INSTITUTO UNIVERSITÁRIO DE LISBOA

Análise de Dados em Big Data: Relação entre o YouTube e o stock market de empresas

Big Data

Docente: Paulo Vieira

David Isaac, Nº 120064

Ricardo Pereira, Nº 120052

Janeiro 2025

Índice

Resumo

Este relatório explora a relação entre as visualizações de vídeos no YouTube e o comportamento do mercado de ações de empresas específicas, analisando como a exposição digital pode influenciar o crescimento das organizações e o valor de suas ações. A pesquisa foi realizada através da análise de dados de 14 datasets provenientes de vídeos do YouTube e dos preços de ações de empresas como Sony, Nvidia, Microsoft, Dell, IBM e Intel. A metodologia envolveu a aplicação de processos de ELT (Extract, Load, Transform) para tratar e combinar os dados, seguido por análises de correlação entre as visualizações e os movimentos do mercado de ações.

Os resultados mostram que empresas como a Sony têm uma correlação significativa entre as visualizações dos seus vídeos e o valor das suas ações, especialmente em momentos estratégicos como lançamentos de produtos. Por outro lado, empresas como a Dell, Nvidia e Intel apresentaram uma relação mais fraca, sugerindo que outros fatores externos influenciam mais diretamente o desempenho das ações. O trabalho também destaca a importância de dados consistentes e a implementação eficaz de uma arquitetura técnica com PySpark, MongoDB e Streamlit para processar e visualizar grandes volumes de dados. Conclui-se que a análise de plataformas digitais pode ser uma ferramenta útil para prever tendências de mercado, embora fatores externos também desempenhem um papel significativo.

Introdução

Desde a estabilização do comércio internacional a nível mundial que as empresas têm tido um crescimento exponencial, impulsionado por diversos fatores, incluindo a inovação tecnológica e a globalização. Nos últimos anos, o crescimento de plataformas de divulgação e compartilhamento de vídeos têm um papel significativo na promoção de marcas, produtos e serviços, influenciando a percepção dos consumidores e, consequentemente, o desempenho financeiro das empresas.

Este projeto tem como objetivo analisar o impacto das visualizações de vídeos no *YouTube* sobre empresas específicas, avaliando como a exposição digital pode influenciar o crescimento das organizações e o valor das suas ações no mercado bolsista. Através de uma abordagem que combina análise de dados, economia e marketing digital, pretende-se compreender de que forma métricas como visualizações, interações e popularidade de conteúdo podem ser indicadores de tendências de investimento e valorização empresarial.

Metodologia

Datasets

Para a realização deste trabalho, foram utilizados 14 *datasets* que se dividem em dois grupos principais: **os datasets de vídeos do *YouTube*** e **datasets de stocks de empresas.**

Datasets de vídeos do *YouTube*

Os dados destes *datasets* são provenientes de oito diferentes países, dos quais: **Brasil**, **Estados Unidos** (US), **Canadá**, **França**, **Inglaterra**, **Índia**, **Coreia do Sul** e **México**.

Estes dados incluem título do vídeo (string), titulo do canal que publicou o vídeo (string), data de publicação (string), tags (string), número de visualizações (integer), número de *likes* e *dislikes* (integer), descrição (string) e número de comentários (integer).

Datasets de stocks de empresas

Os *datasets* sobre os *stocks* de empresas fornecem dados de seis organizações distintas: **Sony**, **Nvidia**, **Microsoft**, **Dell**, **IBM**, **Intel**.

Estes dados incluem data (string), preço de abertura (float), valor máximo atingido durante o período em que o stock esteve aberto (float), valor mínimo atingido durante o período em que o stock esteve aberto (float), valor aquando do fecho do stock (float), valor ajustado de fecho do stock (float), volume (integer).

Preparação dos Dados

Os dados passaram por um processo de **ELT (Extract, Load, Transform)**, com cada grupo de datasets submetido a um conjunto específico de processos para atingir o estado pretendido.

Abordagem ELT

A abordagem **ELT** demonstra-se mais vantajosa em relação à ETL (Extract, Transform, Load) uma vez que permite uma maior flexibilidade no tratamento dos dados conforme o objetivo dos mesmos, para além disto, como os dados de múltiplos países e empresas foram combinados e processados em um único ambiente, ELT facilita o manuseamento de grandes volumes de dados. O facto de mantermos os dados no seu estado bruto garante que informações importantes não sejam perdidas durante transformações iniciais, possibilitando reprocessamento ou análises diferentes no futuro.

Transformações

Grupo de Stocks

No grupo de datasets relacionados com stocks, foram realizadas as seguintes etapas:

1. Verificação de Dados: foi analisada a presença de valores nulos e duplicados eliminando ou tratando estes dados para garantir a integridade dos dados.
2. Validação e Correção de Tipos: os tipos de dados foram examinados para verificar se coincidiam com os formatos esperados. Por exemplo, colunas que representavam datas mas estavam classificadas como strings foram convertidas para o tipo date.
3. Combinação de Datasets: após a limpeza e validação dos dados, os diferentes datasets deste grupo foram combinados numa única estrutura. Durante este processo, foi adicionada uma nova coluna que identifica a organização a que cada conjunto de dados faz referência.

Grupo de Vídeos do YouTube

No grupo de datasets relacionados com vídeos do YouTube, o processo incluiu:

1. Filtragem de Dados: os dados foram filtrados com base no título do vídeo e no nome do canal em cada região, mantendo apenas aqueles que mencionam diretamente as empresas alvo. Esta etapa foi essencial para garantir que os dados refletiam a relação com as empresas analisadas.
2. Combinação de Dados: os dados filtrados de diferentes países foram combinados, sendo também adicionada uma nova coluna que identifica a organização relacionada com cada vídeo.

3. Verificação e Tratamento de Dados: Após a combinação, foram verificadas as quantidades de valores nulos e duplicados, e estes casos foram tratados adequadamente.
4. Validação e Correção de Tipos: assim como no grupo de stocks, os tipos de dados foram analisados e corrigidos para garantir a consistência, ajustando colunas a tipos apropriados.

Análise de Resultados

Neste capítulo busca-se compreender o impacto digital dos vídeos do Youtube no comportamento do mercado financeiro, a partir da identificação de tendências nos dados.

Tendo os dados já transformados e carregados na base de dados do Mongo, procedeu-se à interpretação e análise dos dados através de gráficos feitos a partir da Biblioteca "matplotlib". Inicialmente, o objetivo foi criar diversos gráficos lineares que ilustram o comportamento dos stocks ao longo dos dias, considerando as views de todos os vídeos no Youtube relacionados às empresas que estão a ser analisadas.

Entre as colunas disponíveis nos dados, a métrica que apresenta a maior possibilidade de comparação, capaz de ajudar a explicar a subida ou descida dos stocks, são as views. Até porque as demais colunas não apresentam uma correlação evidente ou significativa para justificar determinadas variações. Com isto em mente, formulamos vários gráficos com focos nessas duas variáveis.

Views por data em cada empresa

Neste gráfico o foco principal é nas views diárias de cada empresa em vídeos do YouTube. Por culpa da grande quantidade de vídeos associados à Sony, as restantes empresas aparecem com menor significado no gráfico. Com isto, é possível perceber que a Sony tem um foco de investimento maior em publicidade (seja diretamente ou por meio de terceiros) nas redes sociais, neste caso, como analisado no YouTube do que em comparação com as outras empresas que demonstram a sua utilização do YouTube de forma mais pontual, seja para compartilhar informações sobre avanços internos e novos produtos, seja devido à menor quantidade de vídeos de terceiros que mencionam essas empresas.

