

Relatório do trabalho da disciplina de Integração de Sistemas de
Informação

Processos de ETL em KNIME

Hugo Filipe Nogueira Silva – a16368

Licenciatura em Engenharia Sistemas Informáticos (Pós-Laboral)

Outubro de 2024

Afirmo por minha honra que não recebi qualquer apoio não autorizado na realização deste trabalho prático. Afirmo igualmente que não copiei qualquer material de livro, artigo, documento web ou de qualquer outra fonte exceto onde a origem estiver expressamente citada.

Hugo Filipe Nogueira Silva – a16368

Índice

ENQUADRAMENTO	5
PROBLEMA	6
PROTOTIPAGEM	7
ESTRATÉGIA UTILIZADA	8
VÍDEOS DEMONSTRATIVOS	9
1º Parte	9
2ª Parte (Correções e melhorias da 1ª Parte)	9
TRANSFORMAÇÕES	10
Diagramas:	10
Explicação das principais ferramentas utilizadas:	14
JOBS	16
Explicação	16
RESULTADOS	20
CONSTRANGIMENTOS/DIFICULDADES	25
CONCLUSÃO E TRABALHOS FUTUROS	26
BIBLIOGRAFIA	27

Lista de Figuras

Figura 1 - Components PSI-20	5
Figura 2 - Financial Company Profile API	6
Figura 3 – 1ª Parte - Código QR - Vídeo Demonstrativo	9
Figura 4 - 2ª Parte - Código QR - Vídeo Demonstrativo	9
Figura 5 - Only Stocks Lisbon	10
Figura 6 - Details All Stock EURONEXT	11
Figura 7 - Yahoo Finance CSV Stock Lisbon Retail	11
Figura 8 - JOIN Stocks Lisbon with details	12
Figura 9 - JOIN Stocks Lisbon Retails with details	12
Figura 10 - Reports with Charts	13
Figura 11 - Global Stocks to Iterative Dashboards with moving average	13
Figura 12 - Global Stocks with symbol to API Get Request	14
Figura 13 - Job nº1	16
Figura 14 - Job nº2	16
Figura 15 - Job nº3	17
Figura 16 - Job nº4	17
Figura 17 - Job nº5	18
Figura 18 - Job nº6	18
Figura 19 - Job nº7	19
Figura 20 - Resultados - TableView	20
Figura 21 - Resultados - Ganho/Perda Stocks PSI20	21
Figura 22 - Resultados – GoogleSheet	21
Figura 23 - Resultados – EmailSender	22
Figura 24 - Resultados - MA Ação	22
Figura 25 - Resultados - MA Comparativo entre Ações	23
Figura 26 - Resultados - Combinação de API's para gráficos iterativos com dados atuais	24

Enquadramento

Este trabalho prático insere-se na disciplina de Integração de Sistemas de Informação, parte do curso de Licenciatura em Engenharia de Sistemas Informáticos. O principal objetivo é aplicar e explorar as diversas técnicas e ferramentas de ETL (Extração, Transformação e Carregamento de dados) em cenários práticos que envolvem a integração e manipulação de grandes volumes de dados provenientes de diferentes fontes.

O trabalho foi realizado utilizando a plataforma KNIME, uma ferramenta poderosa e versátil de ETL que permite a criação de workflows visuais para integrar, transformar e analisar dados de diversas origens. O foco deste projeto foi demonstrar como, através de processos automáticos, é possível integrar dados provenientes de APIs remotas, processá-los e gerar resultados que podem ser visualizados de forma intuitiva através de gráficos e relatórios.

Neste contexto, optou-se pela utilização de dados financeiros públicos relacionados com as ações da bolsa de valores portuguesa (PSI-20) e um exemplo de ações globais extraídos através de APIs de fontes como o Yahoo Finance e a Financial Modeling Prep. O objetivo central foi criar um fluxo de trabalho que extraia os dados dessas APIs, realize a transformação necessária para filtrar e organizar as informações relevantes, e apresente os resultados de forma clara e visual, através de gráficos de desempenho diário das ações.

































































<input type="checkbox"/>	 Galp Energia	16,77	16,82	16,64	+1,33%	169,28K			11:43:30	
<input type="checkbox"/>	 Jeronimo Martins	16,93	17,17	16,88	-0,12%	112,75K			11:43:16	
<input type="checkbox"/>	 Energias de Portugal	3,823	3,915	3,821	-2,05%	1,53M			11:45:55	
<input type="checkbox"/>	 Altri	5,130	5,145	5,120	0,00%	24,08K			11:45:14	
<input type="checkbox"/>	 BCP	0,4287	0,4317	0,4275	-0,69%	10,73M			11:45:09	
<input type="checkbox"/>	 The Navigator	3,658	3,688	3,658	-0,81%	106,77K			11:44:58	
<input type="checkbox"/>	 Mota-Engil	2,552	2,596	2,542	-1,31%	425,45K			11:45:07	
<input type="checkbox"/>	 REN	2,375	2,390	2,370	-0,63%	323,42K			11:38:49	
<input type="checkbox"/>	 Semapa	14,92	15,00	14,76	+0,67%	6,47K			11:42:55	
<input type="checkbox"/>	 Sonae	0,9360	0,9380	0,9310	+0,54%	328,65K			11:45:44	
<input type="checkbox"/>	 Nos SGPS SA	3,62	3,63	3,62	-0,28%	17,97K			11:35:39	
<input type="checkbox"/>	 EDP Renováveis	13,79	14,08	13,79	-1,15%	111,42K			11:41:09	
<input type="checkbox"/>	 Corticeira Amorim	8,89	8,93	8,86	+0,23%	20,04K			11:28:48	
<input type="checkbox"/>	 Ibersol Reg	7,10	7,28	7,10	-2,20%	14,7K			10:40:07	
<input type="checkbox"/>	 CTT Correios de Portug...	4,27	4,29	4,26	-0,35%	11,71K			11:33:05	
<input type="checkbox"/>	 Greenvolt Energias Ren...	8,250	8,280	8,215	-0,96%	30,35K			11:37:28	

Figura 1 - Components PSI-20

Problema

O presente trabalho pretende resolver o desafio da integração e transformação de dados financeiros provenientes de diversas fontes de informação, como APIs públicas. Neste caso específico, o foco recai sobre a análise do mercado de ações. O problema a abordar consiste na extração automática de dados financeiros referentes às ações listadas, a sua transformação e organização em dados utilizáveis para posterior análise, e a visualização desses dados de forma acessível e compreensível.

