

Ciência de Dados

Licenciatura Engenharia Informática
2º Semestre – 2021/2022

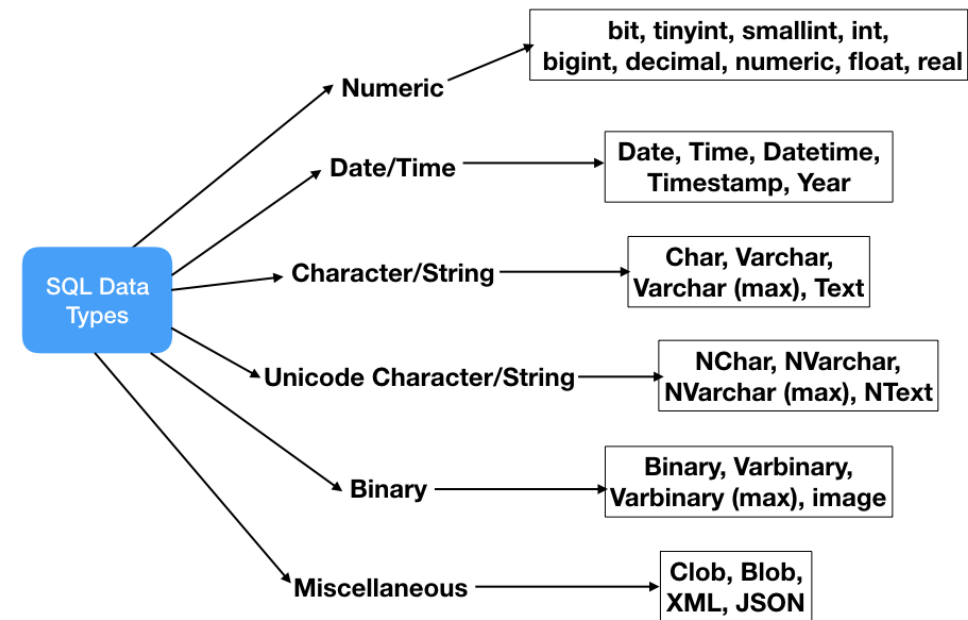
Ricardo Jesus Ferreira
ricardojesus.ferreira@my.istec.pt

Dados

- Tipos de Dados
- Estrutura de Dados
- Tipos Ficheiros Dados

Tipos de Dados

- O tipo de dados é um esquema detalhado da codificação que será realizada
- Esta codificação terá que estar de acordo com o DBMS
- A escolha do tipo de dados terá que garantir:
 - Abrangência dos valores a representar
 - Integridade
 - Permitir todas as manipulações
 - O menor espaço em memória



Date

- Utilizado para armazenar datas e tempo (horas, minutos, segundos)
- Existem diversas variações

DATE

YYYY-MM-DD = 2022-02-05

TIMESTAMP

YYYY-MM-DD HH24:MI:SS.FF
= 2022-02-05 05:24:05.12

Numérico

- Utilizado para armazenar números
- Múltiplos tipos dependendo da linguagem e DBMS

INT = 5

FLOAT/DOUBLE = 5,5

Alfanumérico

- Utilizado para armazenar mistura de números e letras
- Tipicamente associado à `STRING` ou `VARCHAR`
- Utilizado para armazenar IDs, como o produto ID, cliente ID, etc..

PRD0001

CUST666

Moeda

- Utilizado para armazenar dimensão monetária na DBMS
- Exemplos podem ser Euros, dólares, yen, libras, etc..

800€

85,23\$

Texto

- Utilizado para armazenar grandes dados de informação de texto
- Guarda uma string com um comprimento máximo de 65.535 bytes

Asda sda sdas dasda sdasdasdasd as dasda
sd asda sdasda s dasdasd asda asdasd
asdasdasdasd asdasda sda s dasda das sd
as dasd asda sdasdas dasdasda das sdasd

Asda sda sdas dasda sdasdasdasd as dasda
sd asda sdasda s dasdasd asda asdasd
asdasdasdasd asdasda sda s dasda das sd
as dasd asda sdasdas dasdasda das sdasd

