

# Trabalho Final Data Warehouse

Bruno Marques, Érica Yoshiwara e  
Gustavo Jordão



# Tópicos

1. Caracterização dos Dados
2. Modelagem de um Data Warehouse Multidimensional
3. Ferramentas Utilizadas
4. Extração, Transformação e Carga (ETL)
5. Criação de visualizações para os dados (OLAP)



# Bases de Datos

# Movie-Ratings

Número de registros: 562

Número de atributos: 6

Extraída do Kaggle.

<https://www.kaggle.com/nagenderp/movie-ratings/data>

## Descrição dos dados:

Informações básicas sobre os filmes (título, gênero, orçamento, ano). Possui a avaliação do público e da crítica pelo site especializado Rotten Tomatoes.

## Observações:

A nota do Rotten Tomatoes indica a % de críticos que avaliaram o filme como regular ou mais (6-10).



Film	Genre	Rotten Tomatoes Ratings %	Audience Ratings %	Budget (million \$)	Year of release
(500) Days of Summer	Comedy	87	81	8	2009
10,000 B.C.	Adventure	9	44	105	2008
12 Rounds	Action	30	52	20	2009
127 Hours	Adventure	93	84	18	2010
17 Again	Comedy	55	70	20	2009
2012	Action	39	63	200	2009
27 Dresses	Comedy	40	71	30	2008
30 Days of Night	Horror	50	57	32	2007
30 Minutes or Less	Comedy	43	48	28	2011
50/50	Comedy	93	93	8	2011
88 Minutes	Drama	5	51	30	2007
A Dangerous Method	Drama	79	89	20	2011

Visualização da base de dados

# IMDB-Movie-Data

Número de registros: 1000

Número de atributos: 12

Extraída do Kaggle.

<https://www.kaggle.com/PromptCloudHQ/imdb-data/data>

## Descrição dos dados:

Apresenta informações mais completas sobre os filmes, como descrição, atores, diretor e duração.

## Observação:

Traz a avaliação de outro site especializado, o Metascore.



Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
51	Star Wars: Episode VII - The Force Awakens	Action,Adventure, Fantasy	Three decades(...).	J.J. Abrams	Daisy Ridley, (...)	2015	136	81	661608	93663	81
88	Avatar	Action,Adventure, Fantasy	A paraplegic (...).	James Cameron	Sam Worthington,(...)	2009	162	78	935408	76051	83
86	Jurassic World	Action,Adventure, Sci-Fi	A new(...)	Colin Trevorrow	Chris Pratt, (...)	2015	124	7	455169	65218	59
77	The Avengers	Action,Sci-Fi	Earth's might (...)	Joss Whedon	Robert (...)	2012	143	81	1045588	62328	69
55	The Dark Knight	Action,Crime,Dra ma	When the (...)	Christopher Nolan	Christian (...)	2008	152	9	1791916	53332	82

