Trabalho de sistema distribuídos

Alunos: Lucas Meneses, Rafael Vinicius e Tiago Tamaki

Relatório

O seguinte relatório de sistemas distribuídos do curso de ciência da computação da UFMS tem como objetivo analisar e resolver problemas criados para um dataset.

Para isso, configuramos o ambiente distribuído apache hadoop e utilizando na solução o apache pig, ferramenta que permite analisar dados por meio do hadoop.

Problemas:

1 - Qual foi o mês que teve mais episódios?

Descrição: Trazer os totais de episódios para cada mês correspondente.

Código:

```
eps = LOAD './got_imdb.csv' USING PigStorage(',') AS (season:int, episode:int, date:chararray, title:chararray,
dts = FOREACH eps GENERATE FLATTEN(STRSPLIT($2, ' ', 3)) AS (day:chararray,month:chararray,year:chararray);

months = GROUP dts by month;
--DUMP months;
result = FOREACH months GENERATE group as month, COUNT(dts) AS count;
DUMP result;
```

Saída:

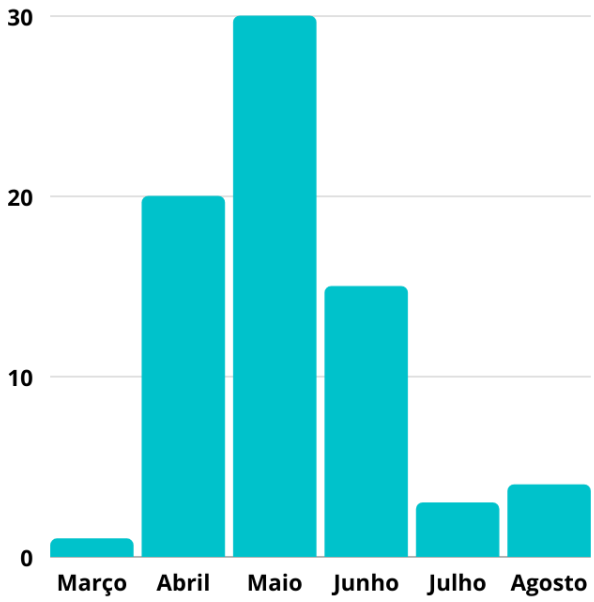
Março	1
Abril	20
Maio	30
Junho	15
Julho	3

Agosto 4

A primeira coluna corresponde ao mês.

A segunda coluna corresponde ao número de episódio lançados por mês

Colocando os dados no gráfico para melhor visualização.



Podemos observar que o mês com maior número de episódios lançado foi em maio com um total de 30.

2 - Qual foi a temporada que teve a melhor média de notas?

Descrição: Trazer todas as temporadas e calcular qual foi a média de cada uma.

Código:

```
A = LOAD '/user/game.csv' USING PigStorage(',') AS (season:int, episode:int, date:chararray, title:chararray
B = FOREACH A GENERATE $0, $4;
C = GROUP B BY season;
D = FOREACH C GENERATE group as season, SUM(B.Rating)/COUNT(B.Rating) as media;
F = ORDER D BY media DESC;
STORE F INTO 'file:///home/rafael/SubProblema' -- MUDAR PARA O SEU SISTEMA
```

Saída:

4 9.310000038146972

1 9.133333418104383

7 9.099999972752162

3 9.099999957614475

6 9.05999984741211

2 8.960000038146973

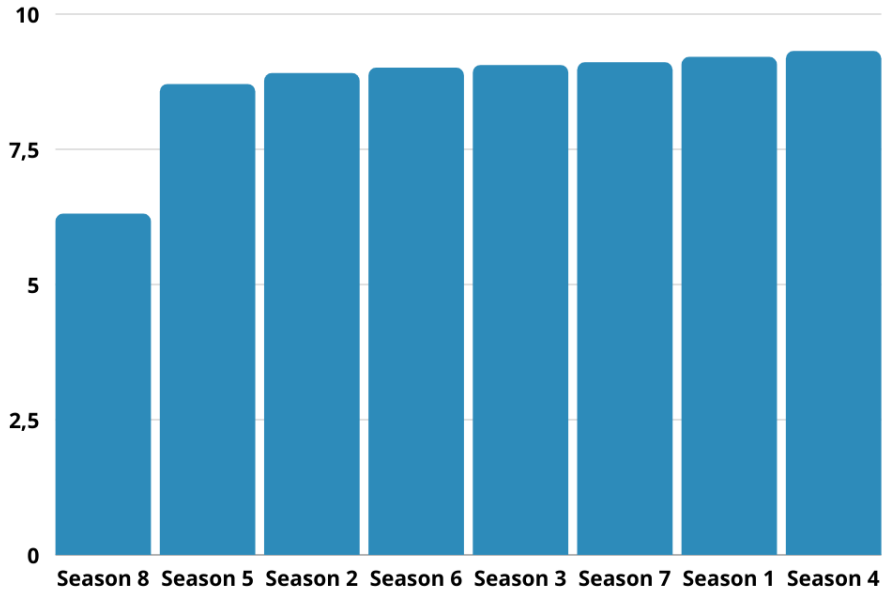
5 8.922222243414986

8 6.333333412806193

A primeira coluna representa a temporada.

A segunda coluna representa a média das notas de cada temporada

Arredondando os números para melhor visualização e colocando em um gráfico.



Com o gráfico fica visualmente melhor para saber quem está na frente.

Podemos também fazer uma pesquisa para saber quais foram os episódios e suas respectivas temporada com a melhor média, com algumas alterações no código.

Código:

```
A = LOAD '/user/game.csv' USING PigStorage(',') AS (season:int, episode:int, date:chararray, title:charar
B = FOREACH A GENERATE $0, $1, $4;
D = GROUP B ALL;
E = FOREACH D GENERATE MAX(B.Rating) as rating;
F = FILTER B BY Rating == (float)E.rating;
STORE F INTO 'file:///home/rafael/P2'; --EDITAR ESSE CAMPO PARA O SEU SISTEMA
```

Saída:

3 9 9.9

5 8 9.9

6 9 9.9

6 10 9.9

A primeira coluna representa o número da temporada.

A segunda coluna representa o número de episódio correspondente a temporada.

A terceira coluna representa a média.

O resultado da saída mostra que não existe nenhum episódio com uma nota acima de 9.9, existem apenas 4 episódios tendo a melhor nota e 2 deles são da temporada 6.

3 - Qual foi a temporada que teve maior número de votos?

Descrição: Trazer as temporadas e o seu episódio com o maior número de votos.

Código:

```
A = LOAD '/user/game.csv' USING PigStorage(',') AS (season:int, episode:int, date:chararray, title:chararray
B = FOREACH A GENERATE $0, $1, $5;
C = GROUB B BY season;
D = FOREACH C GENERATE group as season, COUNT(B.episode) as episode, SUM(B.count) as count;
F = ORDER D BY count ASC;
STORE F INTO 'file:///home/rafael/P4' --EDITAR ESSE CAMPO PARA O SEU SISTEMA
```

Saída:

2 10 272548

1 10 292663

3 10 303286

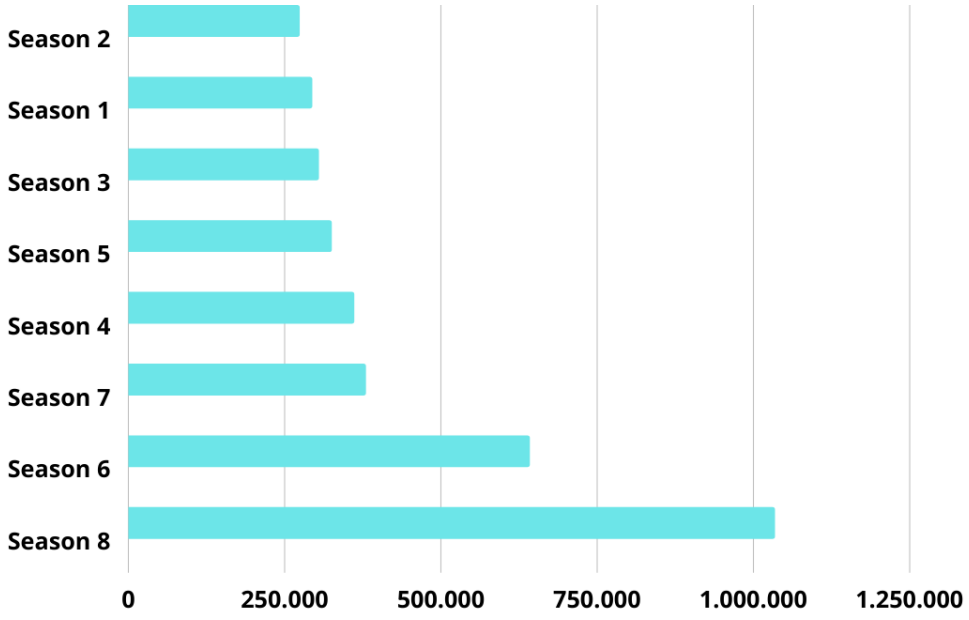
5	10	323871
4	10	359867
7	7	378528
6	10	640744
8	6	1032901

A primeira coluna representa a temporada.

