

Unidade Curricular de Armazenamento para Big Data

Relatório sobre:
Transformação do *dataset*: Toys and Games



Elaborado por:

André Simões Novo, n.º 93343

Filipe Cordeiro Hristovsky, n.º 93949

Luís Miguel dos Santos Pereira, n.º 98398

Sebastião Manuel Inácio Rosalino, n.º 98437

Simão Tadeu Castelo Miguel, n.º 99064

Licenciatura de Ciência de Dados - 2º ano - Turma CD

Ano Letivo 2021/2022 - 1º Semestre

Coordenador da UC:

Professor Pedro Ramos

Docente:

Professora Joana Martinho Costa

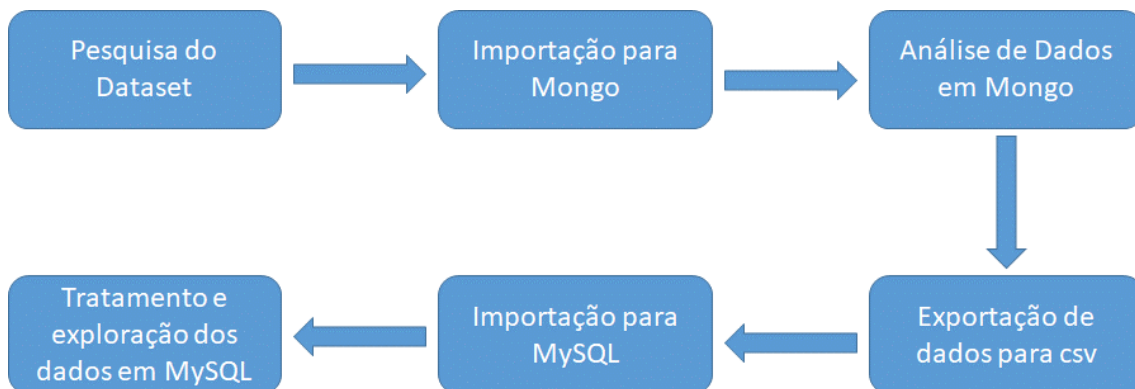
Data de entrega: 12 dezembro de 2021

Índice

1. Plano de trabalho	3
2. Identificação e pesquisa do Dataset	3
3. Processo de importação para mongo	7
4. Análise de Dados em Mongo	8
6. Importação de dados para MySQL e desenho da base de dados	12
7. Tratamento e exploração dos dados em Mysql.....	17
a) Limpeza dos dados em MySQL	17
b) Análise da informação em MySQL.....	18

1. Plano de trabalho

As etapas de desenvolvimento do presente trabalho foram as seguintes:



2. Identificação e pesquisa do Dataset

Como primeira tarefa, procedeu-se à pesquisa de um *dataset* que contivesse, no mínimo, meio milhão de registos, que estivesse no formato “.json” e que pudesse criar um modelo relacional com três tabelas de dados.

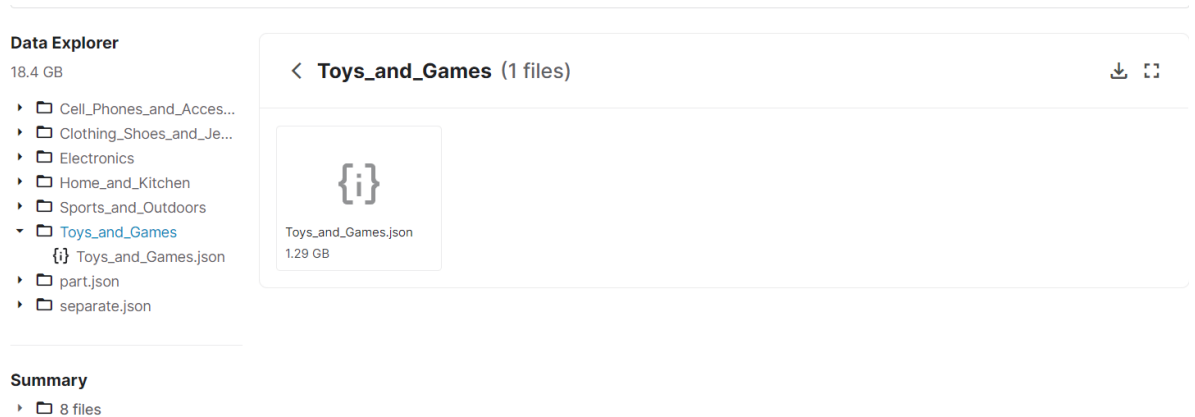
Para além desta condição, procurou-se seleccionar uma base de dados que contivesse informação interessante e, ao mesmo tempo, pertinente para o desenvolvimento das tarefas do projeto proposto.

A pesquisa foi realizada na plataforma Kaggle e foi escolhida uma base de dados relativa a “reviews” sobre compras efetuadas na *store on-line* da Amazon, com a designação geral de “Amazon Product Review (Spam and Non Spam)”.



O conjunto dos dados respeita a avaliações (por clientes) de produtos da Amazon. Trata-se de um conjunto de datasets (8) sobre diferentes categorias de produtos que contém cerca de 26,7 milhões de revisões e 15,4 milhões de revisores.

O atributo classe indica as reviews cujo overall é considerado bom ou muito bom (notação de 4 ou 5), assumindo a classificação de “1”. Para as restantes notações assume a classificação “0”.



O *dataset* objeto do presente trabalho respeita a *reviews* de produtos vendidos da categoria Toys and Games (jogos e brinquedos), no site da Amazon.