Adicionalmente, um dos principais desafios é identificar, entre as ações listadas, aquelas que estão disponíveis para compra no mercado de retalho, filtrando os dados de modo a remover as que não se enquadram nesta categoria. Através deste processo, pretende-se gerar gráficos de desempenho diário, que evidenciem quais as ações com maior valorização e desvalorização no decorrer do dia, proporcionando uma análise clara e intuitiva para o utilizador final. A ideia aqui foi criar métodos de avaliação de investimentos em stocks.

O projeto procura, assim, demonstrar como as ferramentas de ETL podem ser utilizadas para automatizar processos complexos de manipulação de dados, otimizando o fluxo de trabalho desde a extração até à visualização final dos resultados.

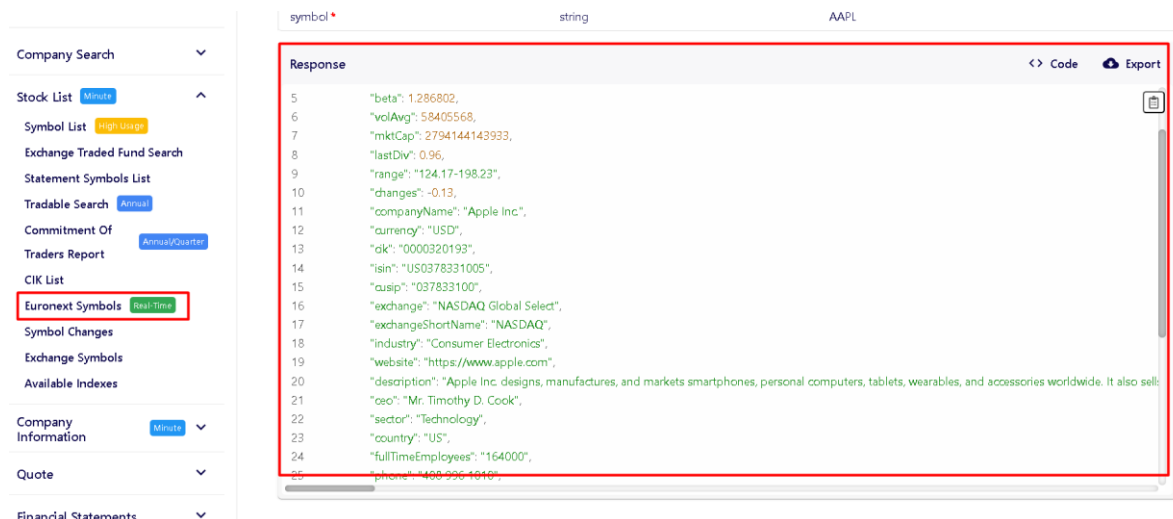
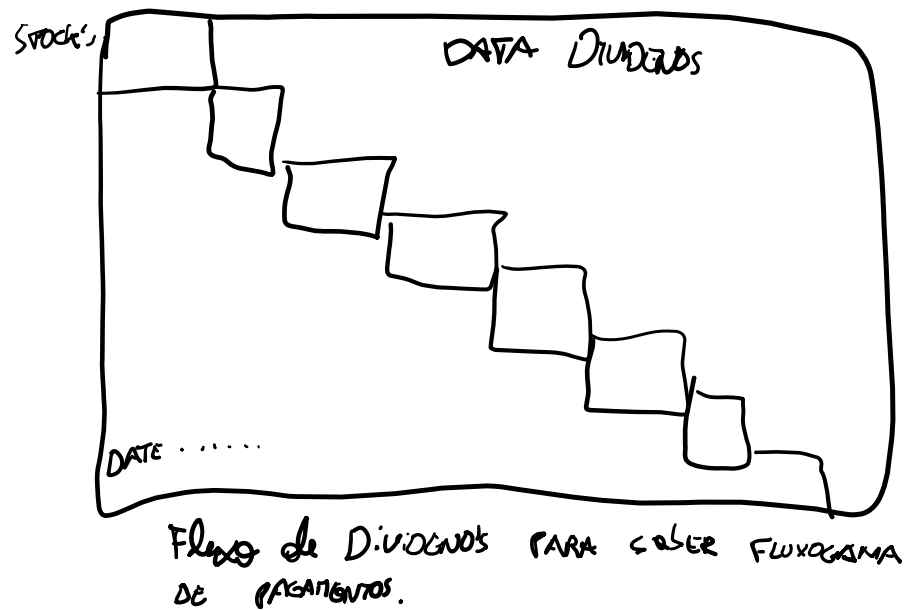
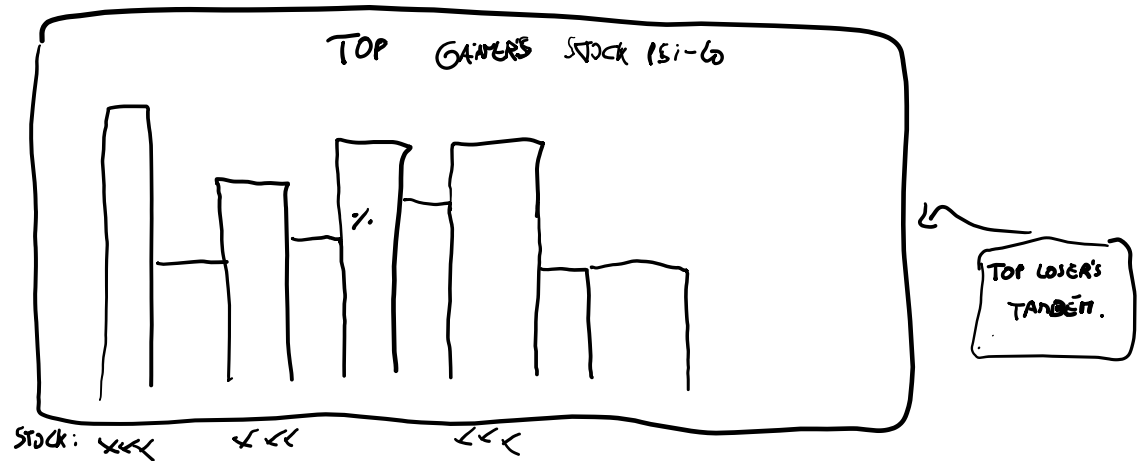