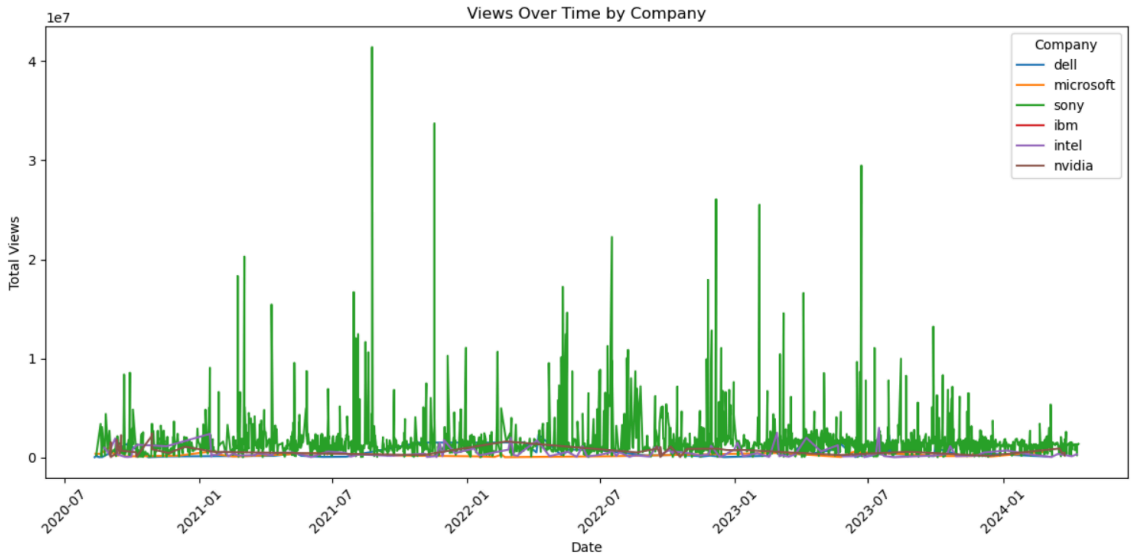


Fig. 1 - Views por data em cada empresa

Stocks por data em cada empresa

Neste gráfico é analisado o comportamento dos stocks de cada empresa ao longo dos anos. Em primeira análise, é possível reparar que algumas empresas possuem dados disponíveis apenas para períodos específicos. Por exemplo, os registros para a NVIDIA começam só no ano de 2000. Também consta que empresas como a NVIDIA e a Microsoft apresentam preços de stocks significativamente elevados em comparação com as outras empresas que demonstram um comportamento mais estável, sem grandes picos de valorização em seus preços de stocks.

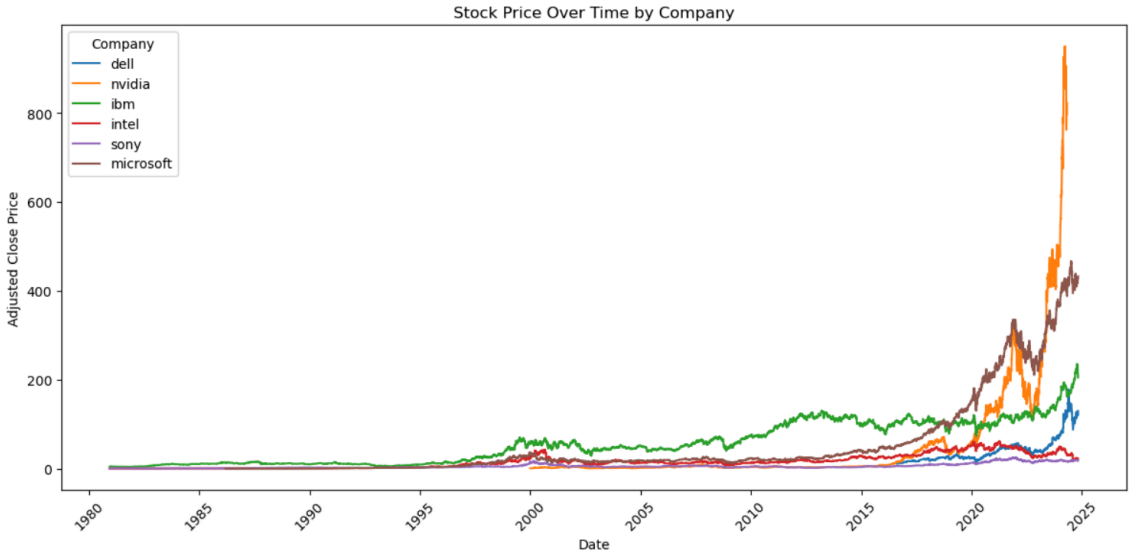


Fig. 2 - Stocks por data em cada empresa

Posteriormente, para facilitar a análise em um intervalo onde todas as empresas possuem dados disponíveis simultaneamente, foi realizado um estudo para identificar-se as datas em comum. Assim, foi gerado um gráfico que, a partir do intervalo de tempo, permite observar com clareza o período em que os dados estão disponíveis para todas as empresas.

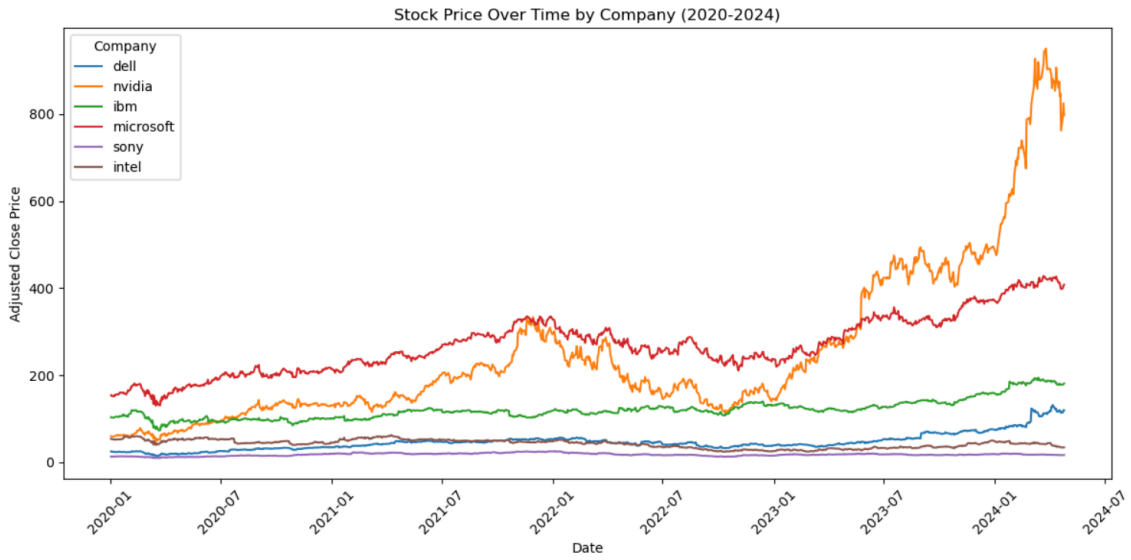


Fig. 3 - Stocks por data em cada empresa (2020-2024)

Comparação entre as Views e os Stocks de cada empresa

Fig. 4 - Quantidade de vídeos por empresa

Como demonstrado na tabela, a Sony é a empresa com maior quantidade de vídeos em seu nome. O que permite uma melhor análise na comparação entre os Stocks e as Views.

Embora outras empresas apresentem uma quantidade menor de vídeos, e portanto dados mais limitados, também incluiremos análises secundárias para elas, visando fornecer uma perspectiva comparativa. Dessa forma, nosso foco principal será a Sony, mas sem deixar de lado a contextualização com as demais empresas (sem contar com a IBM).