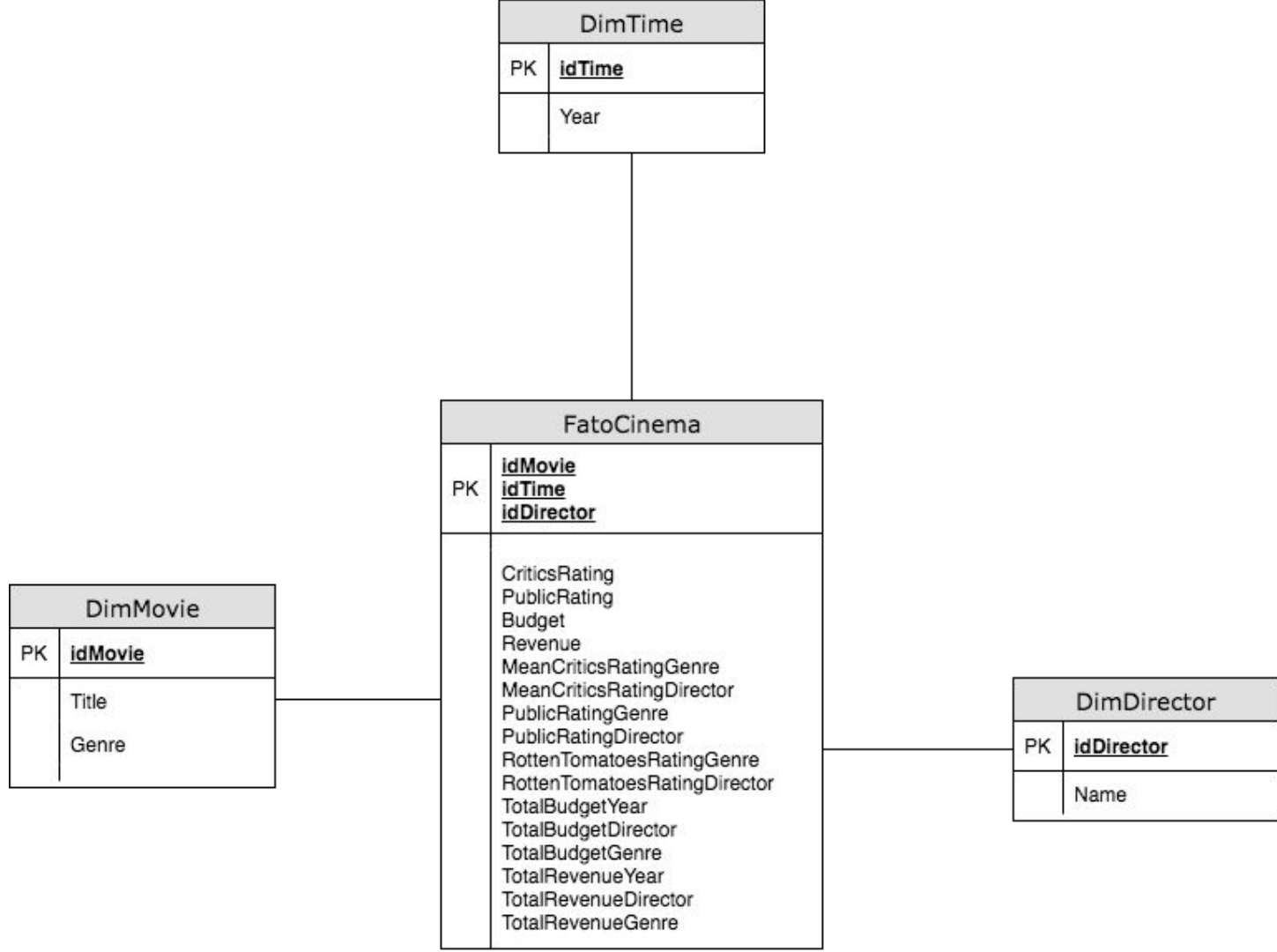
Visualização da base de dados



# Modelagem do Data Warehouse







# Ferramentas Utilizadas

Pentaho



PgAdmin (PostgreSQL)



# Extração, Transformação e Carga

View Design

Steps

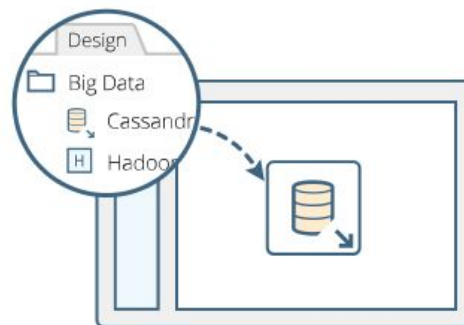
- Input
- Output
- Streaming
- Transform
- Utility
- Flow
- Scripting
- Pentaho Server
- Lookup
- Joins
- Data Warehouse
- Validation
- Statistics
- Big Data
- Agile
- Cryptography
- Palo
- Open ERP
- Job
- Mapping
- Bulk loading

Testes

Transformação 1



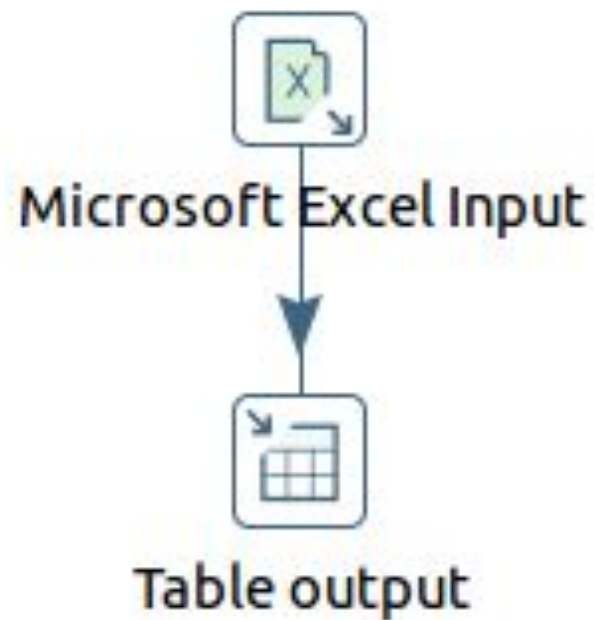
100%



## Drag & Drop a Step

Also try shift + double-click

# Extração



Nome do Step **Microsoft Excel Input**

Files Sheets Content Error Handling Fields Additional output fields

Spread sheet type (engine) Excel 97-2003 XLS (JXL)

File or directory

Regular Expression

Exclude Regular Expression

Selected files:

	File/Directory	Wildcard (RegExp)	Exclude
1	/home/erica/Documentos/CEFET/9º Período/DW/IMDB_excel.xls		

Accept filenames from previous steps

Accept filenames from previous step ☐

Step to read filenames from

Field in the input to use as filename

Show filename(s)...

Nome do Step

Connection  ▼ Edit... New... Wizard...

Target schema  ▼ Navega...

Target table  ▼ Navega...