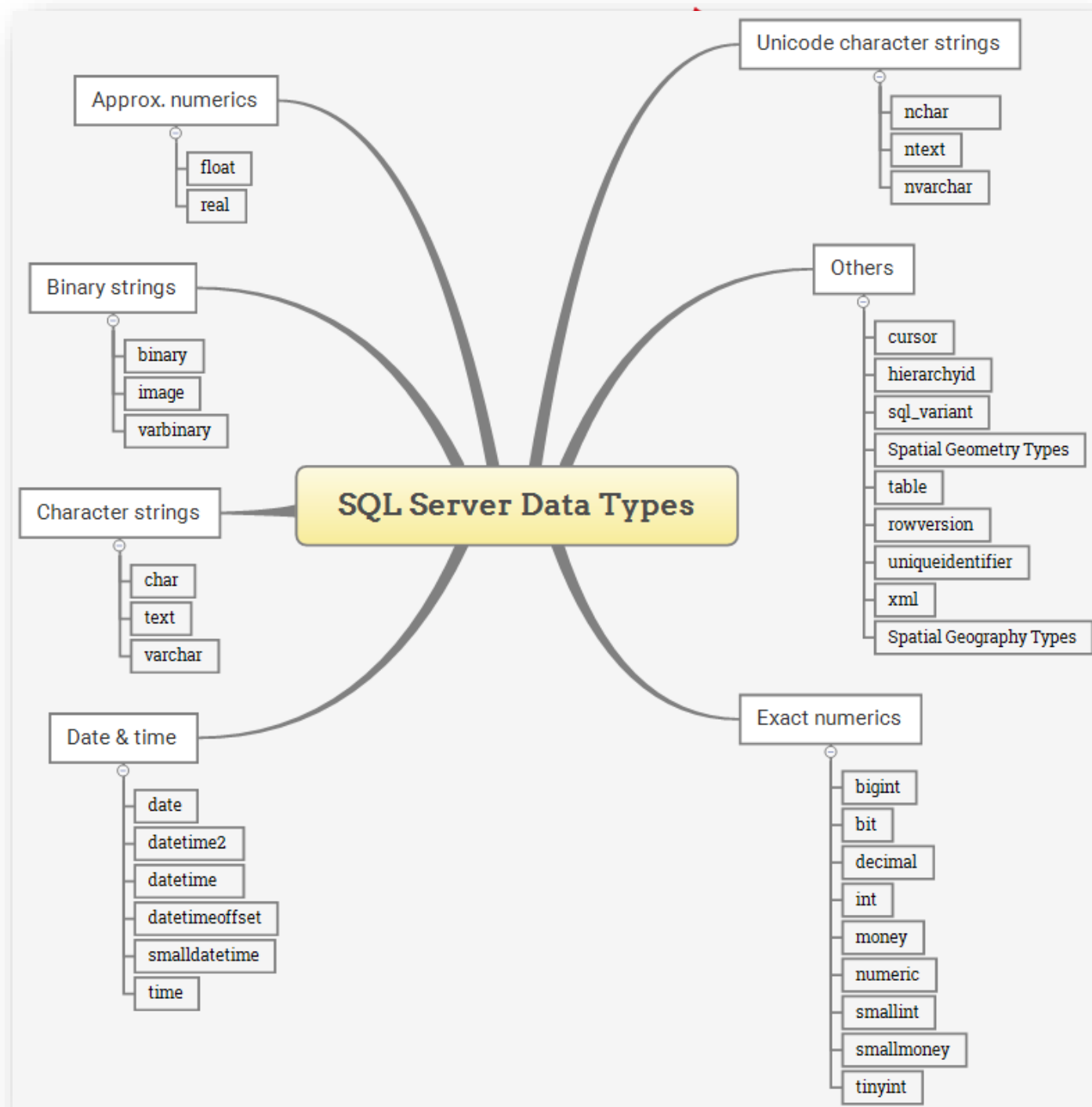
Asda sda sdas dasda sdasdasdasd as dasda
sd asda sdasda s dasdasd asda asdasd
asdasdasdasd asdasda sda s dasda das sd
as dasd asda sdasdas dasdasda das sdasd

Asda sda sdas dasda sdasdasdasd as dasda
sd asda sdasda s dasdasd asda asdasd
asdasdasdasd asdasda sda s dasda das sd
as dasd asda sdasdas dasdasda das sdas

Imagens, vídeos e audios

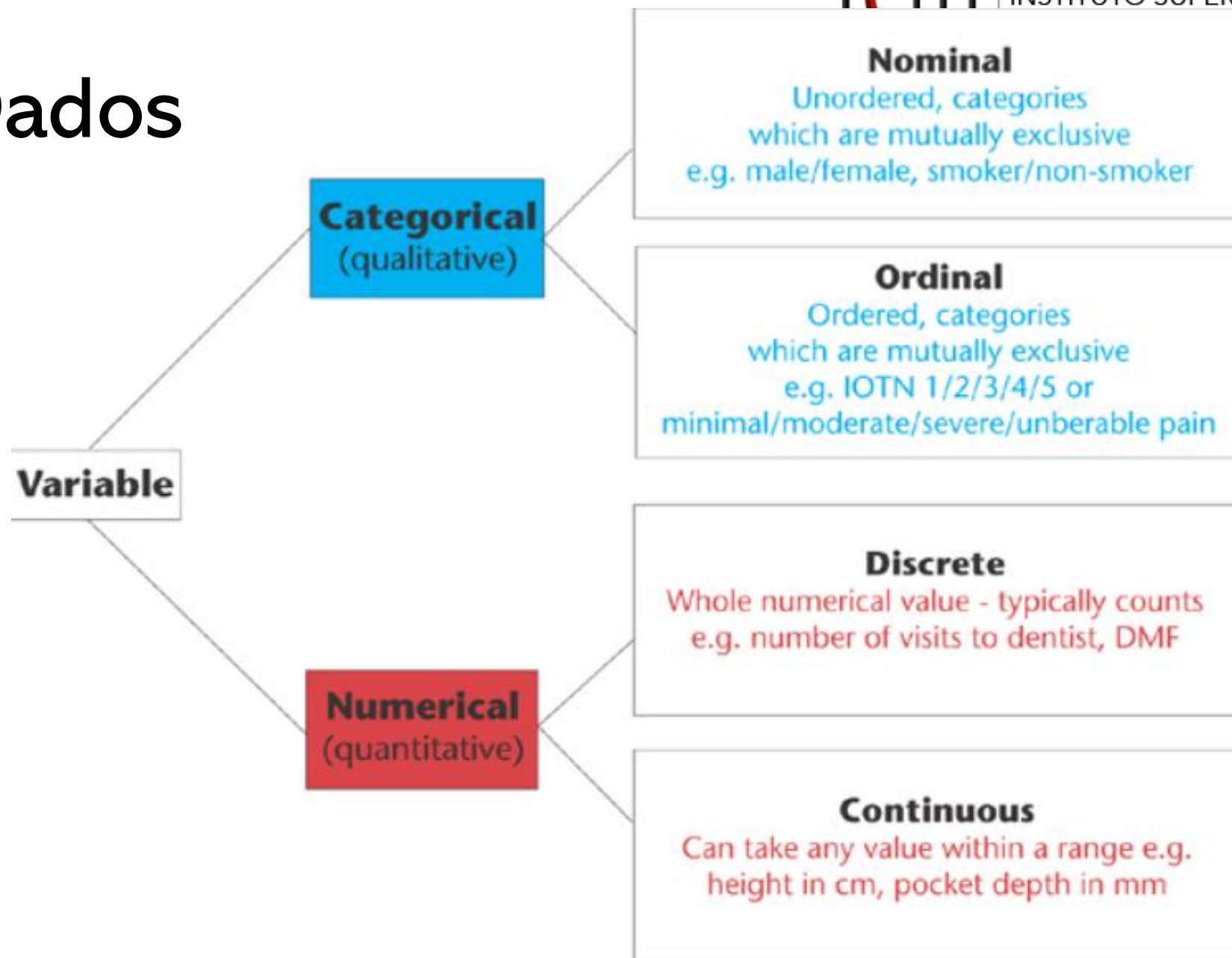
- Estes dados são tipicamente guardados como
 - BLOB (Binary Large Object)
- Os BLOBs são tipicamente utilizados para guardar imagens, vídeos, áudios ou executáveis
- Possui variações:
 - Binary
 - VarBinary(n)
 - VarBinary(max = 2G)