Link:

https://www.kaggle.com/naveedhn/amazon-product-review-spam-and-non-spam?select=Toys_and_Games

Descrição dos campos da coleção:

_id - Id da review

reviewerID - Id do autor da review

asin - Id do produto em review

reviewerName - Nome do autor da review

helpful – Número de likes e dislikes em array

reviewText - Texto completo da review

overall - Classificação do produto (sendo 1 a pior e 5 a melhor)

summary – Título do texto da review

unixReviewTime - Data da review em número

reviewTime - Data da review

category - Categoria do produto em review

class - Classificação “1” (review boa) ou “0” (review má ou razoável)

Exemplo de um documento da coleção:

```

{
  "root": {
    "items": [
      {
        "_id": {
          "$oid": "5a13282b741a2384e879a620"
        },
        "reviewerID": "A3C9CSW3TJITGT",
        "asin": "0005069491",
        "reviewerName": "Renee",
        "helpful": [
          {
            "text": "I love these felt nursery rhyme characters and scenes. The quality of the felt is good, and the illustrations are detailed and pretty. As noted, the figures and scenes are printed on 2 large sheets of flannel and each individual item needs to be cut out. This process took me 2 hours of tiny cutting. To me it does not lend itself to a book form but rather laying out the scenes separately or for use on a flannel board. However, I love the quiet play it offers for my toddler, and as a former Kindergarten teacher, I understand the value of learning rhyme and its connection to future reading. Overall, delightful product with some work involved."
          },
          {
            "text": ""
          }
        ],
        "reviewText": "I love these felt nursery rhyme characters and scenes. The quality of the felt is good, and the illustrations are detailed and pretty. As noted, the figures and scenes are printed on 2 large sheets of flannel and each individual item needs to be cut out. This process took me 2 hours of tiny cutting. To me it does not lend itself to a book form but rather laying out the scenes separately or for use on a flannel board. However, I love the quiet play it offers for my toddler, and as a former Kindergarten teacher, I understand the value of learning rhyme and its connection to future reading. Overall, delightful product with some work involved.",
        "overall": 4,
        "summary": "Charming characters but busy work required",
        "unixReviewTime": 1377561600,
        "reviewTime": "08 27, 2013",
        "category": "Toys_and_Games",
        "class": 1
      }
    ]
  }
}
```

Quantidade coleções/documentos:




Tem uma coleção (Toys_and_Games) e 1.997.140 de documentos

Visualização de um documento (registro) em mongo compass:



Tipo de informação estatística que se pode obter:

Nome e tipo das variáveis:

	#	Nome	Tipo
<input type="checkbox"/>	1	_id 	varchar(100)
<input type="checkbox"/>	2	asin 	varchar(100)
<input type="checkbox"/>	3	class	int(1)
<input type="checkbox"/>	4	helpful	longtext
<input type="checkbox"/>	5	overall	int(10)
<input type="checkbox"/>	6	reviewText	text
<input type="checkbox"/>	7	reviewTime	varchar(100)
<input type="checkbox"/>	8	reviewerID 	varchar(100)
<input type="checkbox"/>	9	summary	text
<input type="checkbox"/>	10	unixReviewTime	int(100)
<input type="checkbox"/>	11	likes	int(11)
<input type="checkbox"/>	12	dislikes	int(11)
<input type="checkbox"/>	2	category	text

Algumas das estatísticas que podem ser obtidas:

- a) Número de reviews para cada overall
- b) Percentagem de reviews com um certo overall
- c) Média de likes e dislikes por review

3. Processo de importação para mongo

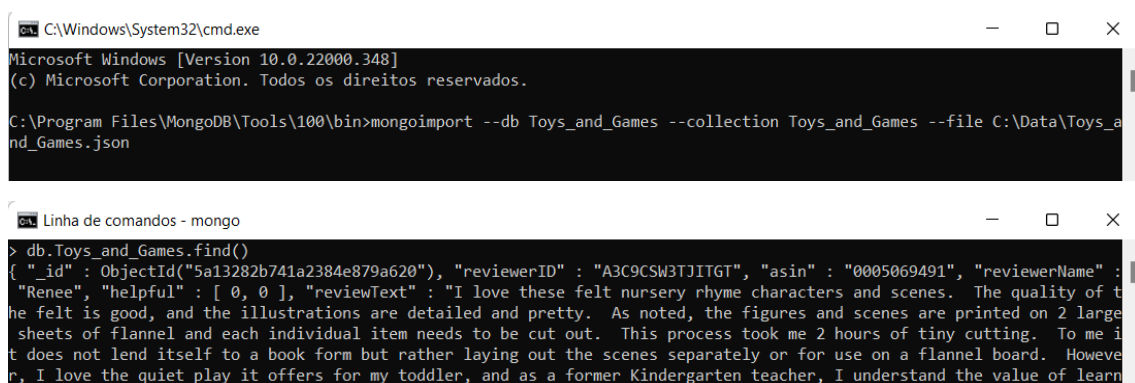
Como segunda etapa, importou-se o ficheiro “.json” (dataset Toys_and_Games) para o mongo com o objetivo de o exportar, posteriormente, para formato “.csv”, para finalmente o transportar para o MySQL.

O desenvolvimento destes procedimentos permitirá, como objetivo final, trabalhar os dados numa base de modelo relacional (com pelo menos 3 tabelas) no MySQL, utilizando os conhecimentos adquiridos na Unidade Curricular Fundamentos de Gestão de Bases de Dados.

Código usado para a importação do dataset:

```
mongoimport --db Toys_and_Games --collection Toys_and_Games --file  
C:\Data\Toys_and_Games.json
```

Imagens dos procedimentos realizados:



- Foi necessário retirar o formato “jsonArray” ao dataset para permitir a sua importação com o comando indicado.

