iImpacto dos Lançamentos de Filmes e Séries nas Quotas de Mercado

ISCTE-Instituto de Lisboa

Bruno Santos nº 125712

Eduardo Bagulho nº

Gonçalo Rosa nº 125684

# Introdução

A aquisição de dados sempre gerou um problema para a sociedade, o seu armazenamento. Nos dias de hoje qualquer pessoa pode comprar dezenas de centenas de gigas de armazenamento. Mesmo assim a quantidade de dados existente atualmente que mesmo as melhores máquinas teriam dificuldades em processar tais dados, a este fenómeno de dados chamamos de BIG DATA, a sua análise não é suportada apenas numa só máquina. Para tal dividimos a carga e os dados por diversas máquinas onde iram armazenar e analisar sempre que necessário, este conceito é chamado de sistemas distribuídos, uma rede de máquinas interligada entre si para que possamos questioná-la ou armazenar nela, uma quantidade de dados que não seria concebível numa única máquina.

Neste trabalho não iremos ter uma quantidade exorbitante de dados, mas pretendemos demonstrar como os sistemas distribuídos armazenam dados e os disponibilizam aos seus utilizadores, quando questionados.

Iremos ingerir no sistema datasets com informações relativas as quotas de mercado e aos lançamentos de filmes/series nas plataformas de entretenimento, por fim iremos realizar um relatório com as informações previamente ingeridas. Também teremos uma breve descrição das camadas (HDFS, YARN, MapReduce, HBase, Phoenix, Hive, Hue) e a sua parte no funcionamento do sistema em geral.

# Sistema distribuído e suas camadas

Um sistema de base de dados distribuído é composto por um array de base de dados interligadas entre si através de uma rede. Não podemos considerar um sistema de base de dados distribuída caso não exista uma rede que a interligue.  
O objetivo destes sistemas é aumentar a capacidade de processamento e de armazenamento para o um nível acima daquilo que uma só máquina alcançaria. Tendo diversos desafios associados, como a redundância dos dados, sua integridade e a forma como estes sistemas partilham informação quando necessário responder a queries onde a informação não se encontra apenas num site.

O beneficio destes sistemas de base de dados é que proporcionam capacidades de armazenamento nunca antes vistos, como também um processamento de dados. Diversas

## HDFS

## YARN

## MapReduce

## HBase

## Phoenix

## Hive

## Hue

# Datasets

Os dados que utilizamos no nosso trabalho foram obtidos na plataforma kaggle, encontramos dois datasets um com a informação relativa aos filmes e series lançadas na plataforma Netflix (por o intervalo de datas do data set ) e outro com quotas diárias no mercado de ações. O objetivo será observar se o lançamento na plataforma impacta de alguma forma as ações da empresa.

## Shares

O dataset Shares contem toda informação relativa à ação da Netflix desde (datas), juntamente com diversos indicadores que possibilitam uma análise mais aprofundada do comportamento da ação. Para o nosso relatório final e para ser ingerido pelo nosso sistema removemos algumas colunas, sendo as seguintes colunas as informações que iremos utilizar.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Date | Open | Close | Volume | sma\_50 | sma\_100 | ema\_50 | ema\_100 | Next day  close |
| Data do mercado | Preço de abertura do dia | Preço de fecho do dia | Quantidade de ações vendidas | *Simple moving*  *Average* de 50 dias | *Simple moving*  *Average* de 100 dias | *Exponential moving*  *Average* de 50 dias | *Exponential moving*  *Average* de 100 dias | Valor do fecho do dia seguinte |

Visto que pretendemos avaliar a performance da quota e como esta se comporta ao longo do tempo, foi necessário utilizar a variável “*Next day close*” que tem o valor do fecho da ação no dia seguinte, juntamente com a variável *close* que possui o valor do fecho do dia ao aplicar a fórmula 1 é possível retirar o crescimento da ação relativamente ao dia seguinte (**Growth)**.

## Netflix movie/series releases

Neste dataset iremos encontrar informações relativas aos filmes/series da Netflix, iremos usar este dataset para realizar perguntas sobre a evolução da quota de mercado relativamente aos filmes/series que são lançados na plataforma.

Os campos que utilizamos na nossa base de dados serão os seguintes:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Type | Title | Country | Date\_added | Release\_year | Durantion | Genre |
| Tipo de filme | Título  Da peça | País | Data de adicionamento à plataforma | Ano de lançamento do filme | ???? | Todos os géneros que a peça foi listada a ter |

Realizamos algumas transformações neste dataset para que fosse possível criar uma relação na base de dados com os diversos dias de lançamento na plataforma com a data de mercado, para tal alteramos a estrutura da Date\_added para que coincidisse com coluna Date de shares.  
Ao analisarmos o dataset da Netflix, descobrimos que podíamos também criar uma tabela de referência com todos os tipos de géneros de peças existentes, possibilitando uma análise futura do mercado pelo género de peça em si.

Todas as alterações que foram mencionadas em ambos datasets foram realizadas em python com ajuda do pandas, extraindo do csv original, aplicamos as transformações e criamos 3 ficheiros csv contendo os dados que pretendemos ingerir no nosso sistema distribuído.   
  
Retirando os seguintes ficheiros, após corremos o nosso código “Netflix\_Clean”, “moviesAltered”, “genres”, “sharesAltered”.   
  
Estando todas as transformações necessárias nos nossos dados podemos então passar à ingestão dos dados para o nosso sistema distribuído pelo HUE.

# Inserção de dados e analise

O passo seguinte terá com objetivo colocarmos todos os dados até agora trabalhados no nosso sistema de base de dados distribuído, podemos executar esta tarefa com ajuda da interface gráfica HUE, que nos permite comunicar com a linguagem SQL com as camadas mais abaixo no nosso sistema. Para tal é necessário seguir os seguintes passos.

1. Iniciar o Docker com o comando “docker-compose up -d”, que irá iniciar a imagem que se encontra no nossa pasta que previamente foi clonada de um repositório git.  
   Também podíamos através do Docker studio iniciar a mesma imagem.

Uma imagem com texto, captura de ecrã, software, Software de multimédia

Descrição gerada automaticamente

Figura 1- ligar o docker

1. Com a imagem a correr, podemos então aceder ao seguinte url [*http://localhost:8888/*](http://localhost:8888/)para que possamos interagir com a interface gráfica. Relembrar que a primeira vez será necessário associar uma palavra-passe. Consequentemente sempre que acedam ao HUE será necessário utilizar a mesma.
2. Uma imagem com texto, software, computador, Sistema operativo

   Descrição gerada automaticamenteCriar uma base de dados com o nome project\_Netflix.

Figura 2 - Criar base de dados

Uma imagem com captura de ecrã, texto, file

Descrição gerada automaticamenteCom a base de dados criada, ao aceder podemos ver que esta se encontra vazia

Figura 3 – base de dados

1. Uma imagem com texto, captura de ecrã, software, número

   Descrição gerada automaticamenteCriação das tabelas com os ficheiros com os dados trabalhados

Importamos os ficheiros csv com os dados trabalhados e definimos os tipos de dados que os campos terão associados dentro da base de dados, este passo é essencial para garantir uma boa gestão de armazenamento e de performance.

1. Repetimos o processo para todos os ficheiro de dados que pretendemos inserir na nossa base de dados.