company	total_videos
microsoft	48
nvidia	48
dell	55
intel	122
sony	2009
ibm	1

Sony

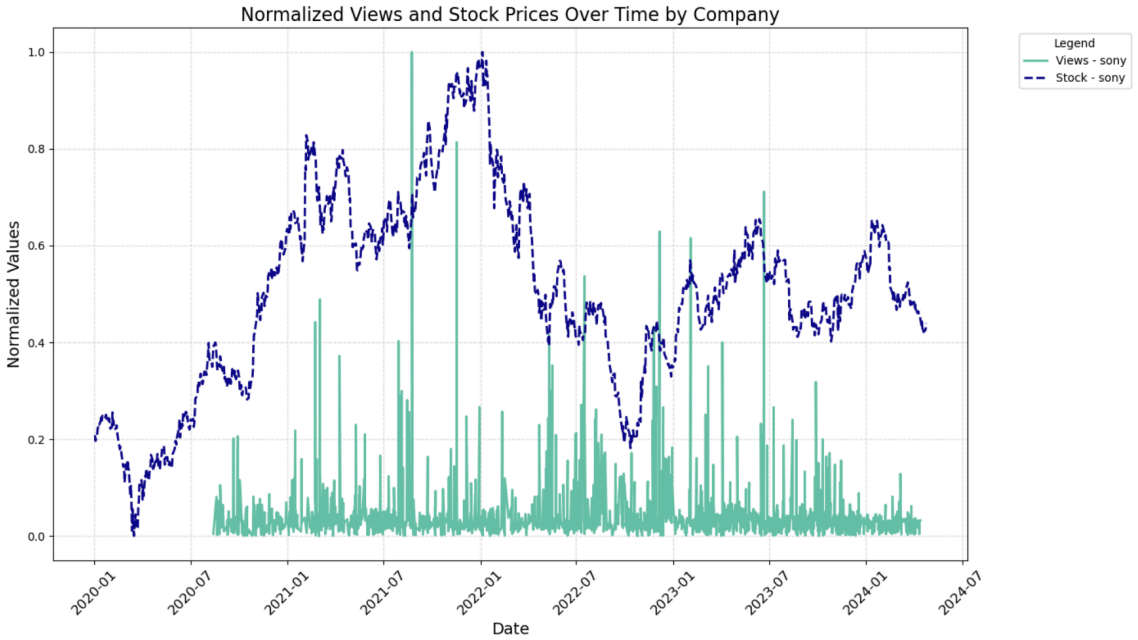


Fig. 5 - Merge entre as Views e os Stocks da Sony (2020-2024)

Analisando o gráfico, com a quantidade de Views por data (Verde) e a quantidade de Stocks (Azul), torna-se visível, a semelhança entre os aumentos e diminuições das views em comparação com os Stocks. Quando ocorre um aumento nas views dos vídeos da Sony, na maioria dos casos, os Stocks da empresa estão a aumentar ou vão nos meses seguintes aumentar. O mesmo é visível quando os valores diminuem.

Os grandes picos de stocks e views são provocados por momentos significativos para a empresa. Exemplos dos mesmos são o lançamento da Playstation 5, uma consola há muito esperada pelos consumidores da Sony. Este evento, que ocorreu no final do ano de 2020, resultou num aumento na quantidade de vídeos relacionados à empresa e nos seus stocks.

Outra data importante foi a divulgação de informações, através de trailers e vídeos promocionais e posterior lançamento do filme do Spider-Man: No Way Home, em dezembro de 2021. Este evento gerou o maior pico registrado tanto em views como em stocks da empresa, o que demonstra o processo estratégico relevante e eficaz da Sony.

DELL

Os *stocks* da Dell apresentaram um crescimento consistente ao longo do tempo, tendo ocorrido picos durante o ano de 2022, com o mais elevado a ocorrer no ano de 2024. As *views* dos vídeos, no entanto, são mais inconsistentes, mostrando picos aleatórios que não seguem de forma direta os movimentos dos *stocks*. Assim, é possível concluir que a relação entre *views* e os stocks da empresa não é tão estreita e que outros fatores têm maior peso no mercado para a Dell.

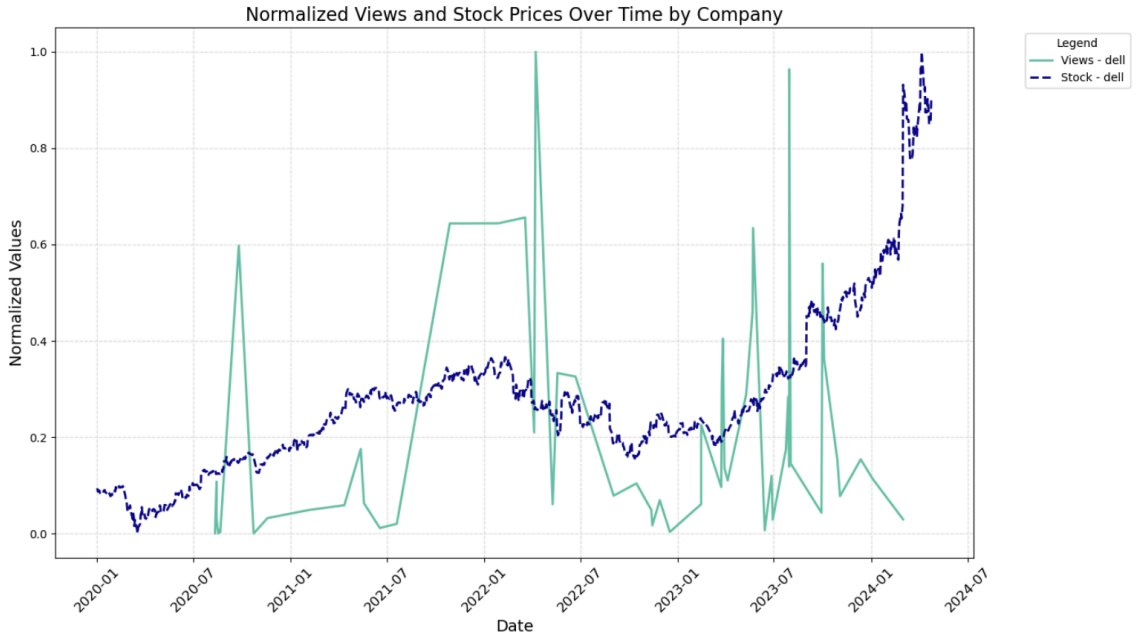


Fig. 6 - Merge entre as Views e os Stocks da Dell(2020-2024)

Microsoft

Os *stocks* da Microsoft têm, normalmente, um crescimento contínuo ao longo dos anos, com algumas diminuições pelo caminho. As *views* também apresentam picos relevantes, embora não sejam tão frequentes quanto os da Dell. Estes picos parecem coincidir com alturas estratégicas da Microsoft, onde o aumento de views, normalmente, ocorre durante o aumento dos stocks.

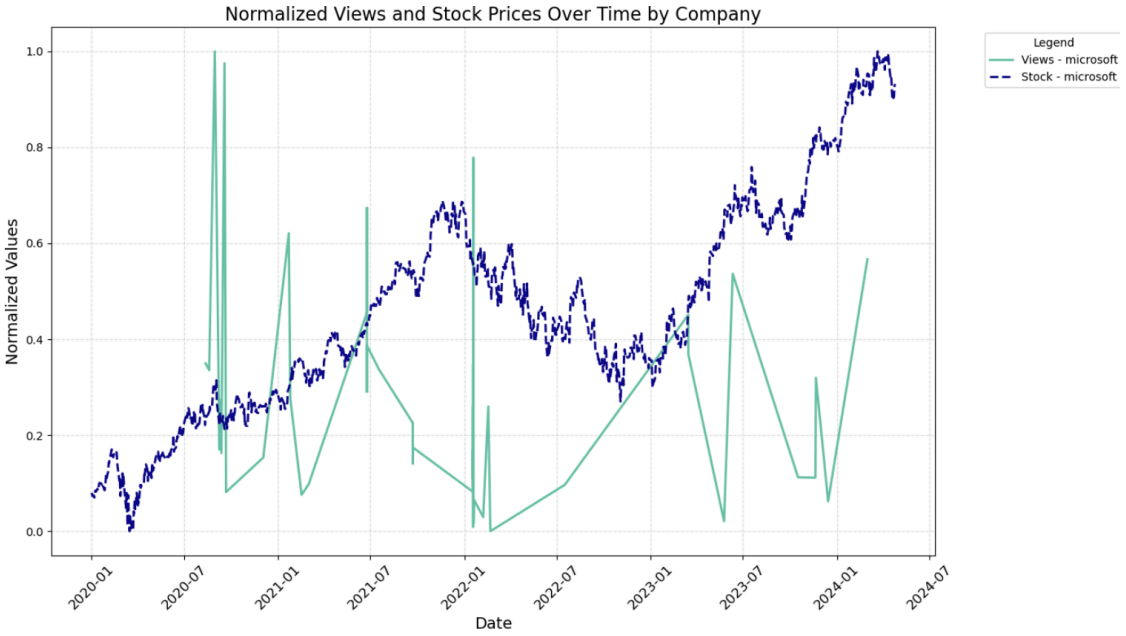


Fig. 7 - Merge entre as Views e os Stocks da Microsoft(2020-2024)

Intel

Os *stocks* mostram uma trajetória irregular, com quedas significativas em determinados períodos, especialmente entre 2021 e 2023, antes de uma leve recuperação em 2024. As *views* são altamente voláteis, com picos frequentes e abruptos. Esses picos, no entanto, parecem não ter uma relação direta ou consistente com os movimentos dos *stocks*. Isso pode indicar que o impacto de vídeos relacionados à Intel no mercado é limitado ou que outros fatores externos (como competição ou desempenho financeiro) influenciam mais os *stocks*.