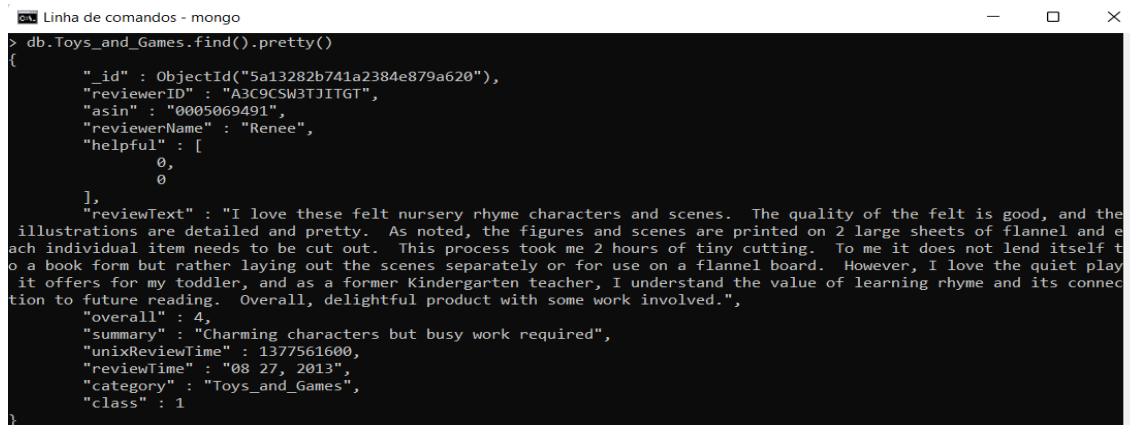
4. Análise de Dados em Mongo

Transportado o ficheiro para Mongo foram realizadas consultas, na linha de comandos, sobre a informação disponível, com o propósito de melhor compreender a estrutura e o conteúdo da coleção para eventual, posterior, limpeza de dados.

Alguns exemplos de consultas realizadas em Mongo:

a) Consulta de todas as reviews

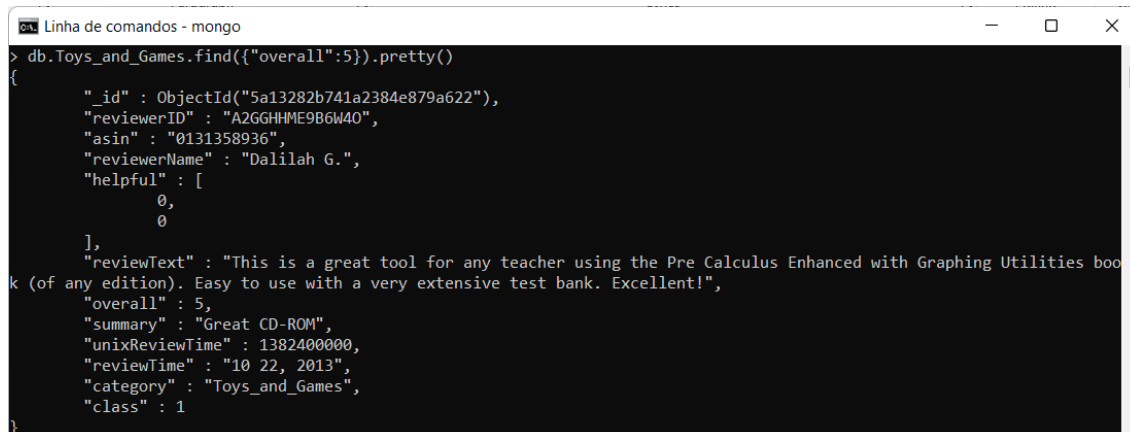
```
db.Toys_and_Games.find().pretty()
```



```
Linha de comandos - mongo
> db.Toys_and_Games.find().pretty()
{
  "_id" : ObjectId("5a13282b741a2384e879a620"),
  "reviewerID" : "A3C9CSW3TJITGT",
  "asin" : "0005069491",
  "reviewerName" : "Renee",
  "helpful" : [
    0,
    0
  ],
  "reviewText" : "I love these felt nursery rhyme characters and scenes. The quality of the felt is good, and the illustrations are detailed and pretty. As noted, the figures and scenes are printed on 2 large sheets of flannel and each individual item needs to be cut out. This process took me 2 hours of tiny cutting. To me it does not lend itself to a book form but rather laying out the scenes separately or for use on a flannel board. However, I love the quiet play it offers for my toddler, and as a former Kindergarten teacher, I understand the value of learning rhyme and its connection to future reading. Overall, delightful product with some work involved.",
  "overall" : 4,
  "summary" : "Charming characters but busy work required",
  "unixReviewTime" : 1377561600,
  "reviewTime" : "08 27, 2013",
  "category" : "Toys_and_Games",
  "class" : 1
}
```

b) Consulta de todas as reviews cujo overall foi máximo

```
db.Toys_and_Games.find({"overall":5}).pretty()
```



```
Linha de comandos - mongo
> db.Toys_and_Games.find({"overall":5}).pretty()
{
  "_id" : ObjectId("5a13282b741a2384e879a622"),
  "reviewerID" : "A2GGHME9B6W40",
  "asin" : "0131358936",
  "reviewerName" : "Dalilah G.",
  "helpful" : [
    0,
    0
  ],
  "reviewText" : "This is a great tool for any teacher using the Pre Calculus Enhanced with Graphing Utilities book (of any edition). Easy to use with a very extensive test bank. Excellent!",
  "overall" : 5,
  "summary" : "Great CD-ROM",
  "unixReviewTime" : 1382400000,
  "reviewTime" : "10 22, 2013",
  "category" : "Toys_and_Games",
  "class" : 1
}
```