Figura 2 - Financial Company Profile API

Prototipagem



Estratégia Utilizada

A estratégia foi dividida nas seguintes fases:

- **Extração de Dados:** Utilizando um GET Request para acessar dados de uma API pública (Yahoo Finance e Financial Modeling Prep).
- **Transformação:** Uso de expressões regulares para normalização de campos e filtros para selecionar ações disponíveis para compra.
- **Carregamento de Dados:** Os dados foram carregados em tabelas processáveis e foram gerados gráficos para visualização dos resultados.
- **Extração dos Dados:** Os dados foram extraídos para uma base de dados SQLite e também para um ficheiro Google Sheets.
- **Visualização:** Geração de gráficos simples ou iterativos para identificar as ações que mais ganharam e perderam no dia, moving average etc..

Vídeos demonstrativos

Para complementar a explicação teórica do processo de ETL desenvolvido neste projeto, foram criados dois vídeos demonstrativo que ilustram, de forma visual e prática, a implementação e execução dos workflows no KNIME. Os vídeos mostram as diferentes fases do processo, desde a extração dos dados via API até à geração dos gráficos e dashboards, permitindo uma melhor compreensão dos procedimentos e resultados alcançados. Este recurso audiovisual facilita a assimilação dos conceitos e oferece uma visão mais dinâmica e interativa do projeto, destacando as principais funcionalidades utilizadas e o fluxo de trabalho aplicado.

1º Parte



Figura 3 – 1ª Parte - Código QR - Vídeo Demonstrativo

2ª Parte (Correções e melhorias da 1ª Parte)



Figura 4 - 2ª Parte - Código QR - Vídeo Demonstrativo

Transformações

Esta secção detalha as operações de transformação aplicadas aos dados extraídos, como a normalização de campos, o uso de expressões regulares, a realização de joins entre diferentes conjuntos de dados, e a filtragem de informações irrelevantes para a análise. As transformações foram essenciais para garantir que os dados estivessem no formato adequado para posterior análise e visualização, assegurando a sua consistência e qualidade ao longo de todo o processo de ETL.

Diagramas:

Este diagrama mostra a extração dos dados da API com informações de todas as ações da bolsa de valores de Lisboa. A transformação principal aqui envolve a filtragem para selecionar apenas ações disponíveis na Euronext Lisboa, descartando as de outras bolsas.

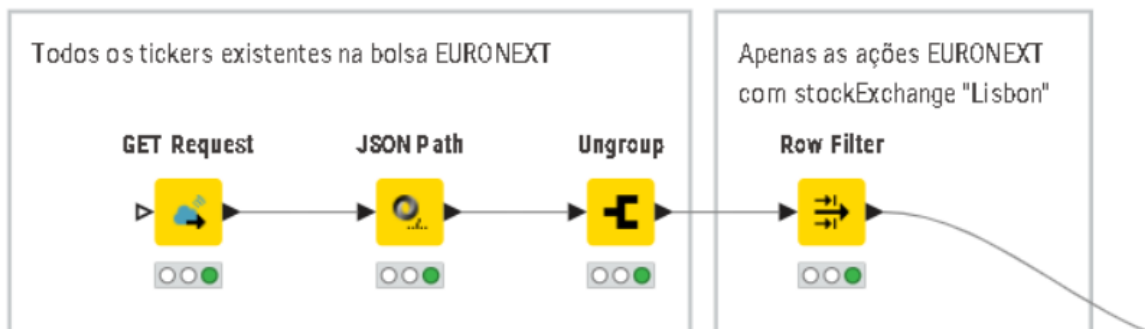


Figura 5 - Only Stocks Lisbon

Neste passo, os dados das ações extraídas são enriquecidos com mais detalhes, como nome, preço, e variação diária, usando operações de junção (join) com outra fonte de dados para conectar informações por símbolo da ação.



Figura 6 - Details All Stock Euronext

Aqui é feita a utilização de um ficheiro CSV proveniente do Yahoo Finance, que contém dados de ações listadas no mercado de retalho de Lisboa. Foi aplicada uma transformação de filtro para selecionar apenas as ações do setor retalhista.

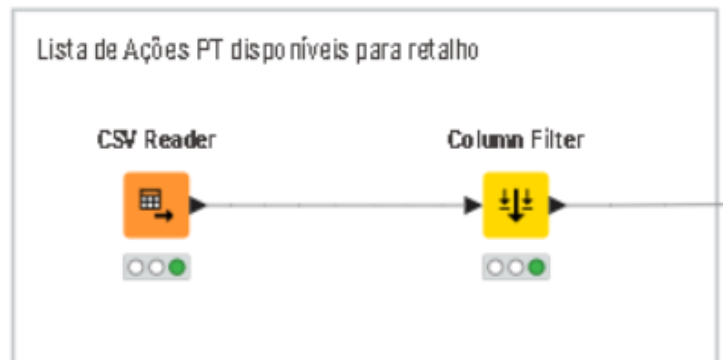


Figura 7 - Yahoo Finance CSV Stock Lisbon Retail

Este diagrama representa a junção dos dados das ações filtradas com mais detalhes obtidos de outra API. A transformação aplicada aqui foi o "Inner Join", conectando as tabelas de acordo com o símbolo da ação.