Tipos de dados SQL



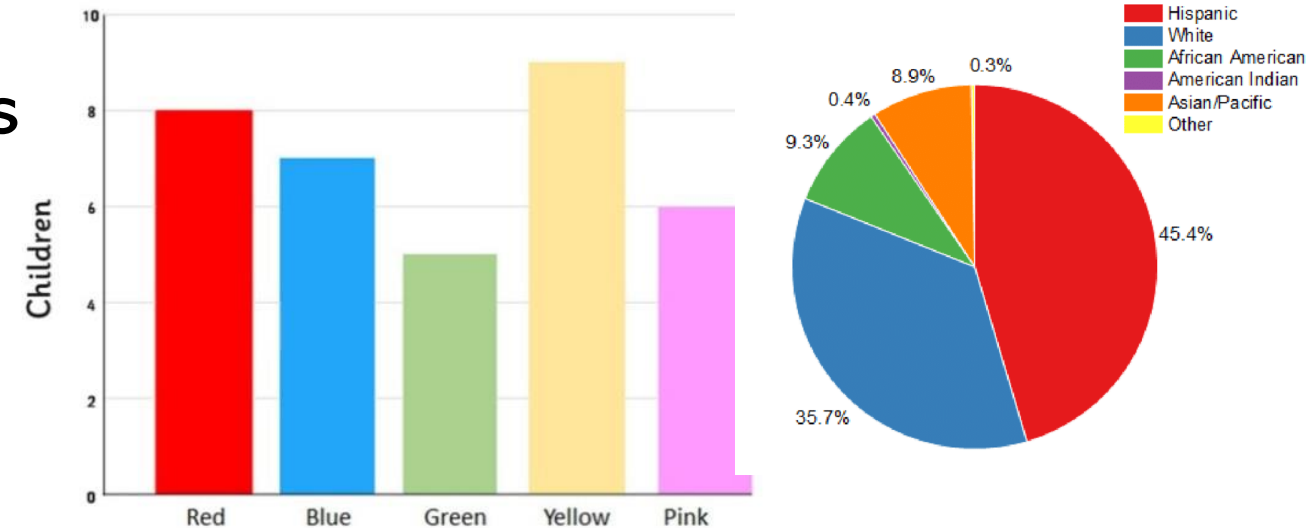
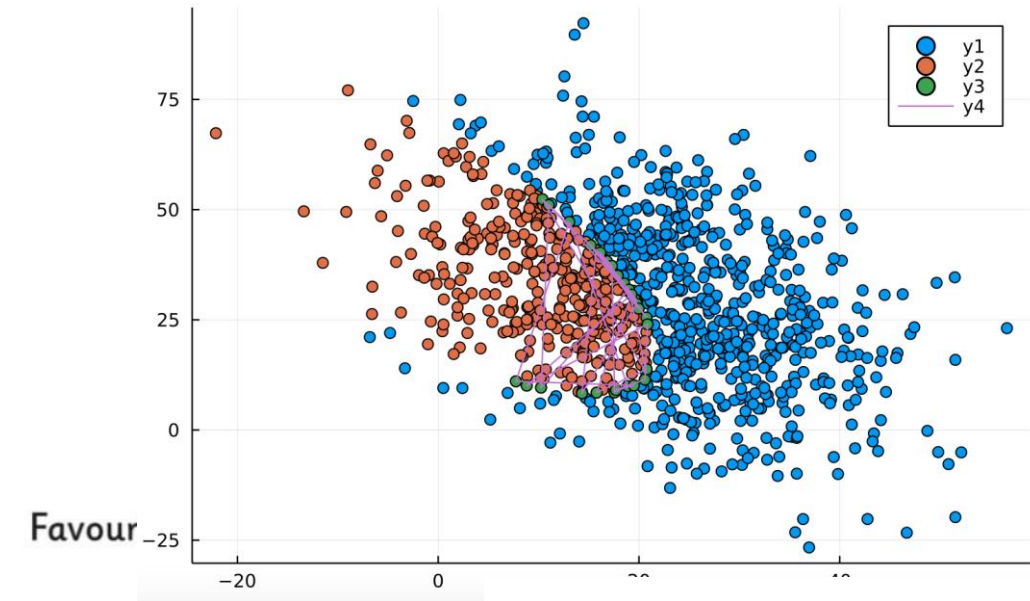
Classes de Dados