c) Número de reviews com overall máximo

```
db.Toys_and_Games.count({"overall":5}) – 1275445 (Classificações máximas)
```

```
Linha de comandos - mongo
> db.Toys_and_Games.count({"overall":5})
1275445
```

d) Número de reviews com pontuação 4 ou 5

```
db.Toys_and_Games.count({"class":1}) – 1662754
```

```
Linha de comandos - mongo
> db.Toys_and_Games.count({"class":1})
1662754
```

e) Número de reviews com pontuação 1,2 ou 3

```
db.Toys_and_Games.count({"class":0}) – 334386
```

```
Linha de comandos - mongo
> db.Toys_and_Games.count({"class":0})
334386
```

f) Consulta de todos os autores de reviews, juntamente com o id de review

```
db.Toys_and_Games.find({}, {"reviewerName":1})
```

```
Linha de comandos - mongo
> db.Toys_and_Games.find( {}, {"reviewerName":1} )
{ "_id" : ObjectId("5a13282b741a2384e879a620"), "reviewerName" : "Renee" }
{ "_id" : ObjectId("5a13282b741a2384e879a621"), "reviewerName" : "So CA Teacher" }
{ "_id" : ObjectId("5a13282b741a2384e879a622"), "reviewerName" : "Dalilah G." }
```

g) Consulta de todas as datas das reviews, sem o seu id

```
db.Toys_and_Games.find({}, {"reviewTime":1, "_id":0})
```

```
Linha de comandos - mongo
> db.Toys_and_Games.find({}, {"reviewTime":1, "_id":0})
{ "reviewTime" : "08 27, 2013" }
{ "reviewTime" : "07 9, 2014" }
{ "reviewTime" : "10 22, 2013" }
```

h) Média do overall de todas as reviews

```
db.Toys_and_Games.aggregate( [ { $group: { _id: "overall", avgOverall: { $avg: "$overall" } } } ] ) - 4.304242566870625
```

```
ca. Linha de comandos - mongo
> db.Toys_and_Games.aggregate( [ { $group: { _id: "overall", avgOverall: { $avg: "$overall" } } } ] )
{ "_id" : "overall", "avgOverall" : 4.304242566870625 }
```

5. Exportação para CSV

Analisada a informação em Mongo, procedeu-se de seguida à exportação do ficheiro “json”, na sua integralidade, para um formato “.csv”,

Comandos de exportação para csv:

```
mongoexport --db Toys_and_Games --collection Toys_and_Games --type=csv --fields _id,reviewerID,asin,reviewerName,helpful,reviewText,overall,summary,unixReviewTime,reviewTime,category,class --out C:\Data\Toys_and_Games.csv
```

Imagens dos procedimentos de importação realizados:

```
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.22000.348]
(c) Microsoft Corporation. Todos os direitos reservados.

C:\Program Files\MongoDB\Tools\100\bin>mongoexport --db Toys_and_Games --collection Toys_and_Games --type=csv --fields _id,reviewerID,asin,reviewerName,helpful,reviewText,overall,summary,unixReviewTime,reviewTime,category,class --out C:\Data\Toys_and_Games.csv
```

Primeiras linhas do ficheiro csv:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	_id	reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime	category	class											
2	Objectid(Sa13282b741a2384e879a620)	A3C9C5W3TJITGT,0005069491	Renee	[0,0]	"I love these felt nursery rhyme characters and scenes. The quality of the felt is good, and the illustrations are detailed and pretty. As noted, the figures and scenes are pr																		
3	Objectid(Sa13282b741a2384e879a621)	A31POTIVKCS295,0076561046	So CA Teacher	[0,0]	"I see no directions for its use. Therefore I have to make up the games, unfortunately,"3, No directions for use....1404864000,"07 5, 2014",Toys_and_Games,0																		
4	Objectid(Sa13282b741a2384e879a622)	A2G6GHME986940,0131358936	Dalliah G.	[0,0]	"This is a great tool for any teacher using the Pre Calculus Enhanced with Graphing Utilities book (of any edition). Easy to use with a very extensive test bank. Exceller																		
5	Objectid(Sa13282b741a2384e879a61f)	AMEVEO2LYVEJA,0000191639	Nicole Soeder	[0,0]	"Great product, thank you! Our son loved the puzzles. They have large pieces yet they are still challenging for a 4 year old."5,Puzzles,1388016000,"12 26, 2013",Toy																		
6	Objectid(Sa13282b741a2384e879a623)	A1FSLDH43ORWZP,0133642984	Dayna English	[0,0]	"Although not as streamlined as the Algebra I materials ... this is extremely helpful for first time teachers ... bulk of materials are prepared for presentations,"5,Alge																		

Dificuldades sentidas:

- O dataset original tem 1.997.140 registos. O ficheiro csv apenas mostra 1.048.576 registos. Para ultrapassar esta dificuldade foi decidido trabalhar apenas com este número de registos. Assim, em relação ao original trabalhou-se com 52,503880% dos dados.

Para se proceder ao tratamento dos dados no phpMyAdmin foram exportados, a partir do ficheiro “.json”, 3 ficheiros para “.csv”, onde cada ficheiro contém apenas as colunas necessárias a cada tabela no modelo relacional a criar.