Commit size  ▼

Truncate table ☐

Ignore insert errors ☐

Specify database fields ☒

Main options **Database fields**

Partition data over tables ☐

Partitioning field  ▼ ▼

Partition data per month ☐

Partition data per day ☐

Use batch update for inserts ☒

O nome da tabela está definido em uma coluna? ☐

Coluna que tem o nome da tabela:  ▼ ▼

Store the tablename field ☒

Return auto-generated key ☐

Name of auto-generated key field  ▼

Help OK Cancela SQL

```
1 SELECT * FROM public.imdb
2
```

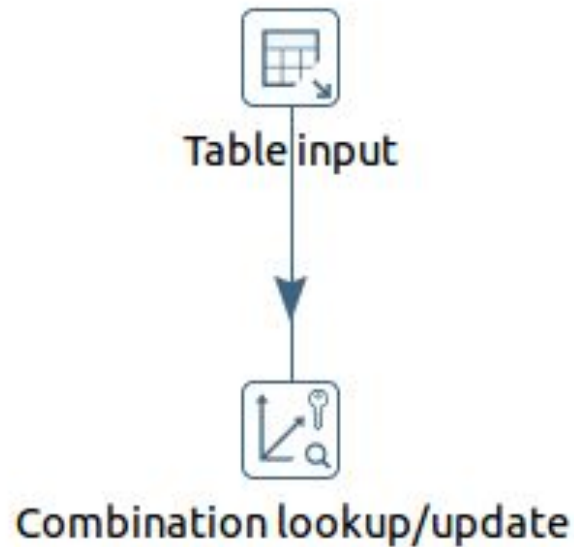
Data Output

[Explain](#)[Messages](#)[Query History](#)

	Rank bigint	title text	genre text	description text	director text	actors text	Year double precision	Runtime (Minutes) double precision	rating double precision	votes double precision
1	51	Star ...	Action,...	Three decade...	J.J. Abrams	Daisy Ri...	2015	136	81	661608
2	88	Avatar	Action,...	A paraplegic ...	James Ca...	Sam W...	2009	162	78	935408
3	86	Juras...	Action,...	A new theme ...	Colin Tre...	Chris Pr...	2015	124	7	455169
4	77	The ...	Action,...	Earth's mighti...	Joss Whe...	Robert ...	2012	143	81	1045588
5	55	The ...	Action,...	When the me...	Christoph...	Christia...	2008	152	9	1791916
6	13	Rog...	Action,...	The Rebel Alli...	Gareth E...	Felicity J...	2016	133	79	323118
7	120	Findi...	Animat...	The friendly b...	Andrew S...	Ellen D...	2016	97	74	157026
8	95	Aven...	Action,...	When Tony St...	Joss Whe...	Robert ...	2015	141	74	516895
9	125	The ...	Action,...	Eight years af...	Christoph...	Christia...	2012	164	85	1222645
10	579	The ...	Action,...	Katniss Everd...	Francis L...	Jennifer...	2013	146	76	525646
11	79	Pirat...	Action,...	Jack Sparrow ...	Gore Ver...	Johnny ...	2006	151	73	552027
12	689	Toy ...	Animat...	The toys are ...	Lee Unkri...	Tom Ha...	2010	103	83	586669
13	280	Iron ...	Action,...	When Tony St...	Shane Bla...	Robert ...	2013	130	72	591023
14	36	Capt...	Action,...	Political interf...	Anthony ...	Chris Ev...	2016	147	79	411656



# Transformação



Nome do Step

Connection



Edit...

New...

Wizard...

Get SQL select statement...

SQL

```
SELECT film, genre
  FROM movieratings
UNION
SELECT title, genre
  FROM imdb
```

Linha 1 Coluna 0

Enable lazy conversion ☐

Replace variables in script? ☐

Insert data from step



Executar para cada linha? ☐

Tamanho limite

Help

OK

Preview

Cancela

Nome do Step

Combination lookup/update

Connection

dw\_movies

Edit...

New...

Wizard...

Target schema

Navega...

Tabela de destino

DimMovies

Navega...

Confirma tamanho

100

Tamanho do Cache

9999

Pre-load the cache?

☐

Campos Chave (para verificar linha na tabela):

	Campo Dimens	Campo no fluxo
1	Title	film
2	genre	genre

Campo chave técnica

idMovie

Criação de chave técnica

☐

Usa tabela máxima + 1

☐

Usa sequência

☒

Usa o campo de auto incremento

Remove campos lookup?

☐

Usa hashcode?