Classes de Dados



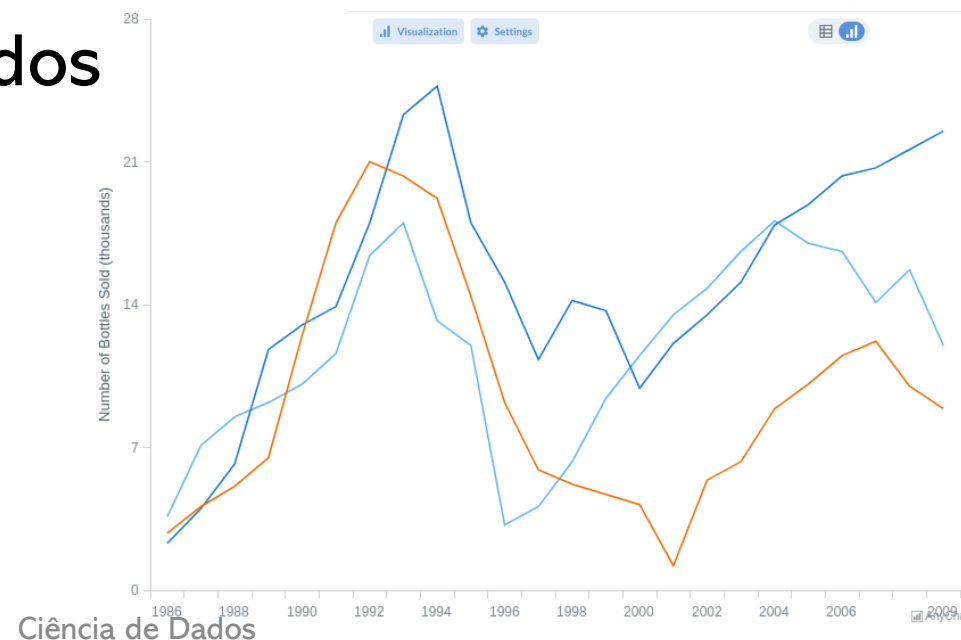
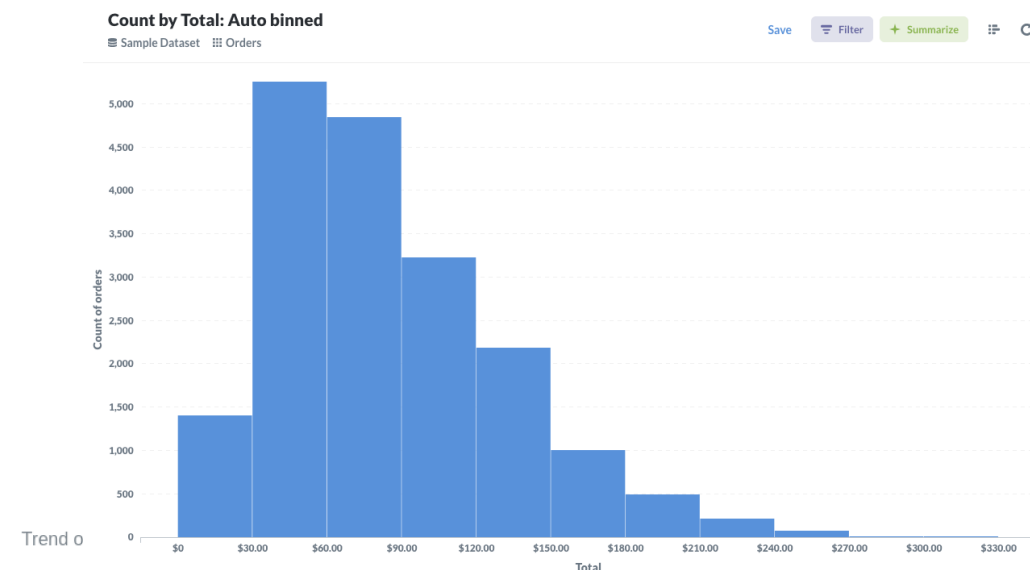
Discretos

- O tipo de dados discreto é um tipo de dados quantitativo que apenas assume valores fixos. Pode ser contado, mas não pode ser medido
- São tipicamente representados por:
 - Pie Charts
 - Bar Charts
 - Scatterplots