Os comandos utilizados para a criação de cada um desses ficheiros “.csv” foram os seguintes:

- Para a tabela **review**:

```
mongoexport --db Toys_and_Games --collection Toys_and_Games --type=csv -  
-fields  
_id,asin,class,helpful,overall,reviewText,reviewTime,reviewerID,summary,unix  
ReviewTime --out Toys_and_Games_tabela_review.csv
```

- Para a tabela **produto**:

```
mongoexport --db Toys_and_Games --collection Toys_and_Games --type=csv -  
-fields asin,category --out Toys_and_Games_tabela_produto.csv
```

- Para a tabela **reviewer**:

```
mongoexport --db Toys_and_Games --collection Toys_and_Games --type=csv -  
-fields reviewerID,reviewerName --out Toys_and_Games_tabela_reviewer.csv
```

Com estes procedimentos, toda a informação foi exportada com sucesso para o formato “.csv”.

6. Importação de dados para MySQL e desenho da base de dados

Uma vez efetuada a transformação e limpeza dos dados, procedeu-se à estruturação do modelo relacional. Para esse efeito, optou-se pela criação de 5 tabelas cuja estrutura será descrita de seguida:

I) A **tabela review** diz respeito a todas as informações relevantes referentes a uma review. São estas: o id da review; o id do produto em revisão; o overall da review; um vetor contendo os gostos (likes) e os não gostos (dislikes); o texto completo da review; a data da review; o id do autor da review; o título da review; a data da review em formato numérico; e o número de gostos e não gostos.

Apresenta-se de seguida a estrutura da tabela:

#	Nome	Tipo	Agrupamento (Collation)	Atributos	Nulo	Predefinido	Comentários	Extra	Ações
<input type="checkbox"/> 1	<u>id</u>	varchar(100)	utf8mb4_general_ci		Não	Nenhum			Muda Elimina Mais
<input type="checkbox"/> 2	<u>asin</u>	varchar(100)	utf8mb4_general_ci		Não	Nenhum			Muda Elimina Mais
<input type="checkbox"/> 3	<u>class</u>	int(1)			Não	Nenhum			Muda Elimina Mais
<input type="checkbox"/> 4	<u>helpful</u>	longtext	utf8mb4_bin		Não	Nenhum			Muda Elimina Mais
<input type="checkbox"/> 5	<u>overall</u>	int(10)			Não	Nenhum			Muda Elimina Mais
<input type="checkbox"/> 6	<u>reviewText</u>	text	utf8mb4_general_ci		Não	Nenhum			Muda Elimina Mais
<input type="checkbox"/> 7	<u>reviewTime</u>	varchar(100)	utf8mb4_general_ci		Não	Nenhum			Muda Elimina Mais
<input type="checkbox"/> 8	<u>reviewerID</u>	varchar(100)	utf8mb4_general_ci		Não	Nenhum			Muda Elimina Mais
<input type="checkbox"/> 9	<u>summary</u>	text	utf8mb4_general_ci		Não	Nenhum			Muda Elimina Mais
<input type="checkbox"/> 10	<u>unixReviewTime</u>	int(100)			Não	Nenhum			Muda Elimina Mais
<input type="checkbox"/> 11	<u>likes</u>	int(11)			Não	Nenhum			Muda Elimina Mais
<input type="checkbox"/> 12	<u>dislikes</u>	int(11)			Não	Nenhum			Muda Elimina Mais

Em relação aos dados presentes no dataset original, foram criados em MySQL o campo *likes* e *dislikes* a partir do array *helpful*.

O código das suas criações foram os seguintes:

- Para o campo *likes*:

```
SELECT CONVERT(REPLACE(SUBSTRING_INDEX(helpful,',',1),"",""),int)
FROM review group BY
CONVERT(REPLACE(SUBSTRING_INDEX(helpful,',',1),"",""),int)
```

- Para o campo *dislikes*:

```
SELECT CONVERT(REPLACE(SUBSTRING_INDEX(helpful,',',-
1),"",""),int) FROM review group BY
CONVERT(REPLACE(SUBSTRING_INDEX(helpful,',',-1),"",""),int)
```

II) A **tabela produto** integra os atributos relevantes para os produtos avaliados.

Os campos são: o id (asin) e a sua categoria (que neste caso serão todas Toys_and_Games)

Apresenta-se de seguida a estrutura da tabela:

#	Nome	Tipo	Agrupamento (Collation)	Atributos	Nulo	Predefinido	Comentários	Extra	Acções
<input type="checkbox"/>	1 asin	varchar(100)	utf8mb4_general_ci		Não	Nenhum			Muda Elimina Mais
<input type="checkbox"/>	2 category	text	utf8mb4_general_ci		Não	Nenhum			Muda Elimina Mais

III) A tabela reviewer diz respeito a todas as informações relevantes referentes a um reviewer. Os campos são: o seu id e o seu nome:

#	Nome	Tipo	Agrupamento (Collation)	Atributos	Nulo	Predefinido	Comentários	Extra	Acções
<input type="checkbox"/>	1 reviewerID	varchar(100)	utf8mb4_general_ci		Não	Nenhum			Muda Elimina Mais
<input type="checkbox"/>	2 reviewerName	text	utf8mb4_general_ci		Não	Nenhum			Muda Elimina Mais

Procedeu-se, de seguida, à importação dos dados para cada tabela, segundo os seguintes comandos.

