos**

- Atributos **categóricos** podem ser:

1. **Binários**, em que só existem duas possíveis categorias.
2. **Nominais**, em que os valores são strings descritivas.
3. **Ordinais**, em que existe ordem entre os valores; seja colocação, ranking ou hierarquia.

Tipos de Atributos

- Atributos **numéricos** podem ser discretos ou contínuos.

1. **Discreto**, quando o conjunto é finito e definido, possivelmente binário. Exemplos: idade de uma pessoa em anos, valor de um refrigerante, e quantidade de carros na loja (colunas Grupo e Colocação na Tabela 2.3).
2. **Contínuo**, quando o conjunto é indefinido em precisão e possivelmente infinito. Exemplos: onda sonora em ambiente analógico. Temperatura exata em um determinado ponto (caso utópico, sem definir precisão decimal).

Tipos de Atributos – Exemplos.

Grupo	Nome	Colocação
1	Michael	1
1	Fernando	3
2	Rubens	5
2	Emerson	4
1	Sebastian	2

Atributo Ordinal

Atributo Nominal

Atributo Binário
(considerando apenas dois grupos)

OLTP e OLAP

- Termos ligados a *BI*
- OLAP (*OnLine Analytical Processing*)
- OLTP (*OnLine Transaction Processing*)

OLTP

- Sistemas que se encarregam de registrar todas as transações contidas em uma determinada operação organizacional.
- Sistemas OLTP, geralmente, tem a característica de ter um grande número de transações curtas on-line (INSERT, UPDATE, DELETE).

OLTP

- Ênfase em processamento de consultas com grande velocidade, mantendo a integridade dos dados em ambientes com múltiplos acessos.

✓ Velocidade e simplicidade

- *Small transactions*
- *Short response time*
- *Data maintenance operations*
- *Large user populations*
- *High concurrency*
- *Large data volumes*
- *High availability*

OLAP

- Sistemas menos usados, mas mais específicos para análise de dados.
- Foco em dados históricos, arquivados e/ou agregados.
- Caracterizado pelo volume relativamente baixo de transações.

OLAP

- As consultas geralmente são complexas e envolvem agregações.
- Aplicativos OLAP são amplamente utilizados em conjunto com técnicas de Data Mining.
- Esquemas são agregados de maneira multidimensional (geralmente esquema em estrela).

OLAP

- Algumas ferramentas permitem visualização ao estilo arrasta e solta.
- É possível navegar por *drill-down* e *roll-up*.