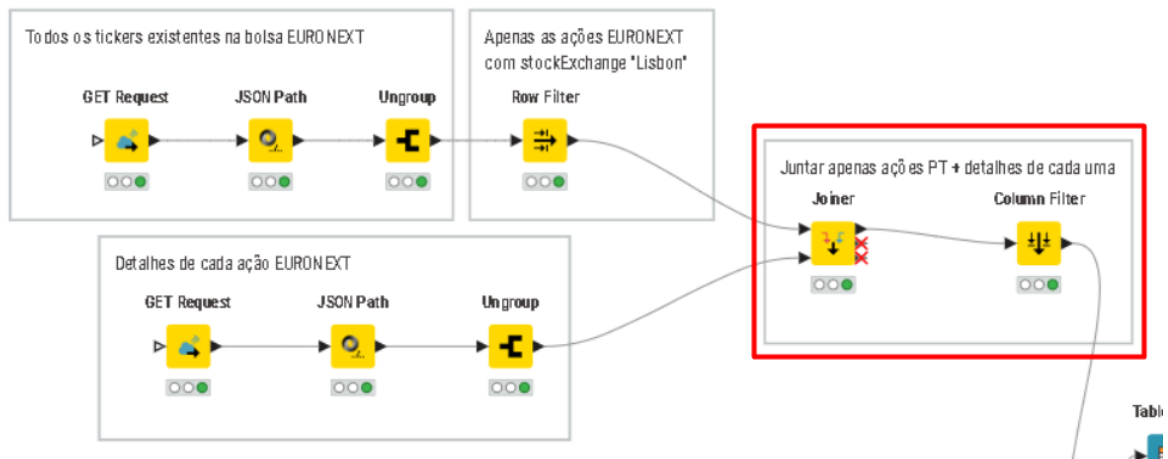


Figura 8 - JOIN Stocks Lisbon with details

Similar ao diagrama anterior, mas focado nas ações retalhistas, onde as informações detalhadas foram combinadas com dados filtrados do mercado de retalho, utilizando novamente operações de junção para criar uma tabela mais completa.

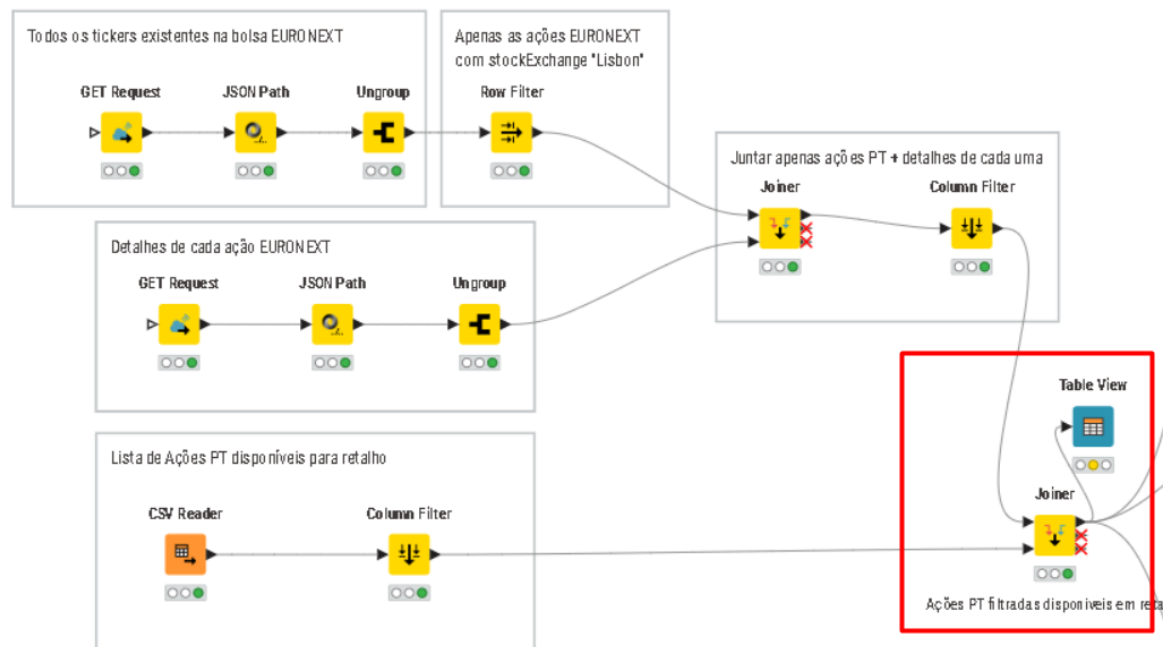


Figura 9 - JOIN Stocks Lisbon Retails with details

A transformação final envolve a criação de gráficos de desempenho das ações, onde foram aplicados cálculos agregados, como a média móvel e variação de preço, para apresentar as tendências de mercado de maneira visual.