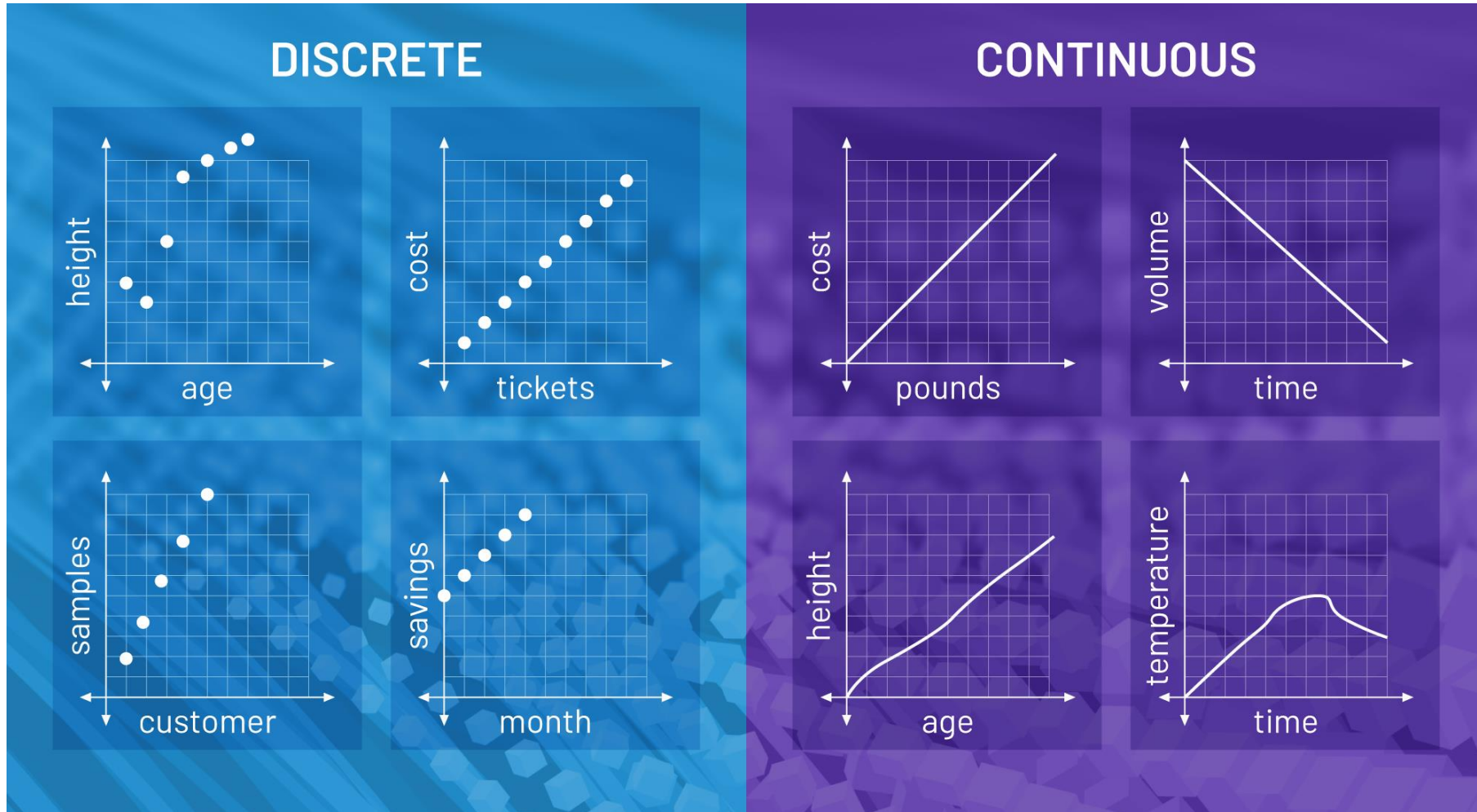


Contínuos

- Dados contínuos podem potencialmente ser medidos com bastante precisão
- São tipicamente representados por:
 - Histograms
 - Line Charts

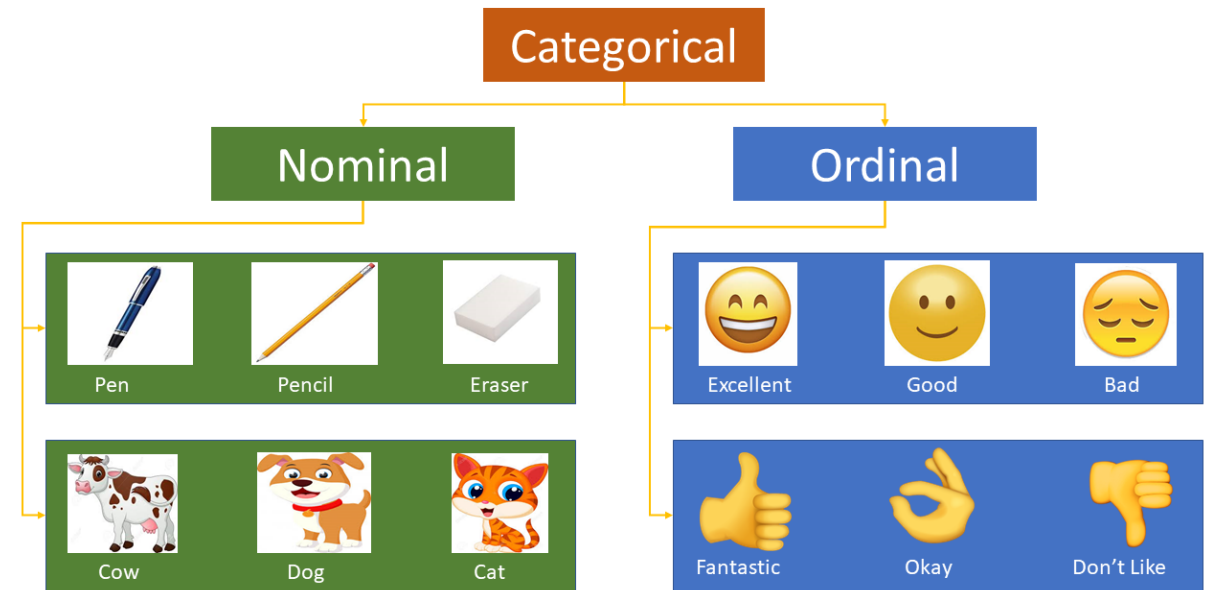


Discretos vs Contínuos




Categorias Dados

- Dados que são tipicamente armazenados em categorias ou grupos, utilizando nomes ou etiquetas
- São tipicamente associados a dados qualitativos
- Existem dois tipos de dados de categoria:
 - Nominais
 - Ordinais



Nominais vs Ordinais

<p>Categorical</p> 	<p>Nominal Can be identified by particular names or categories, and cannot be organized according to any natural order.</p>	<p>Examples</p> <p>Gender : female or male Hair colour: black, blonde etc Favourite sport: soccer, rugby etc</p>	<p>Suitable graphical representation</p> <p>Bar Chart, Pictogram, Pie Chart</p>
	<p>Ordinal Identified by categories which can be ordered in some way</p>	<p>Watching TV: never, rarely, sometimes , a lot</p>	<p>Bar Chart, Pictogram, Pie Chart</p>