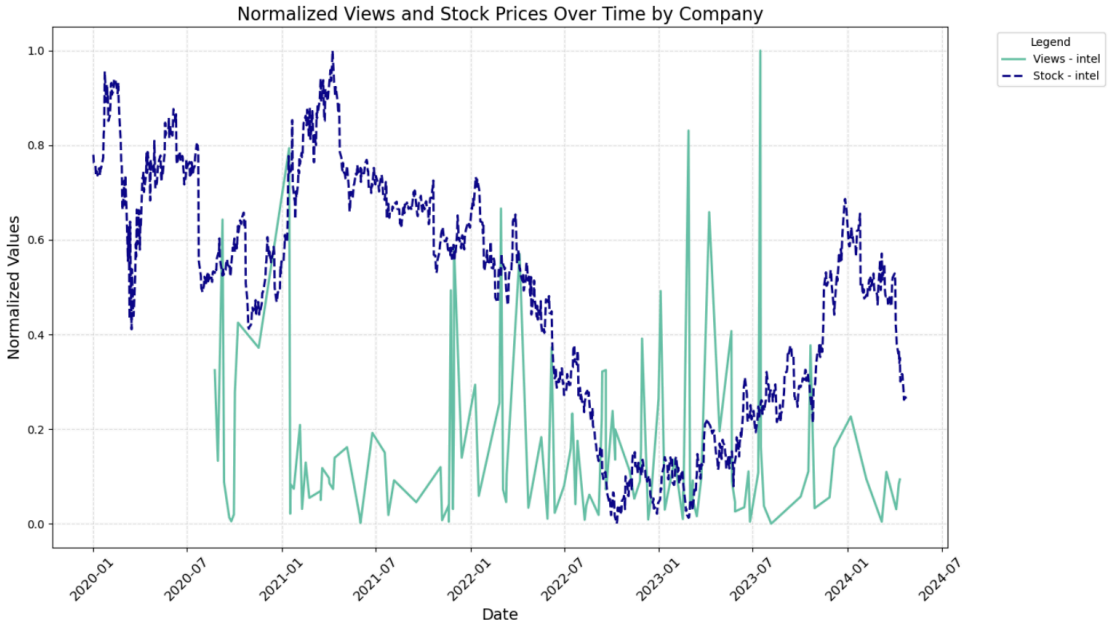


Fig. 8 - Merge entre as Views e os Stocks da Intel(2020-2024)

NVIDIA

A NVIDIA apresenta um crescimento consistente em seus *stocks*, que desde 2023 continuam a aumentar. Um contraste com as *views* dos vídeos, que se mostram inconsistentes, apresentando picos aleatórios ao longo dos anos. Assim, como ocorre com a Dell, é possível concluir que a relação entre *views* e os *stocks* da empresa não é tão estreita, indicando que outros fatores têm maior peso no mercado.

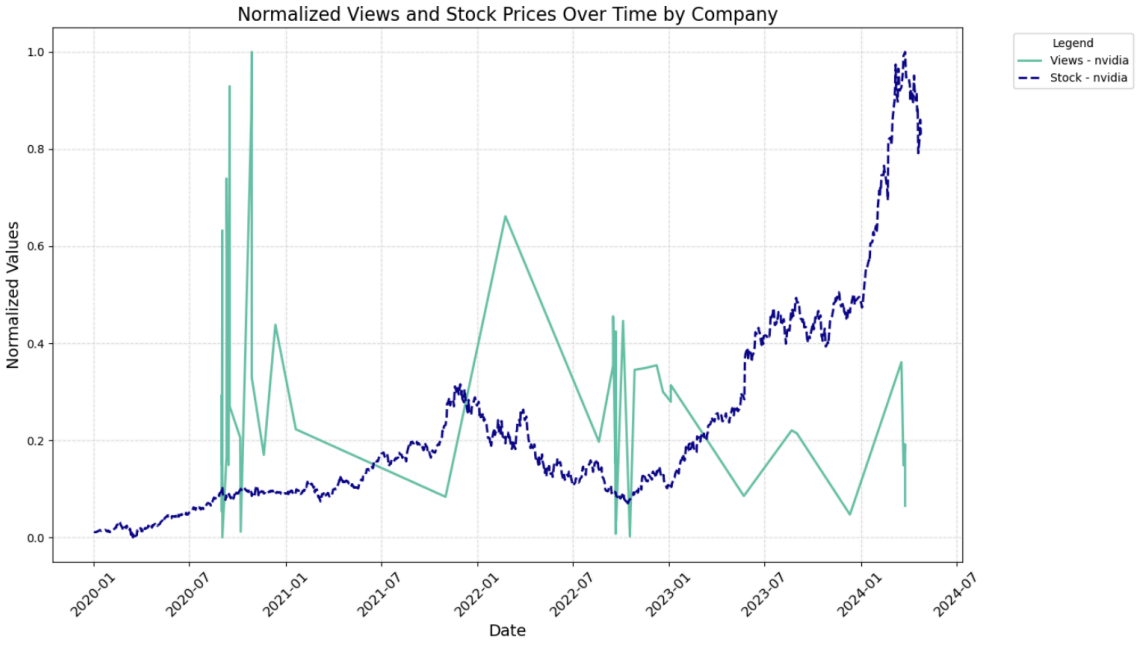


Fig. 9 - Merge entre as Views e os Stocks da Nvidia(2020-2024)

Ausência de Dados

É importante considerar a possível ausência de alguns vídeos nos dados disponíveis, o que significa que as observações atuais podem não refletir completamente a realidade. Além disso, há uma lacuna no conteúdo relacionado às empresas, e alguns desafios iniciais surgiram devido à falta de capacidade do sistema em executar algoritmos complexos. Como resultado, não foi possível implementar algoritmos mais sofisticados que possibilitasse uma busca mais eficaz por vídeos relevantes para cada empresa.

Fatores externos

Os fatores externos, como acontecimentos econômicos, políticos, tecnológicos, também têm poder na influência sobre os stocks e as views. Estes fatores influenciam o comportamento das pessoas e, consequentemente, os valores de visualização e o desempenho do mercado de stocks (torna-se mais imprevisível). Numa crise

econômica, por exemplo, o consumo de vídeos pode aumentar ou diminuir, o que afeta as visualizações, enquanto uma mudança política pode gerar insegurança e influenciar decisões de compra ou de investimento.

Implementação Técnica

Arquitetura

Esta arquitetura foi desenvolvida com o objetivo de possibilitar o processamento de grandes volumes de dados utilizando *PySpark* para este propósito, *MongoDB* para armazenar informações e *Streamlit* para apresentar os resultados obtidos de forma interativa.

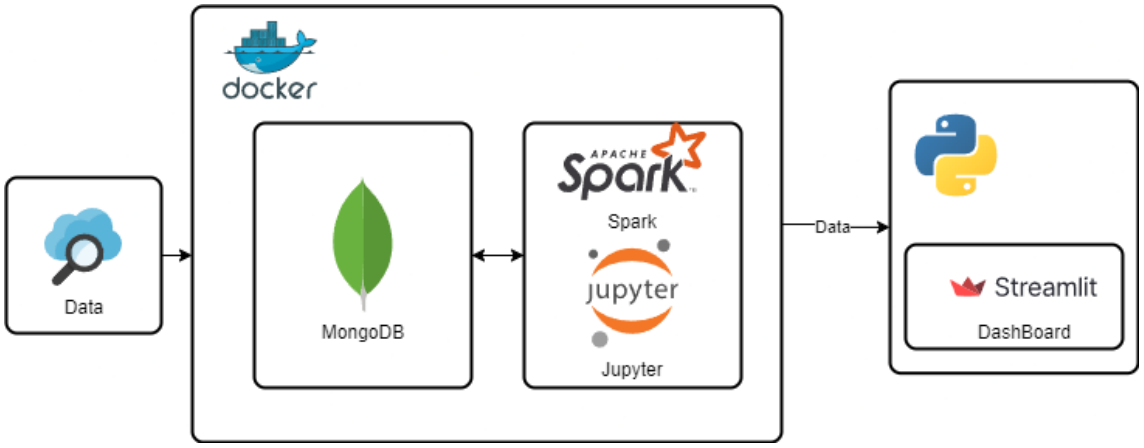


Fig. 10 - Arquitetura do Projeto

Descrição dos Componentes

Dados: Os dados são fornecidos em formato *JSON* e carregados para o *MongoDB*.