- Para a tabela review:

```
1 LOAD DATA INFILE 'C:/Data/Toys_and_Games_tabela_review.csv'
2 INTO TABLE review
3 FIELDS TERMINATED BY ','
4 ENCLOSED BY '"'
5 LINES TERMINATED BY '\n'
6 IGNORE 1 ROWS;
```

- Para a tabela produto:

```
1 LOAD DATA INFILE 'C:/Data/Toys_and_Games_tabela_produto.csv'
2 INTO TABLE produto
3 FIELDS TERMINATED BY ','
4 ENCLOSED BY '"'
5 LINES TERMINATED BY '\n'
6 IGNORE 1 ROWS;
```

- Para a tabela reviewer:

```
1 LOAD DATA INFILE 'C:/Data/Toys_and_Games_tabela_reviewer.csv'
2 INTO TABLE reviewer
3 FIELDS TERMINATED BY ','
4 ENCLOSED BY '"'
5 LINES TERMINATED BY '\n'
6 IGNORE 1 ROWS;
```

Para a importação desta tabela surgiu, no entanto, uma dificuldade que resulta da circunstância de ser possível que um mesmo reviewer tenha feito mais do que uma review, o que levaria à duplicação de entradas (de reviewers).

É também possível que um mesmo produto tenha sido sujeito a mais do que uma review. Não será, então, possível criar as chaves e as relações devido à repetição de registos (que terão que ser únicos).

Para ultrapassar este problema, foram então criadas duas tabelas **produto_clean** e **reviewer_clean** que contêm os dados únicos referentes às suas tabelas originais.

O código usado para criar essas novas tabelas foi o seguinte:

- Tabela produto_clean:

```
1 CREATE TABLE 'produto_clean'
2 AS SELECT DISTINCT(produto.asin), produto.category
3 FROM produto;
```

- Tabela reviewer_clean:

```
1 CREATE TABLE reviewer_clean
2 AS SELECT DISTINCT(reviewee.reviewerID), MAX(reviewee.reviewerName), reviewee.reviewerName
3 FROM reviewer
4 GROUP BY reviewer.reviewerID;
```

Existia anteriormente uma inconsistência de dados para a tabela reviewer: para um mesmo Id de reviewer, existiam vários nomes, mas referentes à mesma pessoa. O comando max foi utilizado para extrair para a nova tabela apenas um nome para cada Id encontrado, sendo este o nome completo pois a **função max** retornará o nome com maior tamanho.

O terceiro argumento do select (reviewer.reviewer.Name) foi posteriormente apagado. Servia apenas de confirmação. O segundo argumento do select (max(reviewer.reviewerName) foi renomeado para apenas reviewerName.

A estrutura das novas tabelas passou a ser a seguinte:

- produto_clean:

#	Nome	Tipo	Agrupamento (Collation)	Atributos	Nulo	Predefinido	Comentários	Extra	Ações
<input type="checkbox"/> 1	asin	varchar(100)	utf8mb4_general_ci		Não	Nenhum			Muda Elimina Mais
<input type="checkbox"/> 2	category	text	utf8mb4_general_ci		Não	Nenhum			Muda Elimina Mais

- reviewer_clean:

#	Nome	Tipo	Agrupamento (Collation)	Atributos	Nulo	Predefinido	Comentários	Extra	Ações
<input type="checkbox"/> 1	reviewerID	varchar(100)	utf8mb4_general_ci		Não	Nenhum			Muda Elimina Mais
<input type="checkbox"/> 2	reviewerName	text	utf8mb4_general_ci		Sim	NULL			Muda Elimina Mais

As tabelas ficaram, assim, prontas a serem utilizadas para estabelecimento de relações.

As chaves primárias criadas foram:

- para a tabela reviewer: ***_id***,
- para a tabela produto_clean: ***asin***
- para a tabela reviewer_clean: ***reviewerID***

As chaves estrangeiras criadas foram as seguintes:

- Para a relação entre a coluna **asin** da tabela review e a coluna **asin** da tabela produto_clean:

```
ALTER TABLE `review` ADD CONSTRAINT `fk_asin` FOREIGN KEY (`asin`) REFERENCES `produto_clean` (`asin`) ON DELETE CASCADE ON UPDATE CASCADE
```

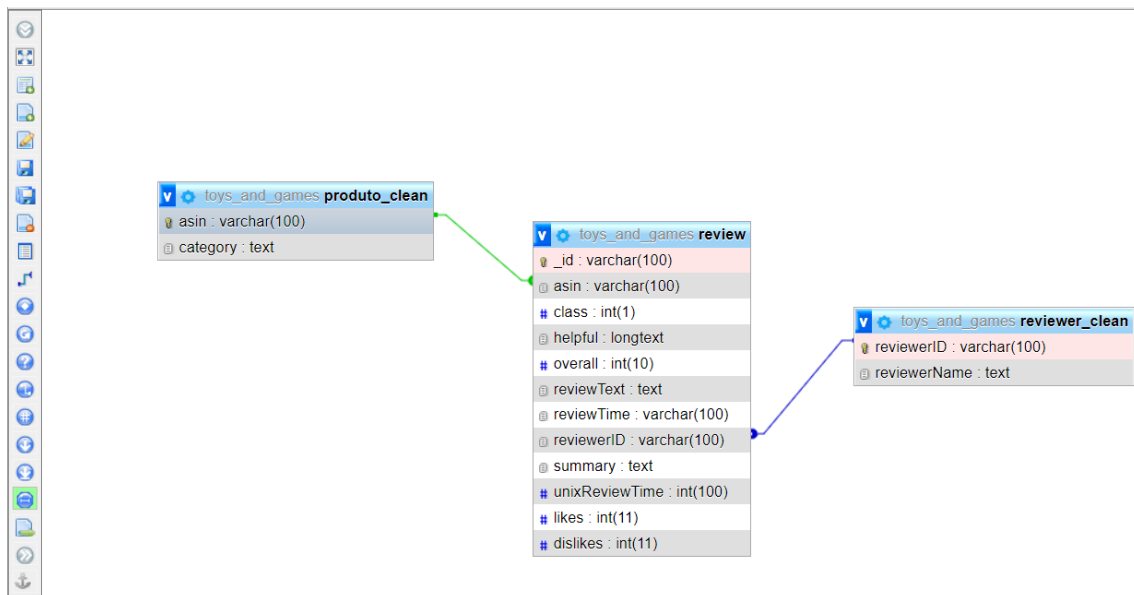
- Para a relação entre a coluna **reviewerID** da tabela review e a coluna **reviewerID** da tabela reviewer_clean:

```
ALTER TABLE `review` ADD CONSTRAINT fk_reviewerid FOREIGN
KEY (`reviewerID`) REFERENCES `reviewer_clean`(`reviewerID`) ON
DELETE CASCADE ON UPDATE CASCADE;
```