Estruturas de Dados

Dados Estruturados

- Os dados estruturados têm por natureza 2 dimensões, e são organizados em linhas e colunas.
- Estão altamente normalizados
- Têm uma elevada integridade e consistência

Dados Não Estruturados

- Não possuem estrutura
- Podem ser textuais ou não textuais
- São tipicamente armazenados em base de dados NoSQL
- Por exemplo:
 - Emails
 - Imagens
 - Redes sociais

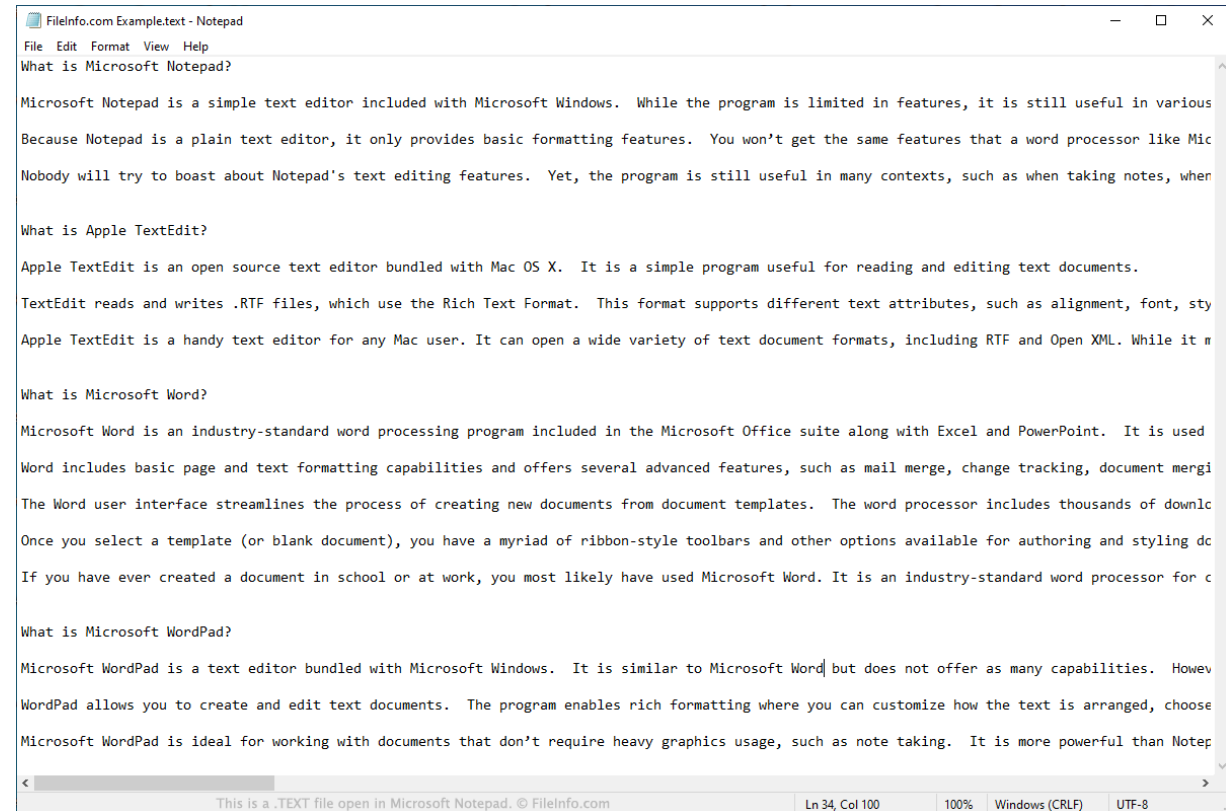
Estruturados vs Não Estruturados

	Structured Data	Unstructured Data
Characteristics	<ul style="list-style-type: none"> • Pre-defined data models • Usually text only • Easy to search 	<ul style="list-style-type: none"> • No pre-defined data model • May be text, images, sound, video or other formats • Difficult to search
Resides in	<ul style="list-style-type: none"> • Relational databases • Data warehouses 	<ul style="list-style-type: none"> • Applications • NoSQL databases • Data warehouses • Data lakes
Generated by	Humans or machines	Humans or machines
Typical applications	<ul style="list-style-type: none"> • Airline reservation systems • Inventory control • CRM systems • ERP systems 	<ul style="list-style-type: none"> • Word processing • Presentation software • Email clients • Tools for viewing or editing media
Examples	<ul style="list-style-type: none"> • Dates • Phone numbers • Social security numbers • Credit card numbers • Customer names • Addresses • Product names and numbers • Transaction information 	<ul style="list-style-type: none"> • Text files • Reports • Email messages • Audio files • Video files • Images • Surveillance imagery