A segunda coluna representa a quantidade de episódios por temporada.

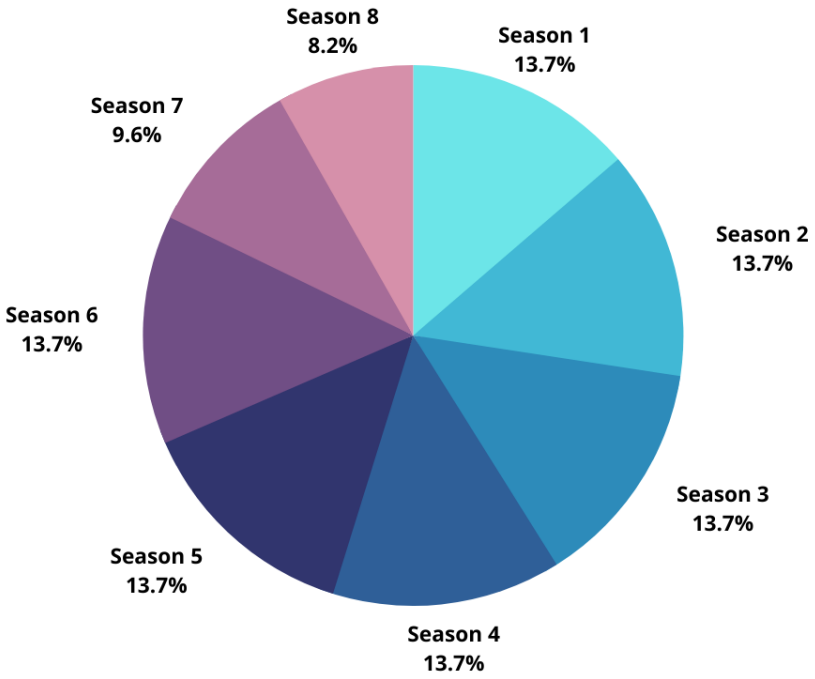
A terceira coluna representa a quantidade total de votos por temporada.

Fazendo uma relação entre a temporada e a quantidade de votos e colocando-os em um gráfico.

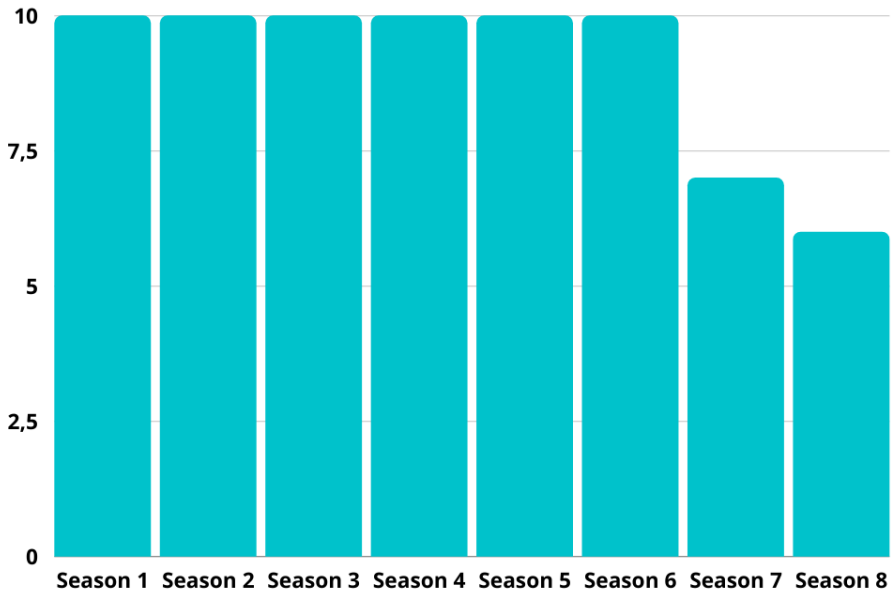


De acordo com o gráfico, mostra que a temporada que tem mais votos é a 8.