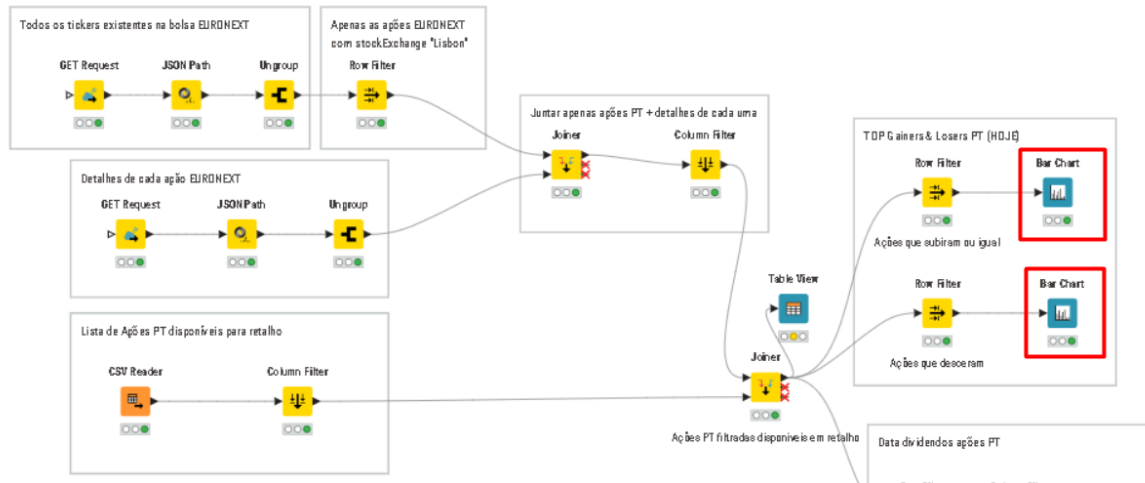


Figura 10 - Reports with Charts

Este diagrama representa a criação de dashboards iterativos utilizando dados globais de ações. A transformação principal aqui envolve a aplicação de uma "média móvel" (moving average), que permite suavizar as flutuações diárias dos preços das ações, destacando tendências ao longo do tempo. Este cálculo de média móvel é aplicado para tornar os dados mais compreensíveis e identificar padrões de comportamento.

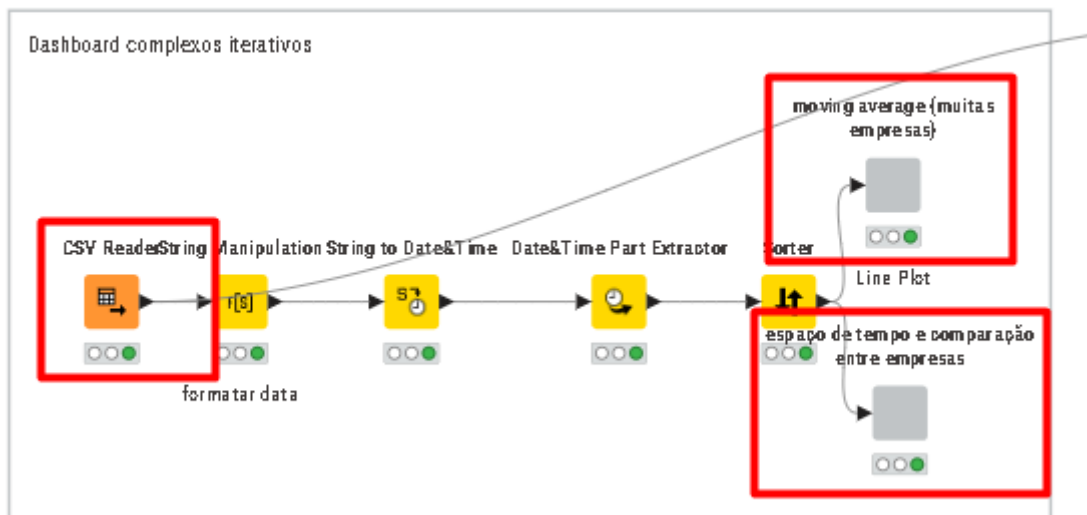


Figura 11 - Global Stocks to Iterative Dashboards with moving average

Neste passo, os dados globais das ações são usados para criar um loop de "GET Requests" a uma API externa. Cada símbolo de ação é utilizado para gerar pedidos de dados adicionais, automatizando a obtenção de informações detalhadas para cada ação específica. Essa transformação facilita a atualização contínua dos dados, assegurando que os dashboards recebam dados mais atuais sobre as ações globais.

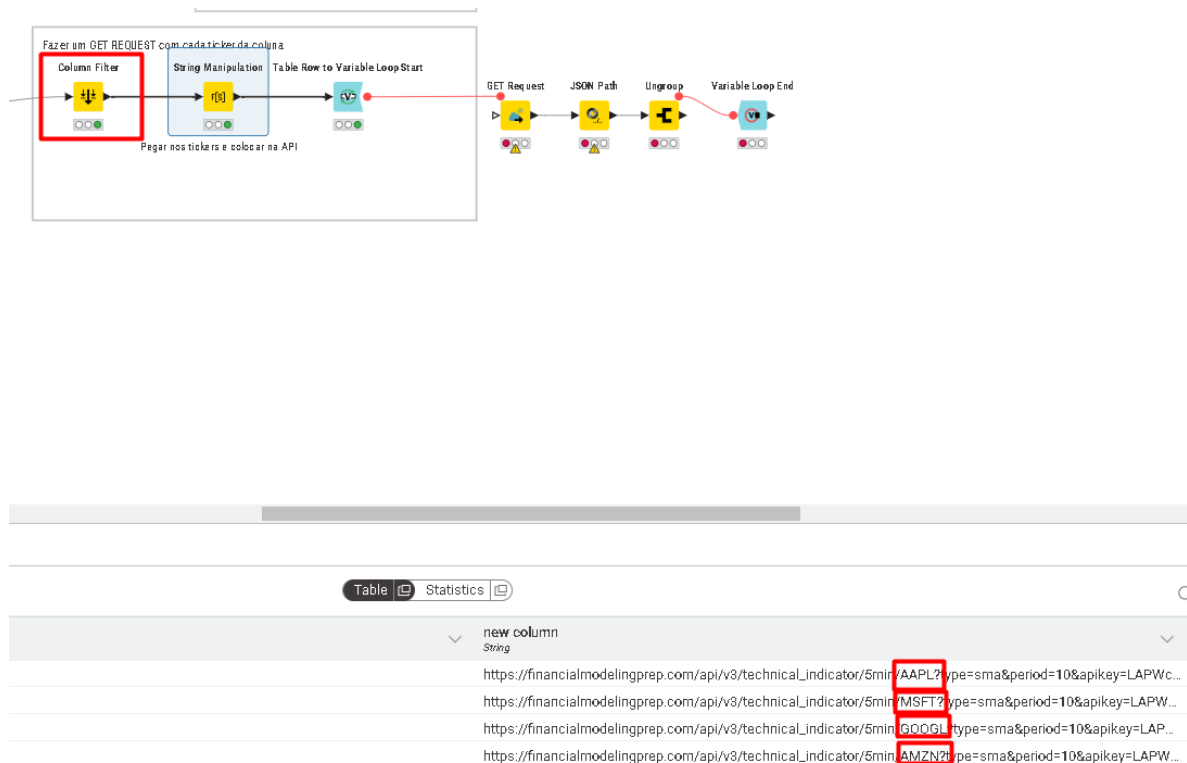


Figura 12 - Global Stocks with symbol to API Get Request

Explicação das principais ferramentas utilizadas:

- **Extração de Dados (GET Request):** Para obter dados das APIs públicas, como a Yahoo Finance e a Financial Modeling Prep, utilizou-se o node GET Request. Esta ferramenta permitiu a realização de chamadas HTTP e a captura dos dados de ações em formato JSON. Através deste processo, foram extraídos dados detalhados sobre as ações do PSI-20 e ações globais.
- **Leitura de Ficheiros (CSV Reader):** Para integrar dados locais, como os ficheiros CSV provenientes do Yahoo Finance, utilizou-se o CSV Reader, uma ferramenta que converte os ficheiros CSV em tabelas processáveis dentro do KNIME. Este node permitiu incorporar informações do mercado de retalho de Lisboa no workflow.
- **Filtragem de Dados (Row Filter):** Para garantir que apenas as ações de interesse fossem processadas, foram aplicados filtros. O node Row Filter foi usado para selecionar dados relevantes, como as ações da Euronext Lisboa ou ações disponíveis para compra no

retalho. Este processo de filtragem foi crucial para focar apenas nos dados necessários à análise.

- **Combinação de Dados (Joiner):** Um dos pontos importantes foi a junção de várias fontes de dados, combinando informações das ações com detalhes adicionais. O node Joiner foi utilizado para realizar operações de junção entre tabelas com base em chaves comuns, como o símbolo das ações. Esta operação permitiu agregar informações complementares de várias origens numa única tabela.
- **Transformação de Dados (Math Formula, String Manipulation):** Várias operações de transformação foram aplicadas aos dados. O node Math Formula permitiu realizar cálculos, como o preço médio ou variações de preços ao longo do tempo, enquanto o String Manipulation foi utilizado para normalizar e ajustar campos textuais, assegurando a consistência dos dados.
- **Agregação e Análise (GroupBy, Moving Average):** Para facilitar a análise dos dados financeiros, foi necessário agrupar e calcular médias. O node GroupBy permitiu agrupar as ações com base em critérios específicos (por exemplo, ações da mesma bolsa), enquanto o Moving Average foi utilizado para suavizar as variações diárias e identificar tendências ao longo do tempo.
- **Automatização de Processos (Loops):** Um aspeto relevante foi a automatização da extração de dados. Utilizando nodes de loop como o Table Row to Variable Loop Start e Loop End, foi possível iterar sobre cada ação e realizar chamadas a APIs de forma automática, garantindo a obtenção de dados atualizados para cada símbolo de ação.
- **Visualização e Relatórios (Data to Report, JavaScript View):** A fase final do processo envolveu a criação de relatórios visuais e dashboards. Utilizou-se o Data to Report para converter os resultados das transformações em relatórios que incluíam gráficos de desempenho e dashboards interativos, onde as médias móveis e outras métricas foram apresentadas de forma clara e acessível.

Jobs

Explicação

Job nº1:

Juntar todas as ações portuguesas extraídas da bolsa europeia e com detalhes de cada uma.

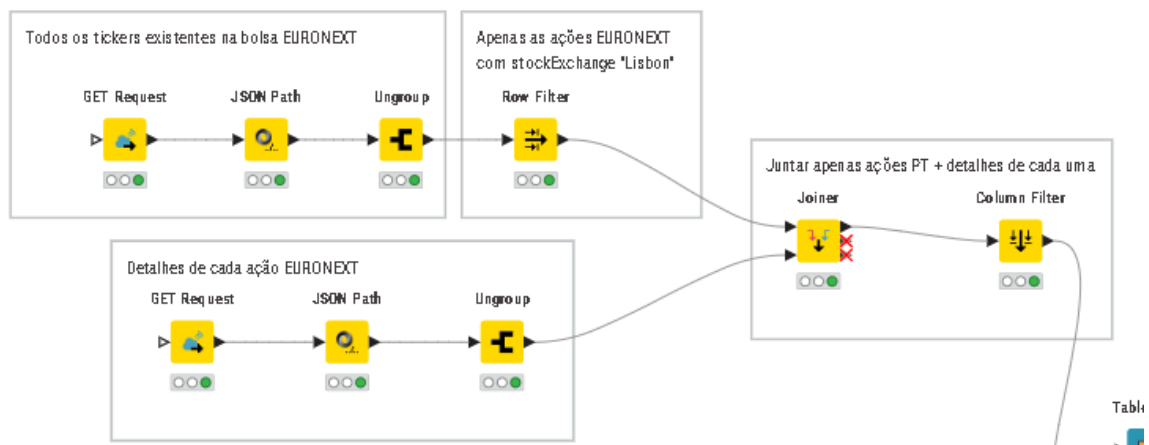


Figura 13 - Job nº1

Job nº2:

Juntar todas as ações portuguesas com detalhe anteriormente conseguidas no Job nº1 mas agora filtrando apenas pelas ações disponíveis em retalho.