Desenho da Base de Dados relacional

Concluído o tratamento, a limpeza e a transformação dos dados, procedeu-se à estruturação do modelo relacional, como apresentado na figura abaixo:



7. Tratamento e exploração dos dados em Mysql

a) Limpeza dos dados em MySQL

Após toda a importação dos dados para as respectivas tabelas, foram detetadas algumas inconsistências nos dados. Nomeadamente, na tabela review, no campo reviewTime que deveria conter datas em formato *string*, estavam presentes 35 registos que não continham datas.

Todos os 35 registos incorretos começavam com “A”, portanto o comando utilizado foi o seguinte:

```
DELETE FROM review  
WHERE reviewTime like 'A%'
```

A segunda inconsistência foi encontrada na tabela review. Existiam reviewers que não tinham o seu ID presente na tabela reviewer_clean.

O código para os encontrar foi o seguinte:

```
SELECT * FROM review  
  
WHERE review.reviewerID NOT IN (SELECT reviewer_clean.reviewerID *  
FROM reviewer_clean
```

Foram, assim, removidos estes registos.

b) Análise da informação em MySQL

Construído a modelo relacional e feita a limpeza dos dados, procedeu-se à exploração de informação através do MySQL.

Apresentam-se, de seguida, um conjunto de comandos realizados para obtenção de informação útil sobre o dataset tratado:

1) Número de reviews para cada overall

✓ A mostrar registos de 0 - 4 (5 total, A consulta demorou 2,9931 segundos.)

```
SELECT review.overall, COUNT(*) as 'número de reviews para cada overall' FROM review GROUP BY review.overall;
```

☐ Perfil [Editar em linha] [Edita] [Explicar SQL] [Criar código PHP] [Actualizar]

☐ Mostrar tudo | Número de registos: 25 ▼ Filtrar registos:

+ Opções

overall	número de reviews para cada overall
1	128149
2	77129
3	129093
4	387303
5	1275424

2) Número de reviews para cada overall e em percentagem

✓ A mostrar registos de 0 - 4 (5 total, A consulta demorou 4,7533 segundos.)

```
SELECT review.overall, COUNT(*) as 'número de reviews para cada overall', (COUNT(*) / (SELECT COUNT(*) FROM review)) * 100 as '%' FROM review GROUP BY review.overall;
```

☐ Perfil [Editar em linha] [Edita] [Explicar SQL] [Criar código PHP] [Actualizar]

☐ Mostrar tudo | Número de registos: 25 ▼ Filtrar registos:

+ Opções

overall	número de reviews para cada overall	%
1	128149	6.4168
2	77129	3.8621
3	129093	6.4640
4	387303	19.3933
5	1275424	63.8639

3) Média de overall em todas as reviews

✓ A mostrar registos de 0 - 0 (1 total, A consulta demorou 2,9184 segundos.)

```
SELECT AVG(review.overall) as 'média de overall' FROM review;
```

☐ Perfil [[Editar em linha](#)] [[Edita](#)] [[Explicar SQL](#)] [[Criar código PHP](#)] [[Actualizar](#)]

☐ Mostrar tudo | Número de registos: 25 ▼ Filtrar registos:

+ Opções

média de overall

4.3043

4) Top 5 reviewers com mais reviews (com nulls)

✓ A mostrar registos de 0 - 4 (5 total, A consulta demorou 14,9984 segundos.)

```
SELECT reviewer_clean.reviewerName as 'nome do cliente', COUNT(*) AS 'número de reviews' FROM reviewer_clean, review WHERE reviewer_clean.reviewerID = review.reviewerID GROUP BY reviewer_clean.reviewerName ORDER BY 2 DESC LIMIT 5;
```

☐ Perfil [[Editar em linha](#)] [[Edita](#)] [[Explicar SQL](#)] [[Criar código PHP](#)] [[Actualizar](#)]

+ Opções

nome do cliente	número de reviews
Amazon Customer	28691
	5907
Pen Name	3649
Jennifer	2022
Chris	1911

5) Top 5 reviewers com mais reviews (sem nulls)

✓ A mostrar registos de 0 - 4 (5 total, A consulta demorou 14,8899 segundos.)

```
SELECT reviewer_clean.reviewerName as 'nome do cliente', COUNT(*) AS 'número de reviews' FROM reviewer_clean, review WHERE reviewer_clean.reviewerID = review.reviewerID AND reviewer_clean.reviewerName != '' GROUP BY reviewer_clean.reviewerName ORDER BY 2 DESC LIMIT 5;
```

☐ Perfil [[Editar em linha](#)] [[Edita](#)] [[Explicar SQL](#)] [[Criar código PHP](#)] [[Actualizar](#)]

+ Opções

nome do cliente	número de reviews
Amazon Customer	28691
Pen Name	3649
Jennifer	2022
Chris	1911
Kindle Customer	1860