☐

Campo Hashcode na tabela

Date of last update field (optional)

Help

OK

Cancela

Obtem Campos

SQL

```

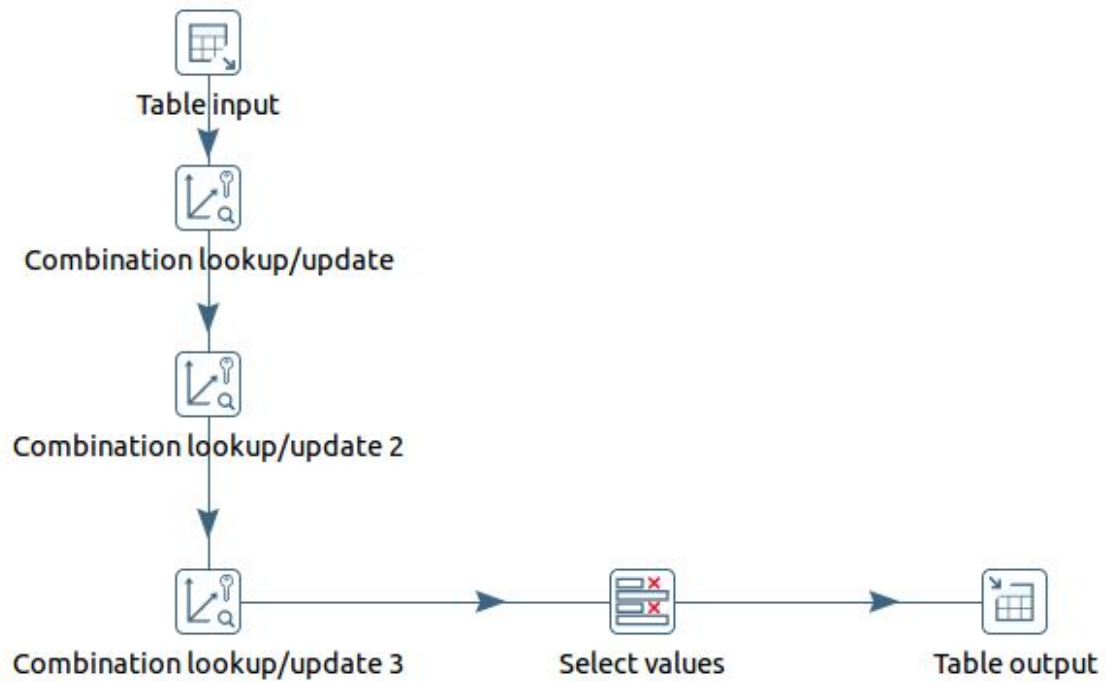
1 SELECT * FROM public.dimmovies
2

```

Data Output Explain Messages Query History

	idmovie bigint	title text	genre text
1807	1807	Leatherheads	Comedy
1808	1808	Collide	Action, Crime, Thriller
1809	1809	Southpaw	Drama, Sport
1810	1810	Jonah Hex	Action
1811	1811	The Green Inferno	Adventure, Horror
1812	1812	Tracktown	Drama, Sport
1813	1813	Despicable Me	Animation, Adventure, Comedy
1814	1814	Cowboys and Aliens	Action
1815	1815	London Has Fallen	Action, Crime, Drama
1816	1816	Queen of Katwe	Biography, Drama, Sport
1817	1817	The Fast and the Furious:...	Action, Crime, Thriller
1818	1818	Shutter Island	Mystery, Thriller
1819	1819	Toni Erdmann	Comedy, Drama
1820	1820	Teenage Mutant Ninja Tu...	Action, Adventure, Comedy

# Carga



Nome do Step

Connection

Target schema

Tabela de destino

Confirma tamanho  Tamanho do Cache

Pre-load the cache? ☐

Campos Chave (para verificar linha na tabela):

	Campo Dimens	Campo no fluxo
1	iddirector	iddirector

Campo chave técnica

Criação de chave técnica

- ☒ Usa tabela máxima + 1
- ☐ Usa sequência
- ☐ Usa o campo de auto incremento

Remove campos lookup? ☐

Usa hashcode? ☐

Campo Hashcode na tabela

Date of last update field (optional)

Step name

Select values

Select & Alter Remove Meta-data

Fields :

▼	Fieldname	Rename to	Length	Precision
1	idmovie			
2	iddirector			
3	idtime			
4	rating			
5	Revenue (Millions)			
6	Rotten Tomatoes Ratings %			
7	Audience Ratings %			
8	Budget (million \$)			

Get fields to select

Edit Mapping

```
1 SELECT * FROM public.fatocinema
2
```

Data Output Explain Messages Query History

	Idmovie bigint	Iddirector bigint	Idtime bigint	rating double precision	Revenue (Millions) double precision	Rotten Tomatoes Ratings % double precision	Audience Ratings % double precision	Budget (million \$) double precision
1	442	450	2	72	25639	69	69	150
2	269	200	2	62	8216	71	52	150
3	504	615	2	69	8005	67	75	20
4	1061	560	2	71	5368	21	82	30
5	560	195	2	81	22714	93	91	110
6	1045	316	2	81	7427	95	84	25
7	1154	549	2	61	7008	42	55	180
8	198	614	2	62	33653	61	54	258
9	683	221	2	55	1754	44	45	10
10	180	375	2	81	1835	82	90	15
11	449	482	2	74	2434	54	84	45
12	710	457	2	76	2503	82	86	53





# Visualizações OLAP



# Visualizações Esperadas

1. Evolução do orçamento dos filmes
2. Evolução da bilheteria dos filmes
3. Orçamento/Bilheteria/Avaliação por gênero
4. Orçamento/Bilheteria/Avaliação por diretor
5. Orçamento/Bilheteria/Avaliação por ano



# Cubo OLAP

Auxílio da ferramenta Mondrian ([mondrian.pentaho.com](http://mondrian.pentaho.com)).