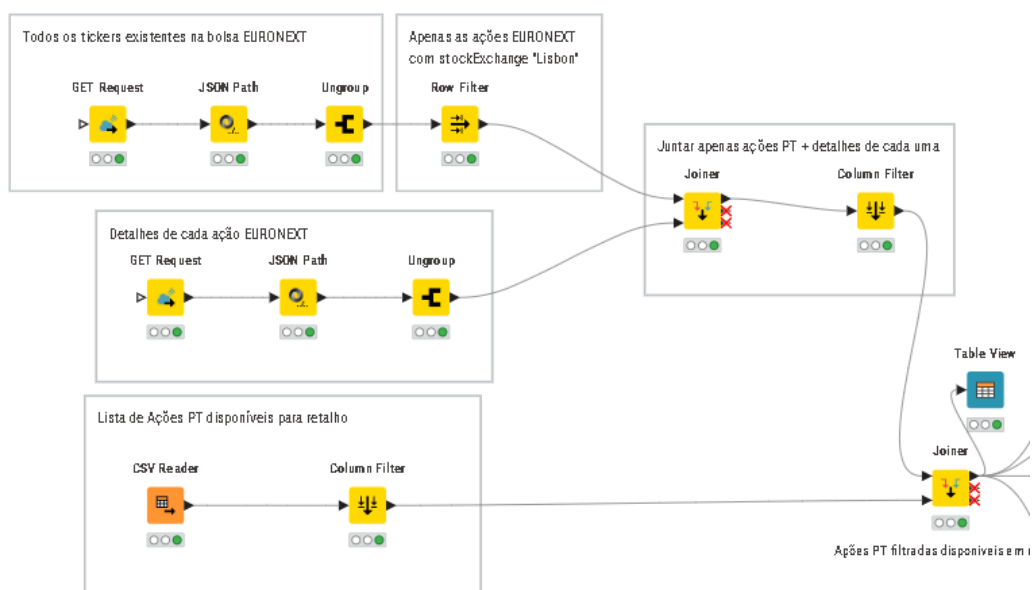


Figura 14 - Job nº2

Job nº3:

Utilizar todas as transformações realizadas no Job nº1 e nº2 para criar dashboards de temas específicos ou criar google sheets com envio da informação via email.

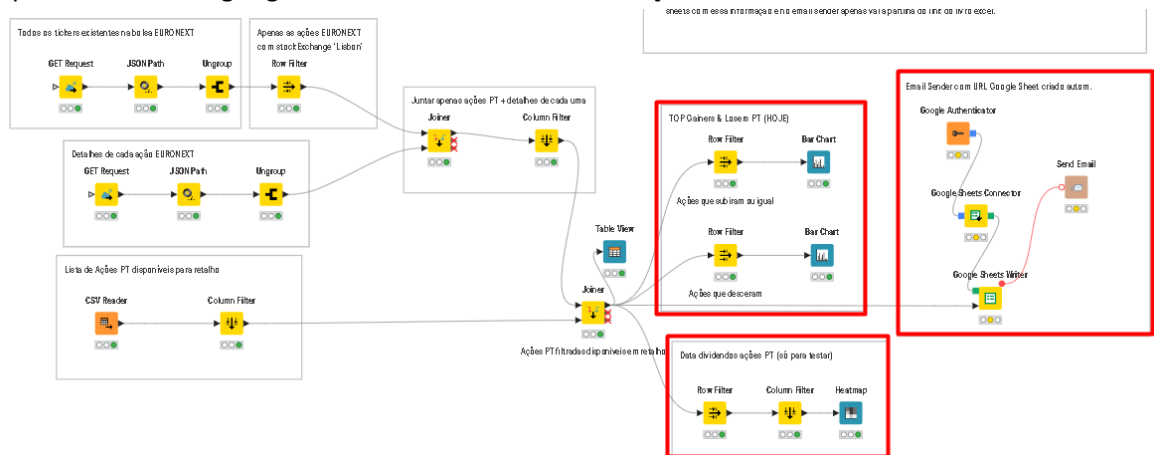


Figura 15 - Job nº3

Job nº4:

A partir de um CSV com várias ações globais criar dashboards interativos onde podemos aplicar os filtros que pretendemos e mostrar a moving average de cada stock ou então a comparação entre duas empresas com a análise do gráfico num intervalo de tempo.

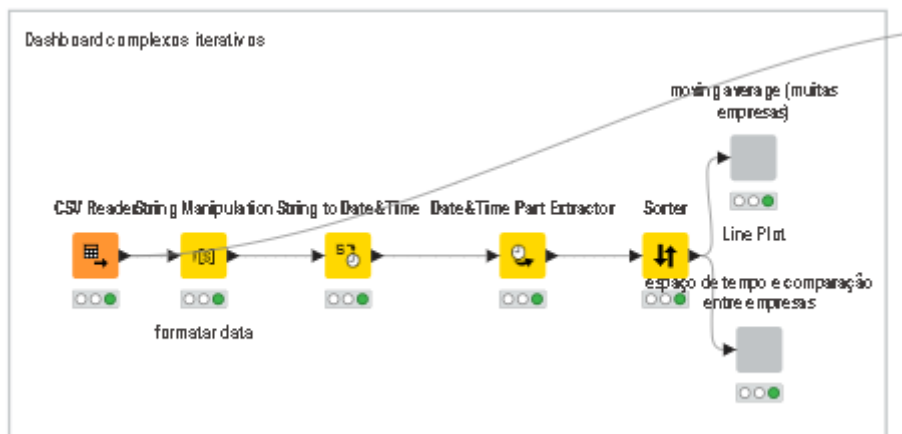


Figura 16 - Job nº4

Job nº5:

Utilização dos dados do CSV do Job nº4 para utilizar o símbolo da ação e criar um loop de Get Requests à API com a ticker correspondente no seu URL.