Ficheiros

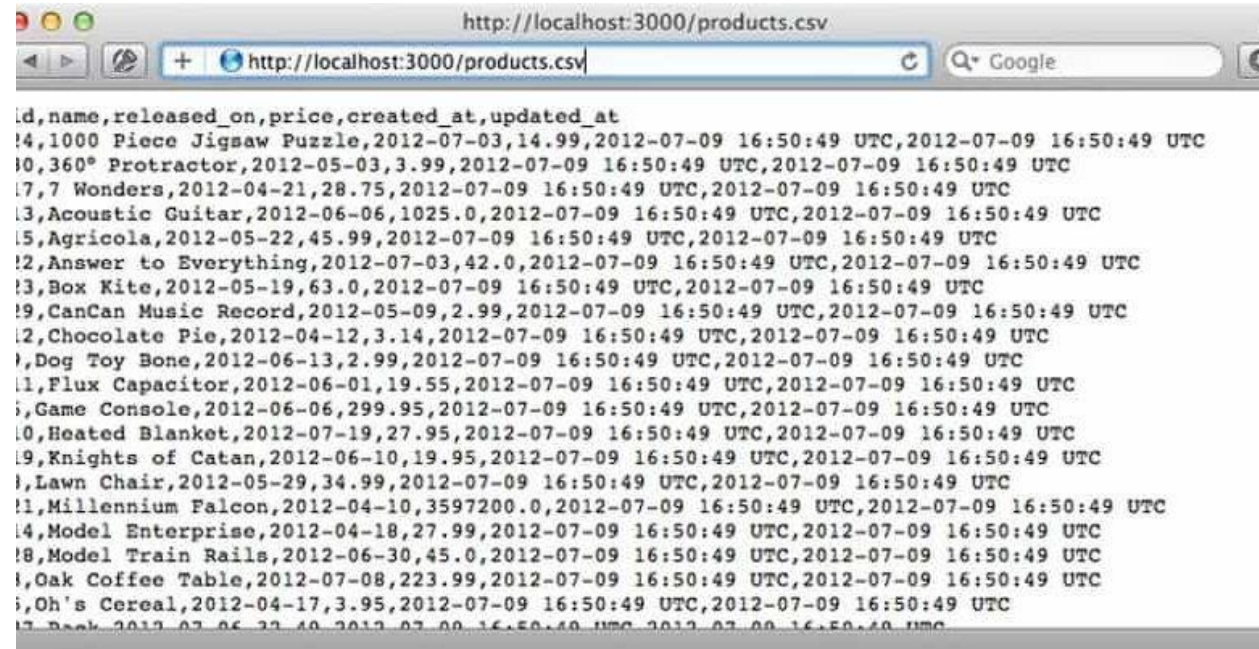
Ficheiros de Texto

- Não contêm ligações para outros ficheiros
- Por exemplo: Um script, um email



CSV

- Comma-Separated Values (CSV), é um ficheiro de texto que é separado por virgulas
- Cada linha do ficheiro é uma linha de informação
- Cada linha contem um ou mais campos separados por virgulas



```

id,name,released_on,price,created_at,updated_at
14,1000 Piece Jigsaw Puzzle,2012-07-03,14.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
10,360° Protractor,2012-05-03,3.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
17,7 Wonders,2012-04-21,28.75,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
13,Acoustic Guitar,2012-06-06,1025.0,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
15,Agricola,2012-05-22,45.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
12,Answer to Everything,2012-07-03,42.0,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
13,Box Kite,2012-05-19,63.0,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
19,CanCan Music Record,2012-05-09,2.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
12,Chocolate Pie,2012-04-12,3.14,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
1, Dog Toy Bone,2012-06-13,2.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
11,Flux Capacitor,2012-06-01,19.55,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
5,Game Console,2012-06-06,299.95,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
10,Heated Blanket,2012-07-19,27.95,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
19,Knights of Catan,2012-06-10,19.95,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
1,Lawn Chair,2012-05-29,34.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
11,Millennium Falcon,2012-04-10,3597200.0,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
14,Model Enterprise,2012-04-18,27.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
18,Model Train Rails,2012-06-30,45.0,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
1,Oak Coffee Table,2012-07-08,223.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
5,Oh's Cereal,2012-04-17,3.95,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
17,Book,2012-07-05,22.40,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
  
```


TSV

- Tab-Separated Value é um ficheiro que usa o “tab” para separar os valores
- Cada linha do ficheiro é uma linha de informação

```

D:\Download\sampleConvertTSVToExcelFormats.tsv - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Window ?
sampleConvertTSVToExcelFormats.tsv
1 Student Science English Math Remarks
2 Simon A C D Do more hard work.
3 Johnson B E A These are nice grades.
4 James C C D These are nice grades.
5 Alex D B E You can perform even better.
6 Emma E A B Well done and keep it up.
7 George D D C Can improve yourself more.
8
Normal text file length: 264 lines: 8 Ln:1 Col:1 Sel:0|0
  
```

XML

- Linguagem semiestruturada
- Utilizador define as suas próprias *tags*
- XML possui algumas regras de formatação
- Flexível