1. Estabelece uma conexão com a base de dados.
2. Cria um cubo.
3. Define tabela, dimensões e métricas.
4. Publica o esquema para o BI Server.



Schema Workbench

File Edit View Options Windows Help

Schema - filmesSchema (SchemaFilme.xml)\*

Schema

- filmesCube
  - Table: fatocinema
  - tempo
    - default
      - year
  - Table: dimtime
  - diretor
    - default
      - director\_name
  - Table: dimdirector
  - filme
    - default
      - title
      - genre
    - Table: dimmovies
    - totalBudget
    - totalRevenue
    - avgAudienceRating
    - avgRottenRating

Measure for 'filmesCube' Cube

Attribute	Value
name	totalRevenue
description	
aggregator	sum
column	Revenue (...)
formatString	
datatype	Numeric
formatter	
caption	
visible	<input checked="" type="checkbox"/>

Database - trabalhofinal (PostgreSQL)

# BI Server



## User Console

User Name:  Password:  [Login](#)

[Login as an Evaluator >](#)

© 2005-2018 Pentaho Corporation. All rights reserved.

File

View

Tools

Help

Home ▾

admin ▾

Browse Files

Create New

Manage Data Sources

Documentation

Recents

You haven't opened anything recently. Browse your files.

Browse Files

Favorites

You haven't selected any favorites yet. Add some favorites.

Browse Files

Pentaho Business Analytics

Try Enterprise Edition

Learn

Contribute

Meet the Family

Upgrade to Enterprise Edition

The Pentaho Business Analytics Platform delivers an open, unified, end-to-end solution including data integration, visualization and consumption of data.

Either by providing ad-hoc tools to analyze and visualize data or by serving pre-created content, different content types including analysis, reports, dashboards, data mining and even community created plugins can be seamlessly connected.

Getting Started

Documentation

Forums

New user? Want to know how to get started? Browse through samples and check out our embedded documentation.