Podemos também saber qual foi a porcentagem de participação de cada temporada na série, e qual delas teve o menor número de episódios, relacionando a temporada com os episódios e colocando no gráfico.



Com esse gráfico podemos ver que a porcentagem de exibição de cada temporada.



Olhando o esse gráfico temos que a temporada 8 foi a que teve menos episódios.

Modificando o código conseguimos saber qual é a quantidade total de episódios que a série teve ao longo da sua exibição.

Código:

```
A = LOAD '/user/game.csv' USING PigStorage(',') AS (season:int, episode:int, date:chararray, title:chararray
B = FOREACH A GENERATE $0, $1, $5;
C = GROUP B ALL;
D = FOREACH C GENERATE COUNT(B.episode);
STORE D INTO 'File:///home/rafael/Pergunta7'; --MUDAR PARA O SEU SISTEMA
```

Saída: 73 episódios no total.

Podemos ir um pouco além na nossa pesquisa e também saber qual foi a temporada e o episódio com mais votos, mudando um pouco o código.

Código:

```
A = LOAD '/user/game.csv' USING PigStorage(',') AS (season:int, episode:int, date:chararray, title:chararray,
B = FOREACH A GENERATE $0, $1, $5;
C = GROUP B ALL;
D = FOREACH C GENERATE MAX(B.count) as count;
F = FILTER B BY count == (int)D.count;
STORE F INTO 'File:///home/rafael/Pergunta3';
```

Saída:

8 6 232767

O resultado da saída mostra que a temporada 8, o episódio 6 foi o mais votado com um total de 233.767 mil votos.

4– Quais foram os melhores meses da série em votos?

Descrição: Trazer os votos de todos os messes.

Código:

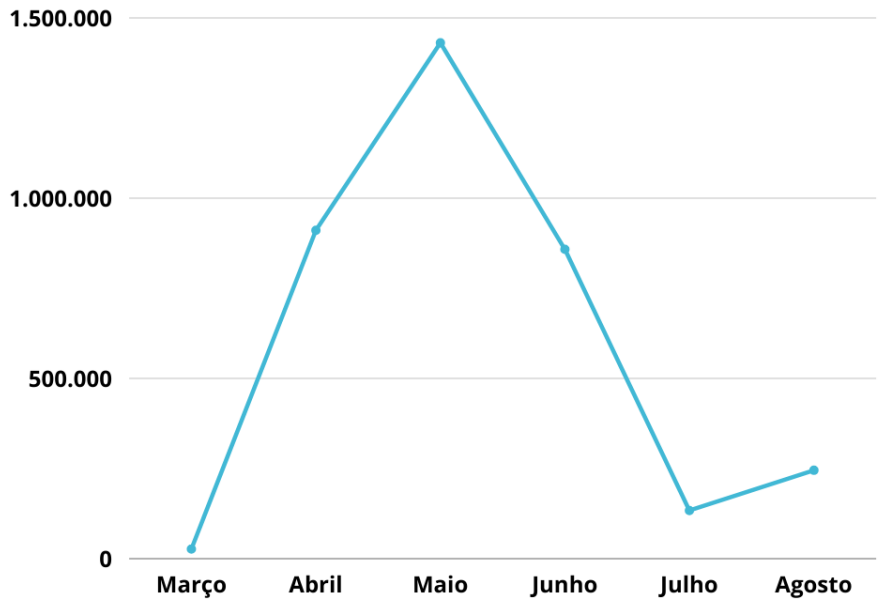
```
eps = LOAD './got_imdb.csv' USING PigStorage(',') AS (season:int, episode:int, date:chararray, title:chararray, Ra
dts = FOREACH eps GENERATE FLATTEN(STRSPLIT($2, ' ', 3)) AS (day:chararray,month:chararray,year:chararray), count;

months = GROUP dts by month;
result = FOREACH months GENERATE group as month, SUM(dts.count);
DUMP result;
```

Saída:

Março	26509
Abril	910717
Maio	1430913
Junho	857741
Julho	133522
Agosto	245006

Colocando em um gráfico para melhor visualização dos dados.



Conseguimos extrair que maio foi o melhor mês em votos, para a série ao longo dos anos de sua exibição.

Bibliografia

Dataset: https://www.kaggle.com/abhijithchandradas/game-of-thrones-imdb-dataset

Nosso repositório do GitHub: https://github.com/LucasGMeneses/got-pigHadoop