Docker: O *MongoDB* e o *Jupyter Notebook* estão encapsulados em containers *Docker* para facilitar a execução em diferentes ambientes.

MongoDB: Armazena os dados e é acessado pelo *PySpark*, para tratamento dos dados, e *Streamlit* para análises e visualização dos dados tratados e resultados.

Jupyter Notebook com PySpark: *PySpark* é utilizado para executar transformações e análises em grande escala nos dados armazenados no *MongoDB*. O *Jupyter Notebook* oferece um ambiente interativo para prototipação e desenvolvimento.

Streamlit: Fornece uma interface *web* interativa para visualização dos dados analisados e resultados.

Fluxo de Dados

Os dados passam por um conjunto de processos para serem apresentados e analisados no final, de forma a que se consigam retirar conclusões.

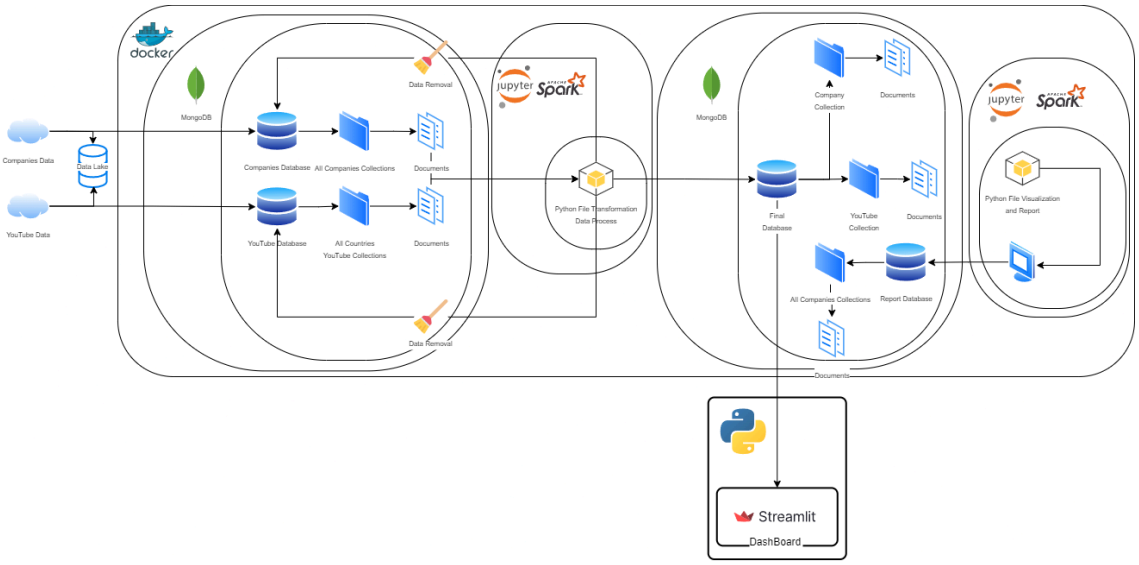


Fig. 11 - Arquitetura do Projeto (WorkFlow)

Inicialmente os dados são obtidos de datasets e guardados na base de dados do MongoDB respeitando a abordagem de **ELT** (o passo de guardar num data lake seria o ideal para manter os dados não tratados para fazer análises diferentes da objetivo deste projeto). Posteriormente são transformados efetuando os processos necessários para a finalidade pretendida utilizando o ambiente *Jupyter* integrado com *PySpark*. De seguida os dados são guardados numa nova base de dados com o propósito de guardar os dados trabalhados, enquanto os restantes são apagados. Com os dados já trabalhados para o objetivo, estes são usados para a visualização utilizando *Streamlit* para apresentação de resultados num ambiente interativo e dinâmico.

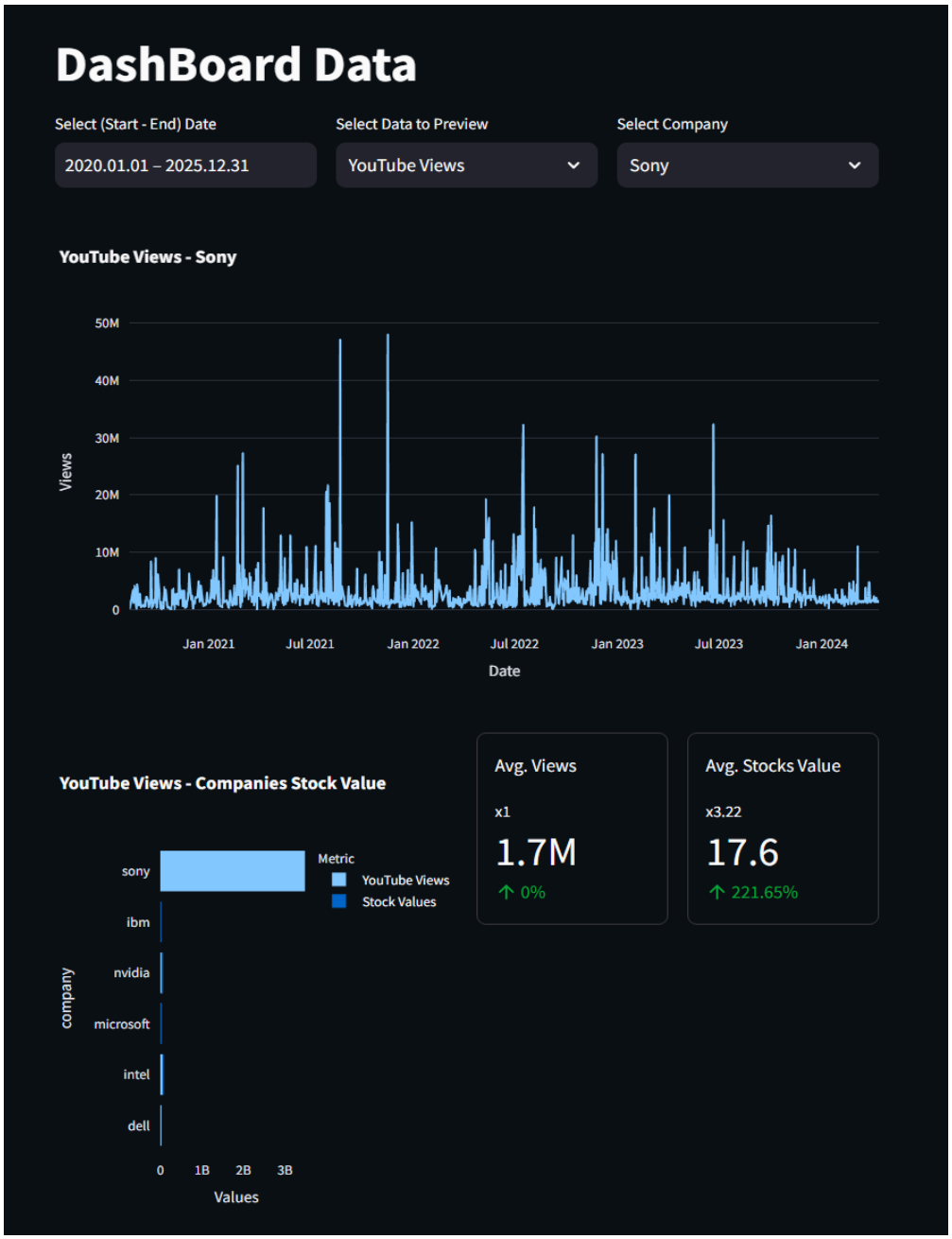
Vantagens da Arquitetura

Esta arquitetura permite o processamento distribuído e em grande escala usando o *PySpark* ideal para lidar com grandes volumes de dados e realizar transformações e análises complexas. Para além disto, a utilização do MongoDB (uma base de dados NoSQL) é altamente escalável e flexível, permitindo o armazenamento de diversas estruturas de dados em formato JSON, agrupado com uma abordagem **ELT** permite manter os dados originais antes do tratamento, possibilitando análises futuras, com outros objetivos. A utilização do *Docker* para encapsular o MongoDB e o *Jupyter Notebook* garante que a arquitetura seja portavel e possa ser executada em diferentes ambientes sem problemas de configuração, reduzindo desta forma a dependência de ambientes específicos. A integração destas ferramentas com *Streamlit* permite adquirir uma visualização e análise dos dados de forma interativa e dinâmica.