• Check samples →

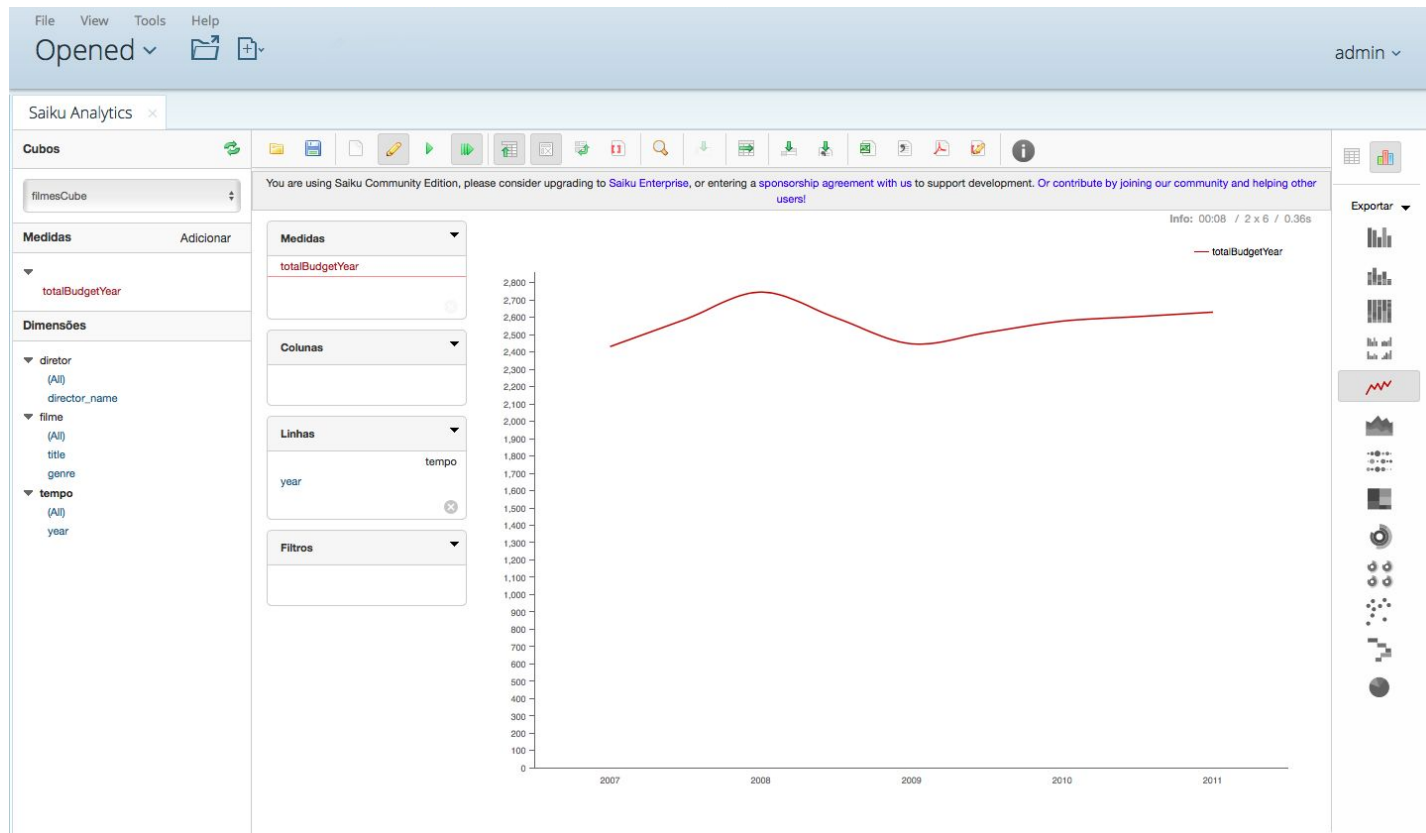
• Pentaho Information Map →

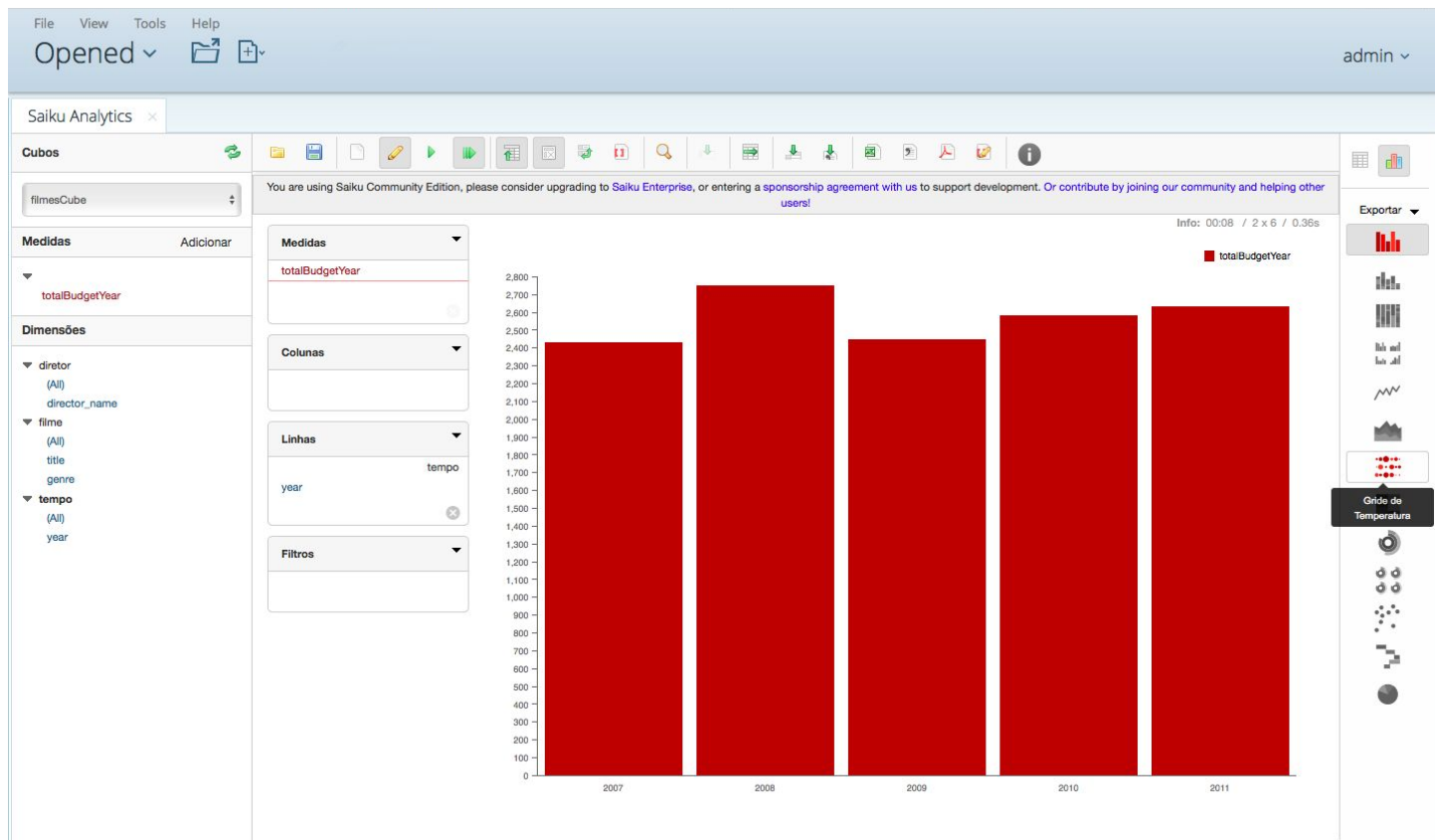
Both the *Pentaho InfoCenter* and the *Wiki* are extensive repositories of information. To learn more spend time navigating content through these important links.