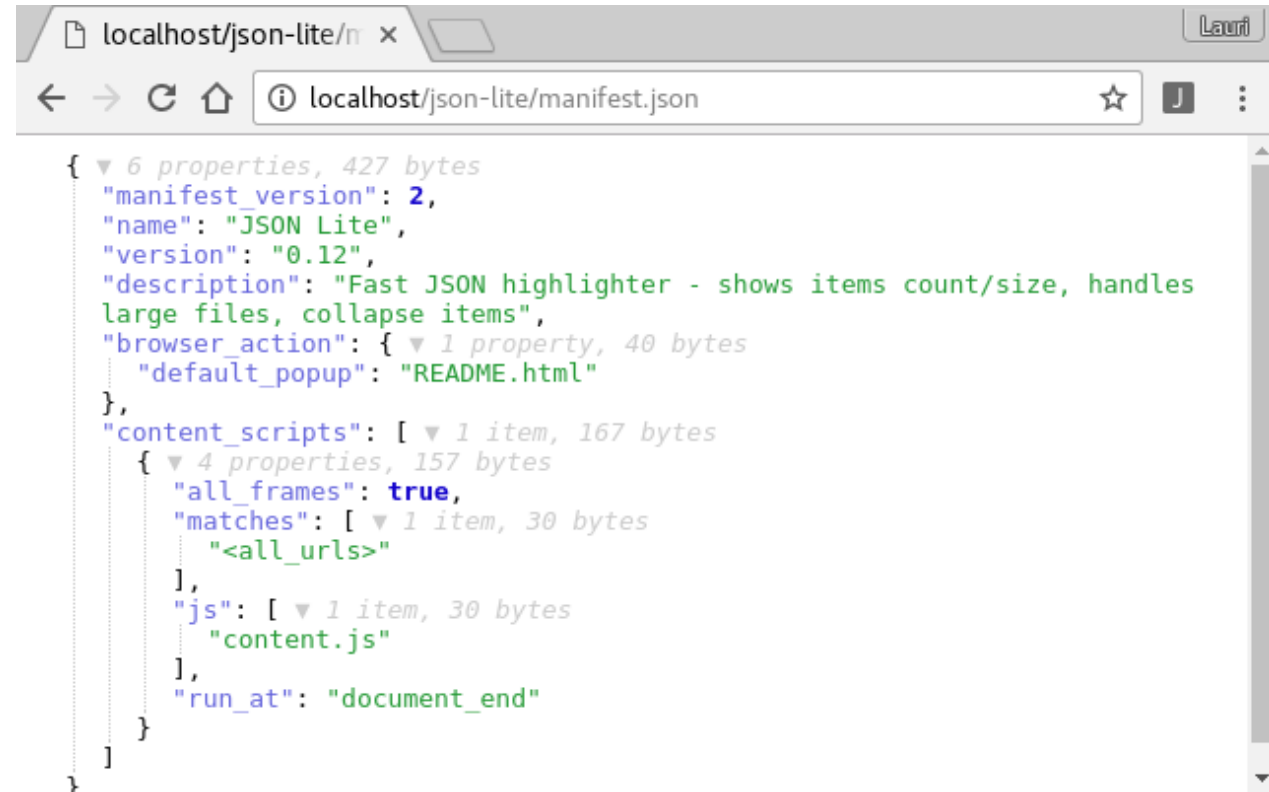
```

1  <?xml version="1.0" encoding="utf-8"?>
2  <Times>
3    <Time Nome="Nome 1" Cidade="Cidade 1">
4      <!--Atributos do time 1-->
5      <Vitorias>Vitorias 1</Vitorias>
6      <Empates>Empates 1</Empates>
7      <Derrotas>Derrotas 1</Derrotas>
8    </Time>
9    <Time Nome="Nome 2" Cidade="Cidade 2">
10     <!--Atributos do time 2-->
11     <Vitorias>Vitorias 2</Vitorias>
12     <Empates>Empates 2</Empates>
13     <Derrotas>Derrotas 2</Derrotas>
14   </Time>
15   <Time Nome="Nome 3" Cidade="Cidade 3">
16     <!--Atributos do time 3-->
17     <Vitorias>Vitorias 3</Vitorias>
18     <Empates>Empates 3</Empates>
19     <Derrotas>Derrotas 3</Derrotas>
20   </Time>
21 </Times>

```

JSON

- JavaScript Object Notation
- Linguagem semiestruturada
- Consiste num par campo/valor
- Utilizado para comunicação entre cliente e servidor
- Extensão .json



```
{
  "manifest_version": 2,
  "name": "JSON Lite",
  "version": "0.12",
  "description": "Fast JSON highlighter - shows items count/size, handles large files, collapse items",
  "browser_action": {
    "default_popup": "README.html"
  },
  "content_scripts": [
    {
      "all_frames": true,
      "matches": [
        "<all_urls>"
      ],
      "js": [
        "content.js"
      ],
      "run_at": "document_end"
    }
  ]
}
```

XML

```
<empinfo>
  <employees>
    <employee>
      <name>James Kirk</name>
      <age>40</age>
    </employee>
    <employee>
      <name>Jean-Luc Picard</name>
      <age>45</age>
    </employee>
    <employee>
      <name>Wesley Crusher</name>
      <age>27</age>
    </employee>
  </employees>
</empinfo>
```

JSON

```
{ "empinfo" :
  {
    "employees" : [
      {
        "name" : "James Kirk",
        "age" : 40,
      },
      {
        "name" : "Jean-Luc Picard",
        "age" : 45,
      },
      {
        "name" : "Wesley Crusher",
        "age" : 27,
      }
    ]
  }
}
```

HTML

- Hypertext Markup Language
- Código utilizado para definir a estrutura de uma pagina web
- Linguagem estruturada

```
font1.html
1 <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/
  xhtml1-strict.dtd">
2
3 <html xmlns="http://www.w3.org/1999/xhtml">
4   <head>
5     <meta name="viewport" content="width=device-width, initial-scale=1.0, maximum-
      scale=1.0, user-scalable=no, target-densitydpi=device-dpi"/>
6     <title>font1.html</title>
7     <link rel="stylesheet" href="../HTMLResources/css/style.css" type="text/css" />
8   </head>
9   <body>
10    <div id="untitled-1">
11      Font Embedding Test - Uses "style.css."
12
13      <div class="myFontOne">
14        <p>Flood Font</p>
15      </div>
16      <p>The text above should be RED and be in the font "Flood." The text is
        styled with "myFontOne."</p>
17
18      <div class="myFontTwo">
19        <p>FetteFraktur Font</p>
20      </div>
21      <p>The text above should be GREEN and be in the font "FetteFraktur." The
        text is styled with "myFontTwo."</p>
22
23      <br>
24      <p class="basic-paragraph"><b>Compare the fonts above to this screen
        shot.</b></p>
25      
26
27    </div>
28  </body>
29 </html>
30
31
```

Bibliografia

- B. Gomez,(2020) “Resolviendo problemas de Big Data”, Alfaomega.
- D. Insua, (2019)“Big data: Conceptos, tecnologías y aplicaciones”, CSIC.
- H. Jones, (2019)“Analítica de datos”, HJ,.
- J. Somed, (2020)“Big Data Analytics”, JLC.
- D. Petković (2020)“Microsoft® SQL Server® 2019 A Beginner's Guide - Seventh Edition”, McGraw Hill.