OLTP & OLAP

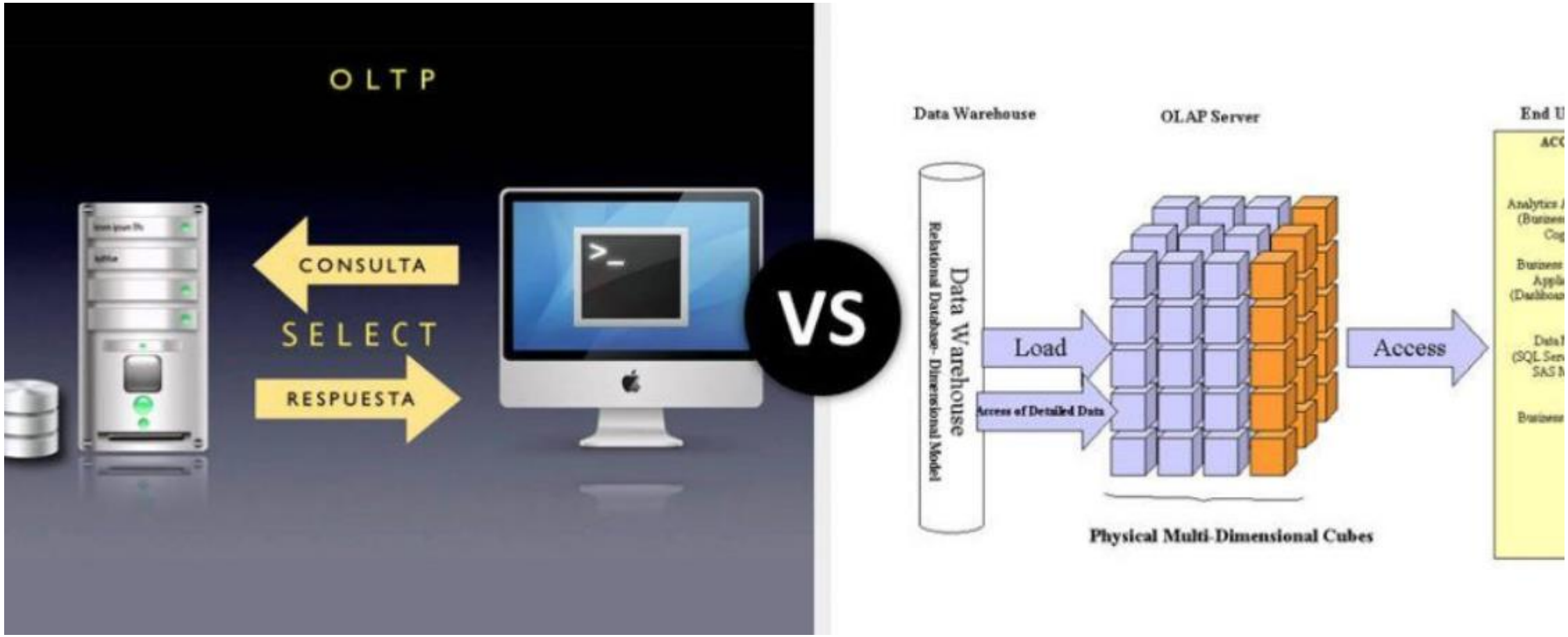

Imagem retirada de: https://www.differencebtw.com/difference-between-oltp-and-olap/

Extra readings

- Leia sobre ***Data werehousing***, ***Star model*** e ***Snowflake model***
- ***Links →*** http://www.oracle.com/webfolder/technetwork/tutorials/obe/db/10g/r2/owb/owb10gr2_gs/owb/lesson3/starandsnowflake.htm

  https://docs.oracle.com/database/121/DWHSG/concept.htm#DWHSG9288

Tidy dataset

- Imagine a tabela no seguinte formato:

Nome	Literatura	Matematica
Steve	8.5	9
Bill	9	9.5
Martin	10	7

- O mesmo conjunto pode ser mostrado de diferentes formas...

Tidy dataset

Nome	Literatura	Matematica
Steve	8.5	9
Bill	9	9.5
Martin	10	7

- O mesmo conjunto pode ser mostrado de diferentes formas...

Disciplina	Steve	Bill	Martin
Literatura	8.5	9	10
Matematica	9	9.5	7

Tidy dataset

- Para facilitar a leitura e manipulação destes dados foi criado o conceito de "*Tidy data*".
- …"Uma variável contém todos os valores que medem o mesmo atributo subjacente entre as unidades. Uma observação contém todos os valores medidos na mesma unidade entre os atributos."

Tidy dataset

- Para o exemplo anterior teríamos:

Nome	Disciplina	Nota
Steve	Literatura	8.5
Steve	Matematica	9
Bill	Literatura	9
Bill	Matematica	9.5
Martin	Literatura	10
Martin	Matematica	7

Tidy dataset

- Trocamos, compactação por clareza. Isto pode ser útil para fins de análise (após as primeiras tarefas de limpeza dos dados).

Hands On!

- Em R, crie a primeira tabela com os comandos:

```r
nomes <- c("Steve", "Bill", "Martin")
literatura <- c(8.5, 9, 10)
matematica <- c(9, 9.5, 7)

DF1 <- data.frame(cbind(nomes,literatura,matematica))
```

1. Crie as tabelas nos demais formatos.

Disciplina	Steve	Bill	Martin
Literatura	8.5	9	10
Matematica	9	9.5	7

Nome	Disciplina	Nota
Steve	Literatura	8.5
Steve	Matematica	9
Bill	Literatura	9
Bill	Matematica	9.5
Martin	Literatura	10
Martin	Matematica	7