Assim esta arquitetura combina eficiência, escalabilidade e interatividade, garantindo um ambiente robusto para o processamento e visualização de grandes volumes de dados.

Dashboard

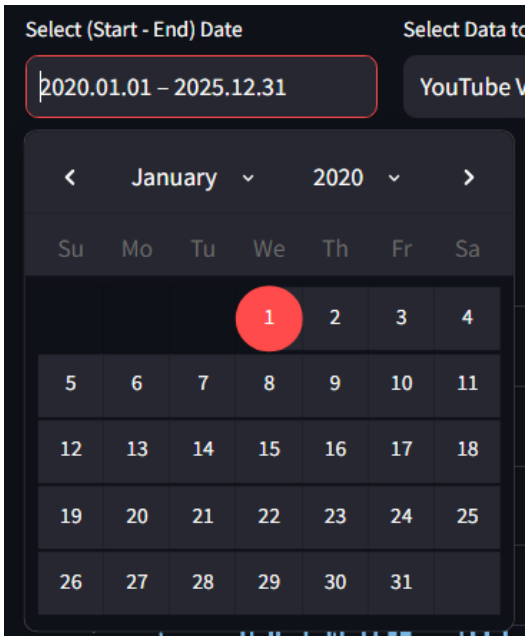
O dashboard foi criado para oferecer um ambiente interativo de visualização e análise de dados. Permite selecionar um intervalo de tempo, os dados a analisar e a empresa de interesse. Além destas opções, também disponibiliza informações adicionais, como o crescimento do valor das ações de uma empresa e o número de visualizações no YouTube.



Campos do Dashboard

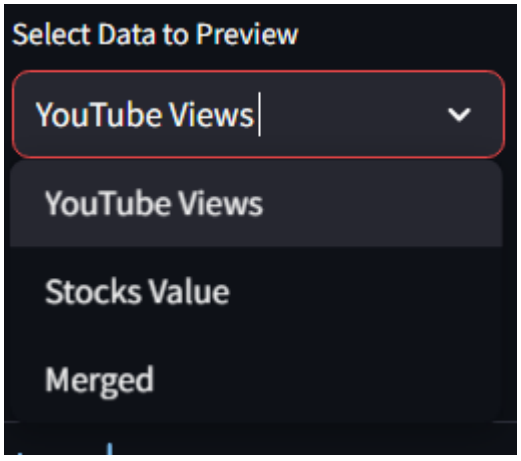
Seleção de Janela Temporal

Este campo permite selecionar o intervalo temporal a ser analisado. O intervalo padrão inicia-se a 1 de janeiro de 2020 e estende-se até 31 de dezembro do ano atual (2025, neste caso). A data mínima permitida é 1 de janeiro de 1998, e a máxima é 31 de dezembro do ano atual.



Seleção de Dados

Este campo permite escolher que dados se pretende analisar. A opção *Merged* possibilita a análise dos dois conjuntos de dados ao mesmo tempo.



Seleção de Empresa

Este campo permite escolher a empresa que se pretende analisar.

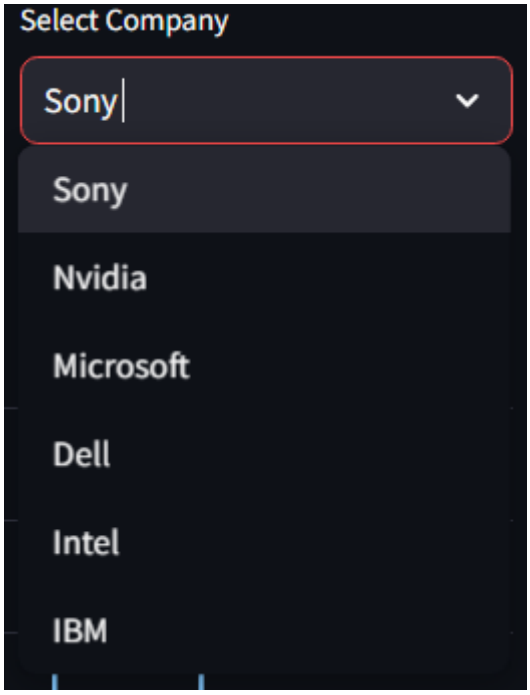
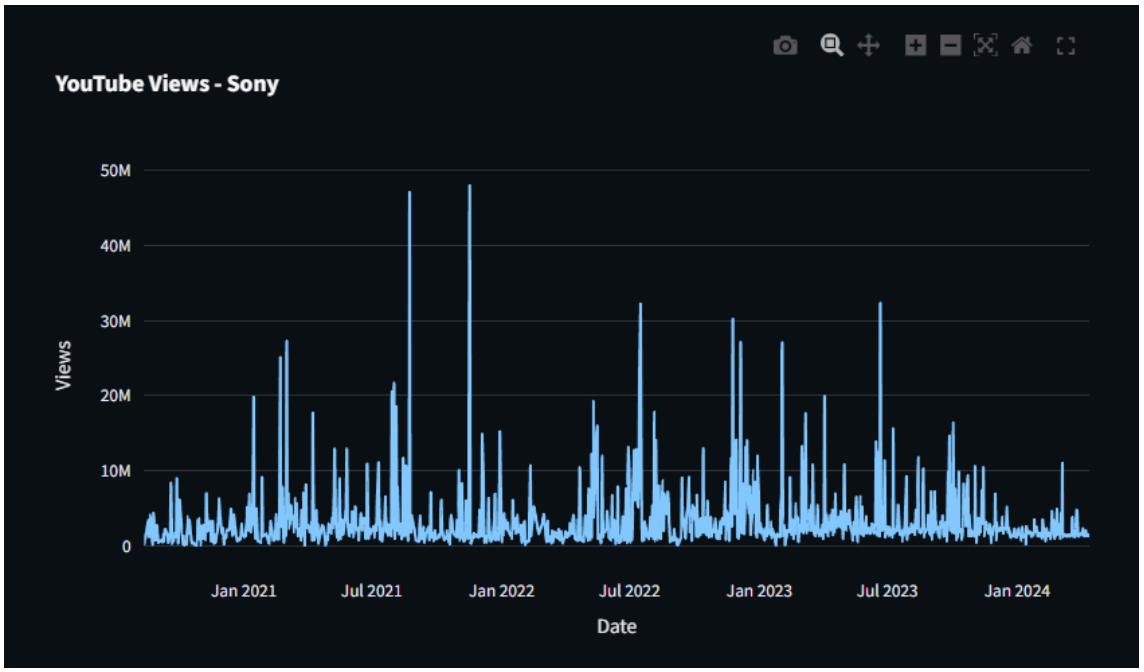


Gráfico Principal

Este gráfico demonstra a distribuição dos dados de acordo com os campos de seleção preenchidos.

YouTube Views - Sony

No seguinte gráfico está representado as *views* da empresa *Sony* na janela temporal padrão.



Stocks Value - Sony

Este gráfico demonstra a distribuição dos dados referentes ao valor dos stocks da empresa *Sony* na janela temporal padrão.

Merged - Sony

No seguinte gráfico está representado os dados de *views* do YouTube da empresa *Sony* com o seu valor de stocks.