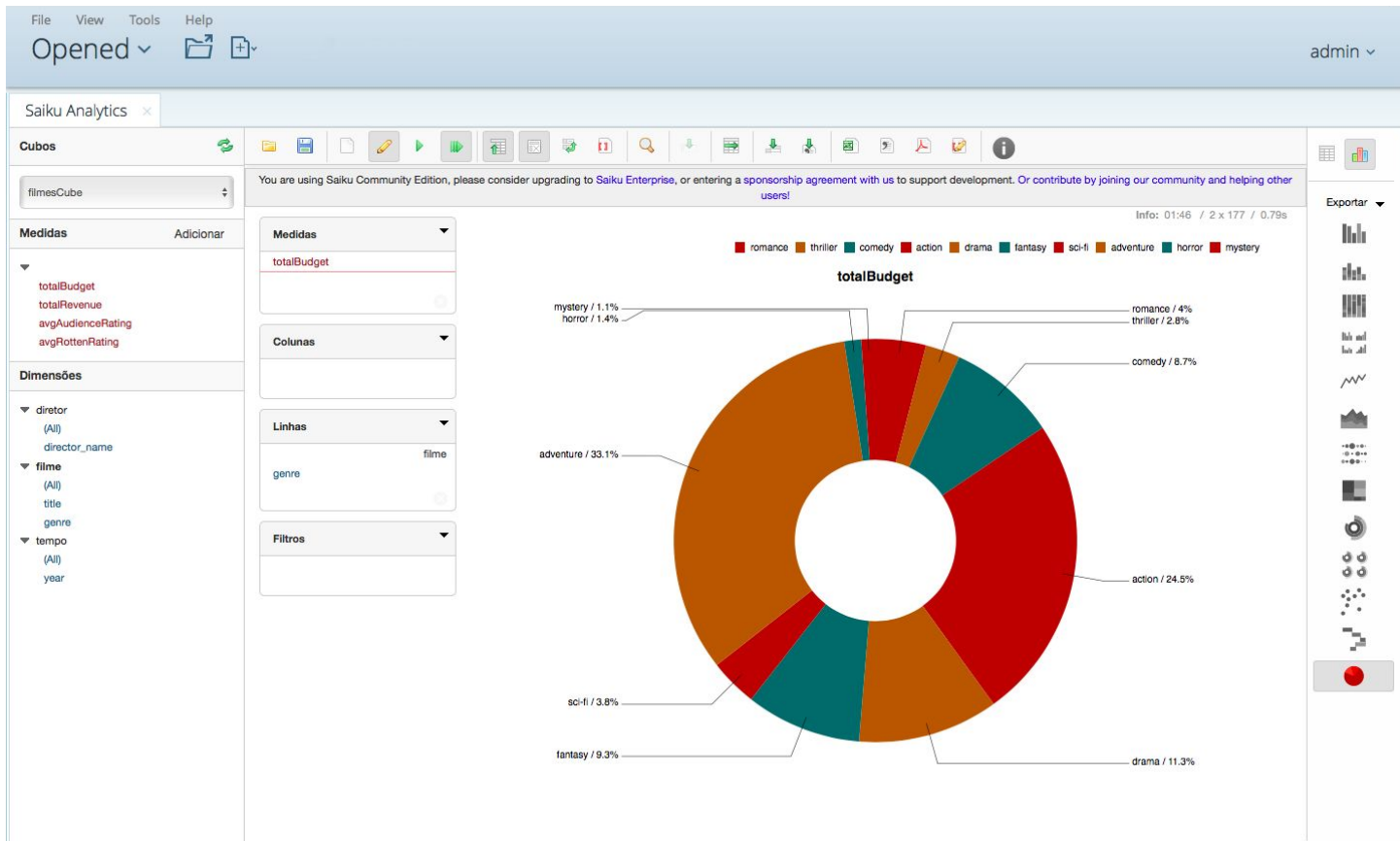
• Pentaho InfoCenter →

The Pentaho forums are an excellent resource not only to get help from experts but also to pass on your expertise to less experienced users.


• Pentaho Forums →









Saiku Analytics 

## Cubos

filmesCube

Adicionar

## Medidas

- totalBudget
- totalRevenue
- avgAudienceRating
- avgRottenRating

### Dimensões

- ▼ **director**
  - (All)
  - director\_name
- ▼ **filme**
  - (All)
  - title
  - genre
- ▼ **tempo**
  - (All)
  - year