6) Top 5 produtos com mais reviews

✓ A mostrar registos de 0 - 4 (5 total, A consulta demorou 3,0508 segundos.)

```
SELECT produto_clean.asin as 'id do produto', COUNT(*) as 'número de reviews' FROM produto_clean, review WHERE produto_clean.asin = review.asin GROUP BY produto_clean.asin ORDER BY 2 DESC LIMIT 5;
```

☐ Perfil [[Editar em linha](#)] [[Edita](#)] [[Explicar SQL](#)] [[Criar código PHP](#)] [[Actualizar](#)]

+ Opções

id do produto	número de reviews
B004S8F7QM	9695
8499000606	2604
B005JFNE8G	2036
B00DPK11ZA	1996
B004A8ZRBA	1838

7) Top 5 produtos com melhores reviews (só para os produtos que têm mais reviews que a média)

```
1 SELECT produto_clean.asin, AVG(review.overall) as 'media de overall', COUNT(*) as total
2 FROM produto_clean, review
3 WHERE review.asin = produto_clean.asin
4 GROUP BY produto_clean.asin
5 HAVING total >=
6     (SELECT AVG(tab1.total)
7      FROM
8      (SELECT COUNT(*) as total
9       FROM review, produto_clean
10        WHERE review.asin = produto_clean.asin
11         GROUP BY produto_clean.asin) as tab1)
12 ORDER BY 2 DESC, 3 DESC
13 LIMIT 5;
```

✓ A mostrar registos de 0 - 4 (5 total, A consulta demorou 11,5534 segundos.)

```
SELECT produto_clean.asin, AVG(review.overall) as 'media de overall', COUNT(*) as total FROM produto_clean, review WHERE review.asin = produto_clean.asin GROUP BY
produto_clean.asin HAVING total >= (SELECT AVG(tab1.total) FROM (SELECT COUNT(*) as total FROM review, produto_clean WHERE review.asin = produto_clean.asin GROUP BY
produto_clean.asin) as tab1) ORDER BY 2 DESC, 3 DESC LIMIT 5;
```

☐ Perfil [\[Editar em linha \]](#) [\[Edita \]](#) [\[Explicar SQL \]](#) [\[Criar código PHP \]](#) [\[Atualizar \]](#)

+ Opções

asin	media de overall	total
B0077Q0NP2	5.0000	42
B00D3Y18WO	5.0000	35
B0006GVA16	5.0000	30
B005QC8N7E	5.0000	29
B007VDI3RQ	5.0000	29

8) Top 5 produtos com piores reviews (só para os produtos que têm mais reviews que a média)

✓ A mostrar registos de 0 - 4 (5 total, A consulta demorou 10,2930 segundos.)

```
SELECT produto_clean.asin, AVG(review.overall) as 'media de overall', COUNT(*) as total FROM produto_clean, review WHERE review.asin = produto_clean.asin GROUP BY
produto_clean.asin HAVING total >= (SELECT AVG(tab1.total) FROM (SELECT COUNT(*) as total FROM review, produto_clean WHERE review.asin = produto_clean.asin GROUP BY
produto_clean.asin) as tab1) ORDER BY 2 ASC, 3 DESC LIMIT 5;
```

☐ Perfil [\[Editar em linha \]](#) [\[Edita \]](#) [\[Explicar SQL \]](#) [\[Criar código PHP \]](#) [\[Atualizar \]](#)

+ Opções

asin	media de overall	total
B003TSD1PG	1.0000	19
B0001W94VI	1.0000	18
B005U00BF0	1.0000	11
B0002MIDLY	1.0000	10
B00022F03S	1.0000	9

9) Média de likes e dislikes

✓ A mostrar registos de 0 - 0 (1 total, A consulta demorou 2,8751 segundos.)

```
SELECT AVG(review.likes) as 'media de likes por review', AVG(review.dislikes) as 'media de dislikes por review' FROM review;
```

☐ Perfil [\[Editar em linha \]](#) [\[Edita \]](#) [\[Explicar SQL \]](#) [\[Criar código PHP \]](#) [\[Atualizar \]](#)

☐ Mostrar tudo | Número de registos: 25 Filtrar registos:

+ Opções

media de likes por review	media de dislikes por review
1.4940	1.8694

10) Número máximo de likes e o seu produto

✓ A mostrar registos de 0 - 0 (1 total, A consulta demorou 5,5790 segundos.)

```
SELECT review.asin, review.likes as 'número máximo de likes' FROM review WHERE review.likes = (SELECT MAX(review.likes) FROM review);
```

☐ Perfil [Editar em linha] [Editar] [Explicar SQL] [Criar código PHP] [Actualizar]

☐ Mostrar tudo | Número de registos: 25 | Filtrar registos: Pesquisar esta tabela

+ Opções

	asin	número máximo de likes
<input type="checkbox"/>	B00BPMMQDG	6984

☐ Editar ☐ Copiar ☐ Apagar

11) Número máximo de dislikes e o seu produto

✓ A mostrar registos de 0 - 0 (1 total, A consulta demorou 5,5908 segundos.)

```
SELECT review.asin, review.dislikes as 'número máximo de dislikes' FROM review WHERE review.dislikes = (SELECT MAX(review.dislikes) FROM review);
```

☐ Perfil [Editar em linha] [Editar] [Explicar SQL] [Criar código PHP] [Actualizar]

☐ Mostrar tudo | Número de registos: 25 | Filtrar registos: Pesquisar esta tabela

+ Opções

	asin	número máximo de dislikes
<input type="checkbox"/>	B00BPMMQDG	7071

☐ Editar ☐ Copiar ☐ Apagar