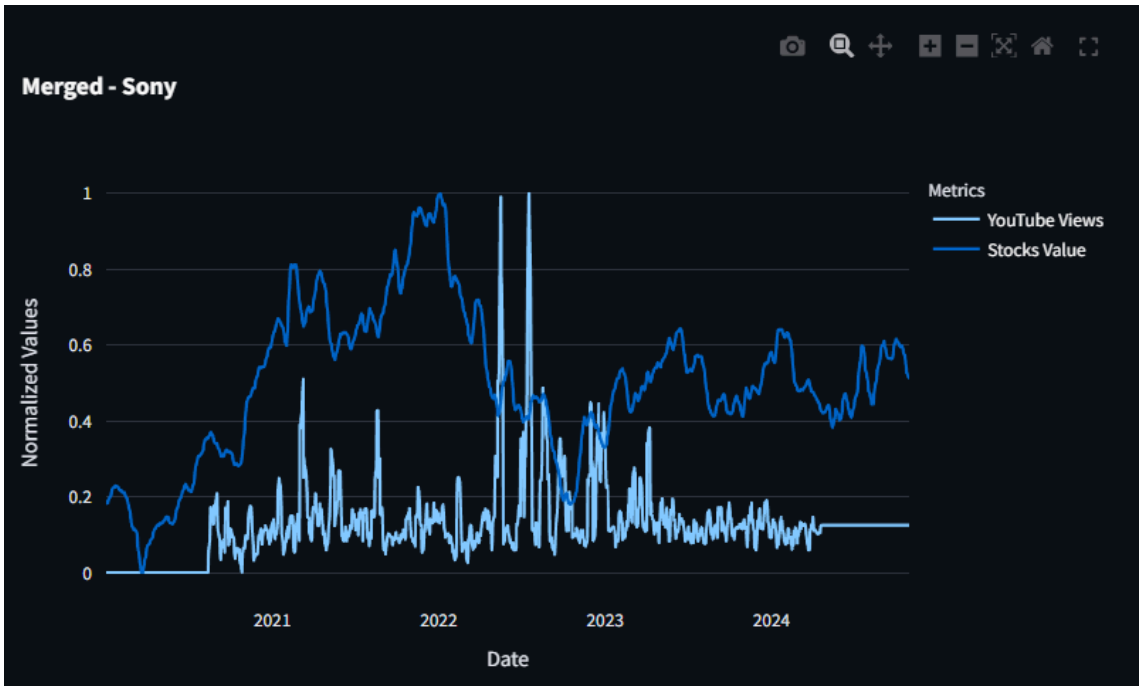
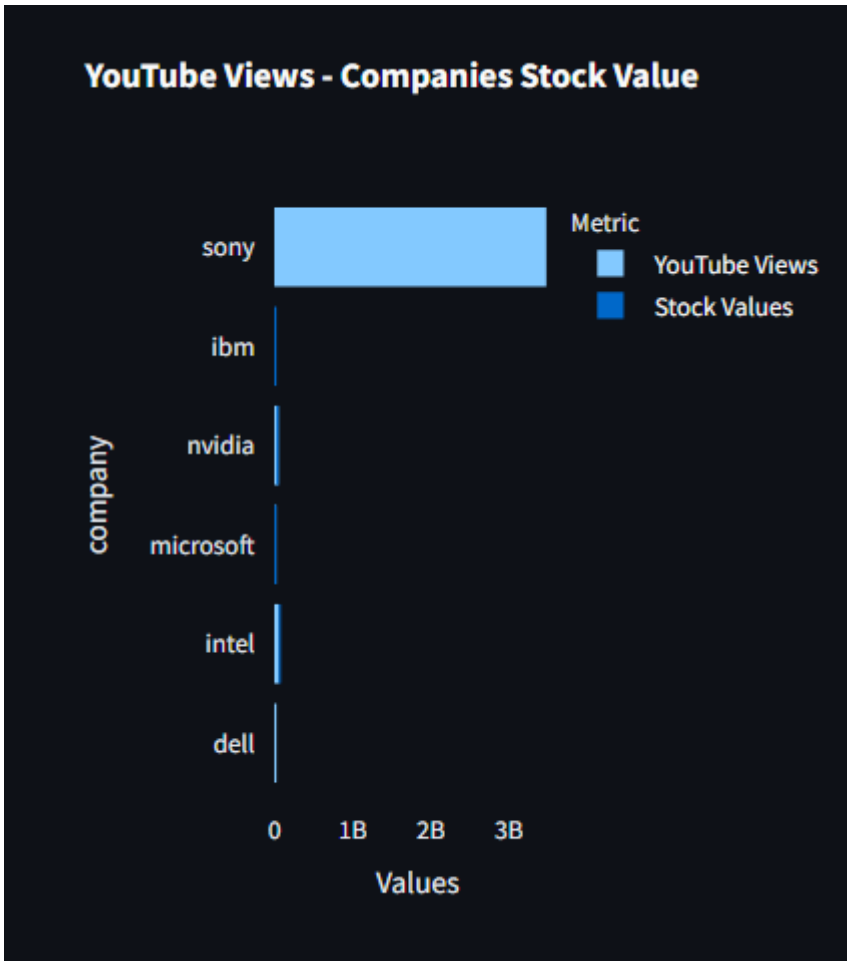


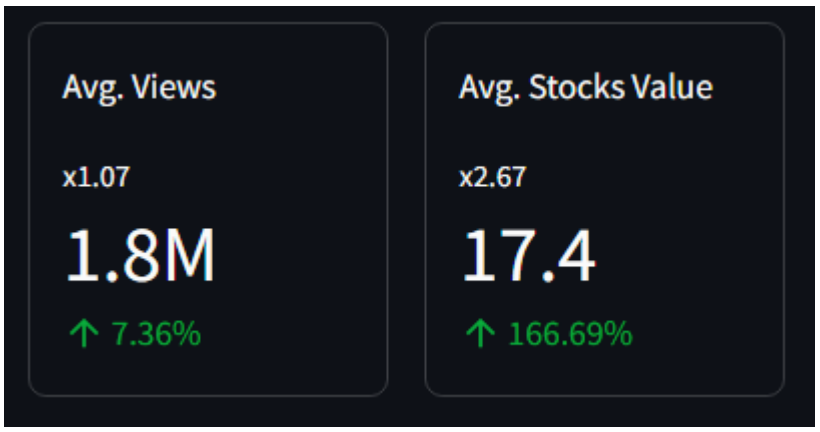
Gráfico Secundário

Este gráfico demonstra a quantidade de visualizações do YouTube de cada empresa comparativamente com o valor dos seus stocks.



Média de Views e Stocks

Esta secção permite analisar qual foi a média de visualizações e valor de stocks na janela temporal selecionada. Para além disso, também avalia em que proporção é que o stock aumentou e a percentagem deste aumento.



Como referido, os dados correspondem à janela temporal pretendida. Os valores apresentados são as médias de visualizações e valor do stock durante esse período.

A proporção é feita comparando a média de dados da janela temporal de 1 de Janeiro de 1998 até à data inicial da janela temporal selecionada, com a média dos dados na janela temporal selecionada.

Exemplo:

Avg. Atual = média dos dados na janela temporal selecionada

Avg. Anterior = média dos dados na janela temporal de 1 de Janeiro de 1998 até o início da janela temporal escolhida

Para o fator multiplicação:

$$fator\ multiplicação = \frac{Avg.\ Atual}{Avg.\ Anterior}$$

Para o fator percentagem:

$$fator\ percentagem = (\frac{100\ Avg.\ Atual}{Avg.\ Anterior}) - 100$$

Utilizemos o exemplo da imagem apresentada.

Avg. Atual = 1.8M

$$fator\ multiplicação = \frac{1.8M}{Avg.\ Anterior} = 1.07$$

$$fator\ percentagem = (\frac{1.8M\ .\ 100}{Avg.Anterior}) - 100 = 7.36$$

Conclusão e Trabalho Futuro

Conclusão

A análise realizada neste estudo demonstra que a relação entre visualizações de vídeos no YouTube e o comportamento do mercado de ações das empresas pode ser observada, embora com variações notáveis entre as diferentes corporações. Ao longo da análise, foi possível perceber que em algumas empresas, como a Sony, as visualizações de vídeos apresentam uma correlação clara com o desempenho dos seus stocks, especialmente durante eventos de grande visibilidade, como lançamentos de produtos e filmes. Em contraste, empresas como Dell, Intel e NVIDIA mostraram uma relação menos pronunciada entre os picos de visualizações e os movimentos do mercado de ações, sugerindo que fatores externos e outros indicadores podem influenciar mais diretamente o comportamento dos seus stocks.