### Medidas

totalRevenue
totalBudget

Colunas

### Linhas

genre

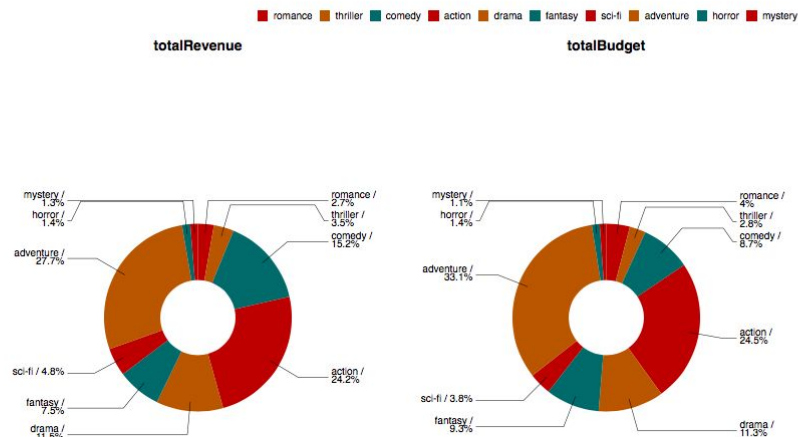
filme

## Filtros

You are using Saiku Community Edition, please consider upgrading to [Saiku Enterprise](#), or entering a [sponsorship agreement with us](#) to support development. Or contribute by joining our community and helping other users!

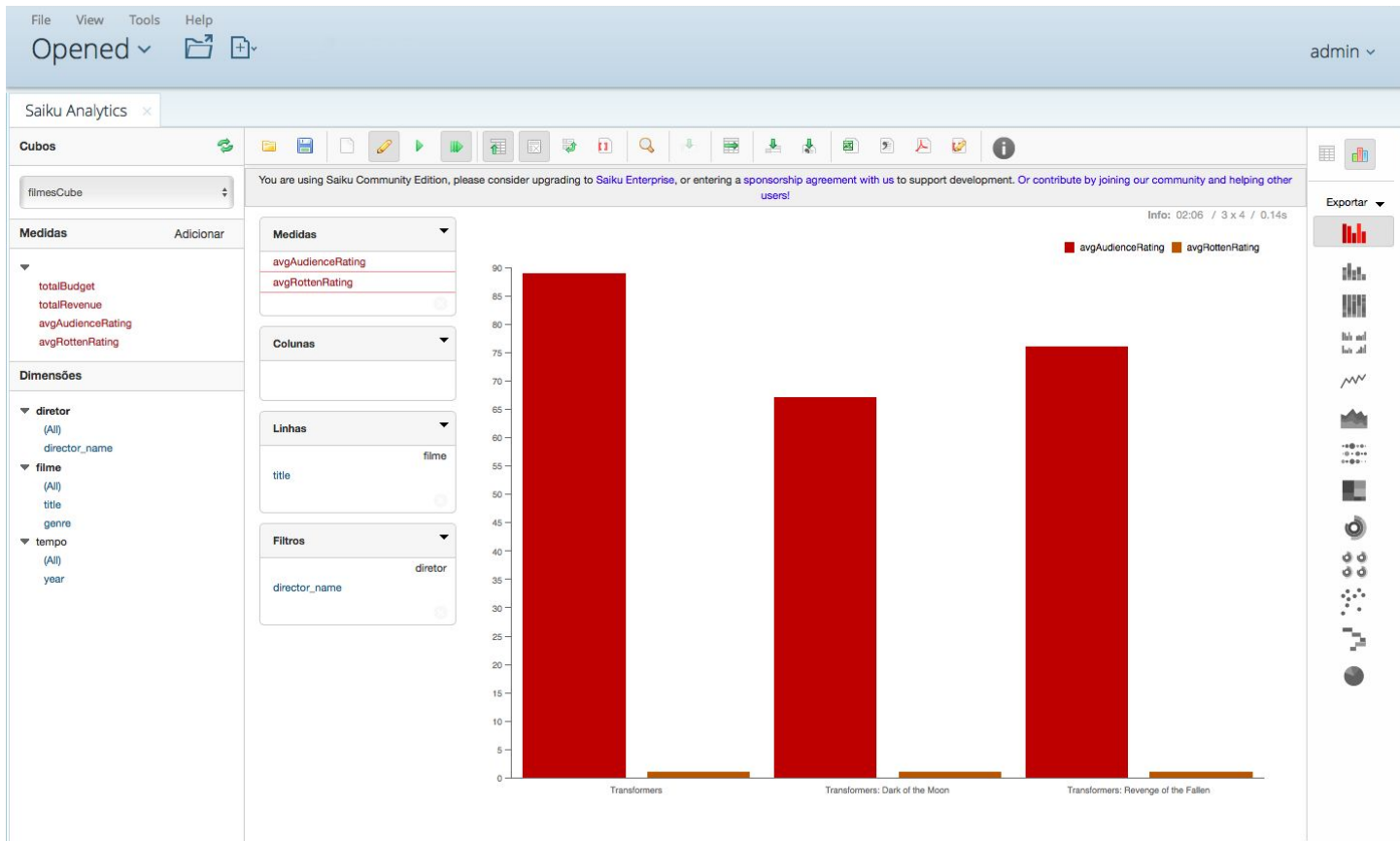
Info: 01:49 / 3 x 177 / 0.20s

totalRevenue

**totalBudget**

Exportar ▼





# Artefatos

Repositório: [https://github.com/brunomaciel/dw\\_movies](https://github.com/brunomaciel/dw_movies)





Obrigado (a) !