Além disso, o trabalho evidenciou a importância de ter acesso a dados consistentes e abrangentes, uma vez que a ausência de dados completos pode impactar as conclusões tiradas a partir da análise. A implementação técnica utilizando MongoDB, PySpark e Streamlit foi eficaz para processar grandes volumes de dados e apresentar insights de maneira interativa, proporcionando um dashboard acessível para análise dinâmica dos dados.

Trabalho Futuro

Como sugestão para o trabalho futuro, seria interessante expandir a análise para outras plataformas de mídia social e explorar como diferentes tipos de conteúdo (não apenas vídeos no YouTube) podem influenciar o mercado de ações. Além disso, seria relevante incorporar outros fatores externos, como indicadores econômicos e políticos, para uma análise mais abrangente da influência das mídias digitais sobre o comportamento do mercado financeiro. Melhorias na busca e filtragem de dados também seriam essenciais para garantir que as análises futuras possam ser ainda mais precisas e detalhadas.

Apêndice Técnico

MongoDB

O MongoDB é uma base de dados NoSQL que foi criada em 2007, com o objetivo de armazenar documentos em coleções que posteriormente armazenam dados no formato de BSON. Por ser uma base de dados NoSQL, o MongoDB é vantajoso em algumas áreas, mas desvantajoso em outras.

Vantagens do MongoDB

- Alta Disponibilidade: Projetado para garantir o acesso contínuo aos dados, mesmo em caso de falhas.
- Escalabilidade Horizontal: Permite o particionamento de dados por vários servidores para lidar com grandes volumes de dados através do sharding.
- Flexibilidade: Permite armazenar e gerenciar dados estruturados, não estruturados e semi-estruturados com facilidade.
- Desempenho Rápido: Ótimo e ideal em cenários que exigem alta simultaneidade e acesso rápido aos dados, como rastreamento de publicidade online.
- Simplicidade de Desenvolvimento: Integração com JSON facilita o trabalho dos desenvolvedores em aplicações modernas.

Desvantagens do MongoDB

- Consistência Eventual: Adota o modelo BASE, o que pode levar a inconsistências temporárias nos dados em sistemas distribuídos.
- Precisão Transacional Limitada: Não é ideal para aplicações que exigem transações complexas e totalmente ACID, como operações financeiras.
- Análise em Tempo Real Restrita: Menos eficiente para cargas de trabalho que exigem análises avançadas em tempo real.
- Uso em Missões Críticas: Bancos de dados relacionais ainda são preferidos em ambientes empresariais que exigem confiabilidade e consistência rigorosa.

Principais comandos utilizados no Projeto em MongoDB

Estas foram as queries mais utilizadas para MongoDB. As mesmas têm o intuito de obter os documentos de acordo com empresa e janela temporal selecionada.

Utilizamos o find() para procurar as informações pedidas, tendo em conta os campos especificados nas condições de filtragem

```
find({"company": company.lower(), "publishedAt": {"$gte": start_date, "$lte": end_date}})
find({"Date": {"$gte": start_date, "$lte": end_date}})
find({"company": company.lower(), "publishedAt": {"$gte": min_date, "$lte": start_date}})
find({"company_name": company.lower(), "Date": {"$gte": min_date, "$lte": start_date}})
```

PySpark

O PySpark é a API Python utilizada no Apache Spark (framework de código aberto para processamento distribuído e em tempo real de grandes volumes de dados). Este permite a criação de análises e o processamento de dados de forma eficiente e flexível, sendo útil para quem já teve experiência com bibliotecas como o Pandas.

Vantagens do PySpark

- Computação Distribuída: Permite processar grandes volumes de dados em clusters de forma eficiente.
- Execução Preguiçosa: Realiza as operações somente quando os resultados são necessários, otimizando o uso de memória e desempenho.
- Suporte a APIs Avançadas: PySparkSQL; MLlib; GraphFrames
- Integração com Ecossistemas Modernos: Compatível com sistemas distribuídos como Hadoop, HDFS, e armazenamento em cloud.
- Flexibilidade com Python: Facilita a transição do Python para PySpark e de bibliotecas como Pandas.

Desvantagens do PySpark

- Complexidade na Configuração: Configurar clusters e ambientes para PySpark as vezes torna-se mais complicado do que em frameworks alternativas.
- Erros Difíceis de Depurar: As mensagens de erro são difíceis de compreender, às vezes, por culpa da existência de Python e Java e também das mensagens de erro pouco informativas.
- Sobrecarga de Processamento: O Spark é, em certas alturas, menos eficiente para pequenos conjuntos de dados ou tarefas simples, onde bibliotecas como Pandas são mais apropriadas.
- Somente prático em servidores horizontais: PySpark só é realmente eficaz em ambientes que suportam escalabilidade horizontal, como clusters distribuídos, o que limita a sua utilização em sistemas menores, como um computador local. Algo que foi observado durante este projeto.

Principais comandos utilizados no Projeto em PySpark

Dentro dos vários tipos de código e sintaxe utilizada a partir do PySpark, as principais e mais "únicas" em comparação com as habituais do python são:

SparkSession: Ponto de entrada para qualquer tipo de aplicação no PySpark. Responsável por criar e orientar as várias conexões com o cluster.

```
spark_ = SparkSession.builder \
    .appName("company_database") \
    .config("spark.mongodb.input.uri", "mongodb://mongodb:27017/Company_Database") \
    .config("spark.mongodb.output.uri", "mongodb://mongodb:27017/Company_Database") \
    .config("spark.mongodb.input.uri", "mongodb://mongodb:27017/Youtube_Database") \
    .config("spark.mongodb.output.uri", "mongodb://mongodb:27017/Youtube_Database") \
    .config("spark.jars.packages", "org.mongodb.spark:mongo-spark-connector_2.12:3.0.1") \
    .config("spark.driver.memory", "7g") \
    .config("spark.executor.memory", "7g") \
    .getOrCreate()
```

.read.format: Usado para especificar o formato de entrada dos dados e de ondes os mesmos são retirados, para posterior adição ao dataframe

```
df_mx = spark_.read.format("mongo").option("uri", "mongodb://mongodb:27017/Youtube_Database.youtube_data_mx").load()
```

.write.format: Usado para gravar DataFrames em formatos de saída específicos, neste caso para o MongoDB

```
df_yt.write.format("mongo") \
    .option("uri", "mongodb://mongodb:27017/Final_Database.youtube") \
    .mode("append") \
    .save()
```

toPandas(): Um método para converter um DataFrame PySpark (distribuído) em um DataFrame Pandas (em memória local), o que permite maior flexibilidade nas visualizações dos dados.

```
plot_data_yt = df_views_over_time.toPandas()
```

Bibliografia

[1] Erickson, J. (2024). **What Is MongoDB? An Expert Guide**. OCI.

https://www.oracle.com/pt/database/mongodb/

[2] Wikipedia contributors. **MongoDB**. Wikipedia.

https://pt.wikipedia.org/wiki/MongoDB

[3] MongoDB. **JSON and BSON Basics**. MongoDB Resources.

https://www.mongodb.com/resources/basics/json-and-bson

[4] Domino Data Lab. **PySpark - Data Science Dictionary**. Domino.ai.

https://domino.ai/data-science-dictionary/pyspark

[5] Wikipedia contributors. **PlayStation 5**. Wikipedia.

https://en.wikipedia.org/wiki/PlayStation_5

[6] Wikipedia contributors. **Spider-Man: No Way Home**. Wikipedia.

https://en.wikipedia.org/wiki/Spider-Man:_No_Way_Home

[7] RBC Global Asset Management. **What's the Relationship Between the Stock Market and the Economy?**

https://www.rbcgam.com/en/ca/learn-plan/investment-basics/whats-the-relationship-between-the-stock-market-and-the-economy/detail