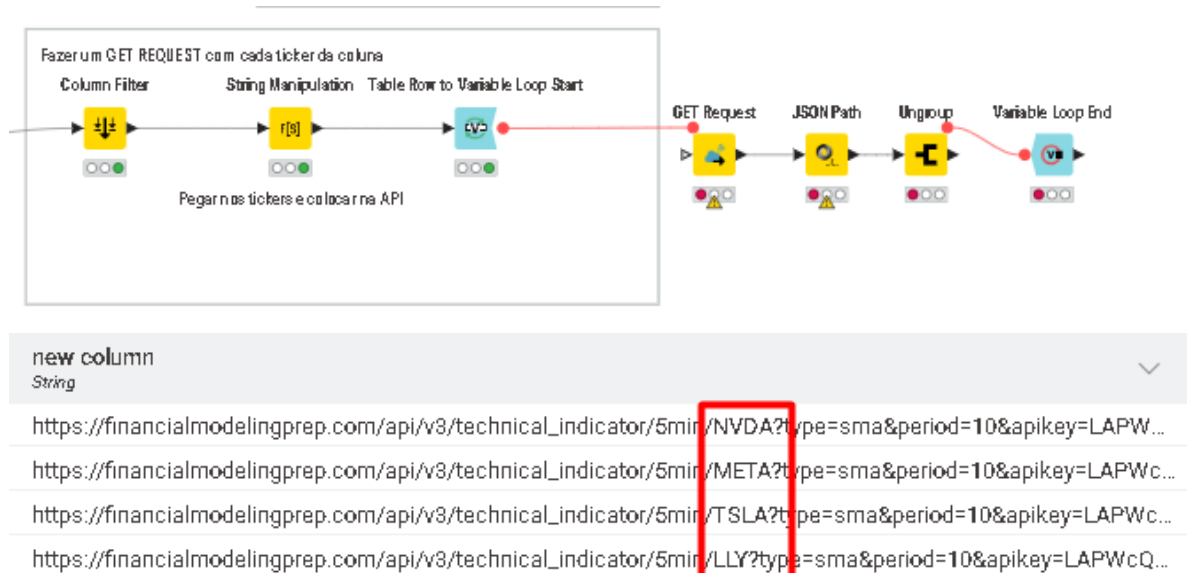


Figura 17 - Job nº5

Job nº6:

Outra solução com o mesmo objetivo do job nº5 e que ficou totalmente funcional, criando assim um loop de GET Requests por várias API's com stock distintos para obter dados em tempo real.

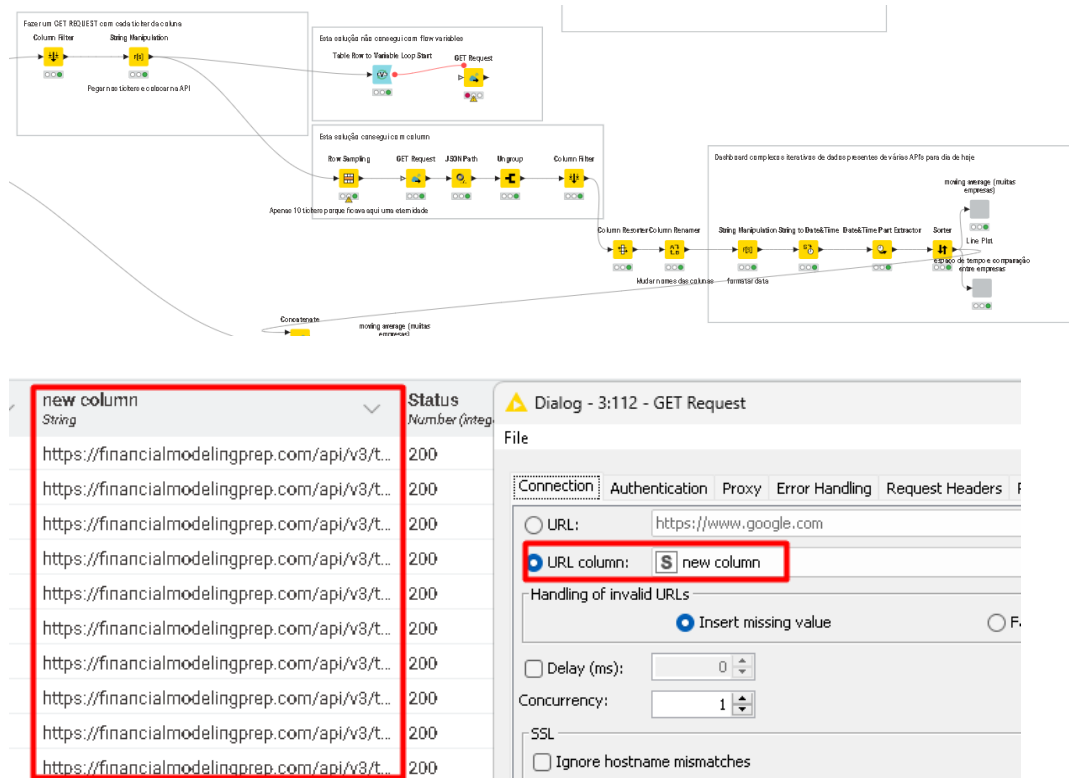


Figura 18 - Job nº6

Job nº7:

Criação de uma base de dados com o carregamento de todas as informações das ações da bolsa portuguesa e com os seus detalhes de forma automática no workflow.

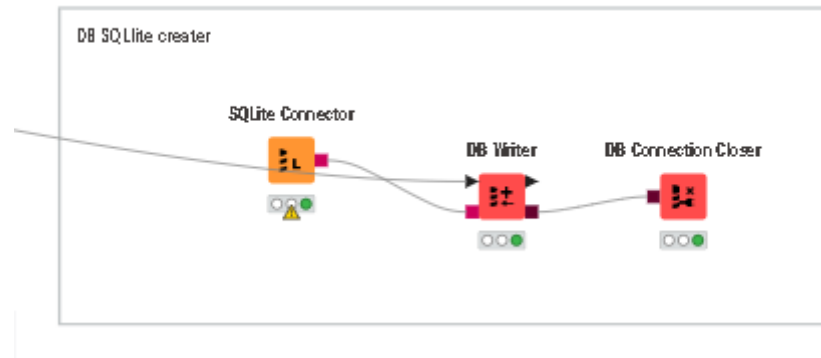
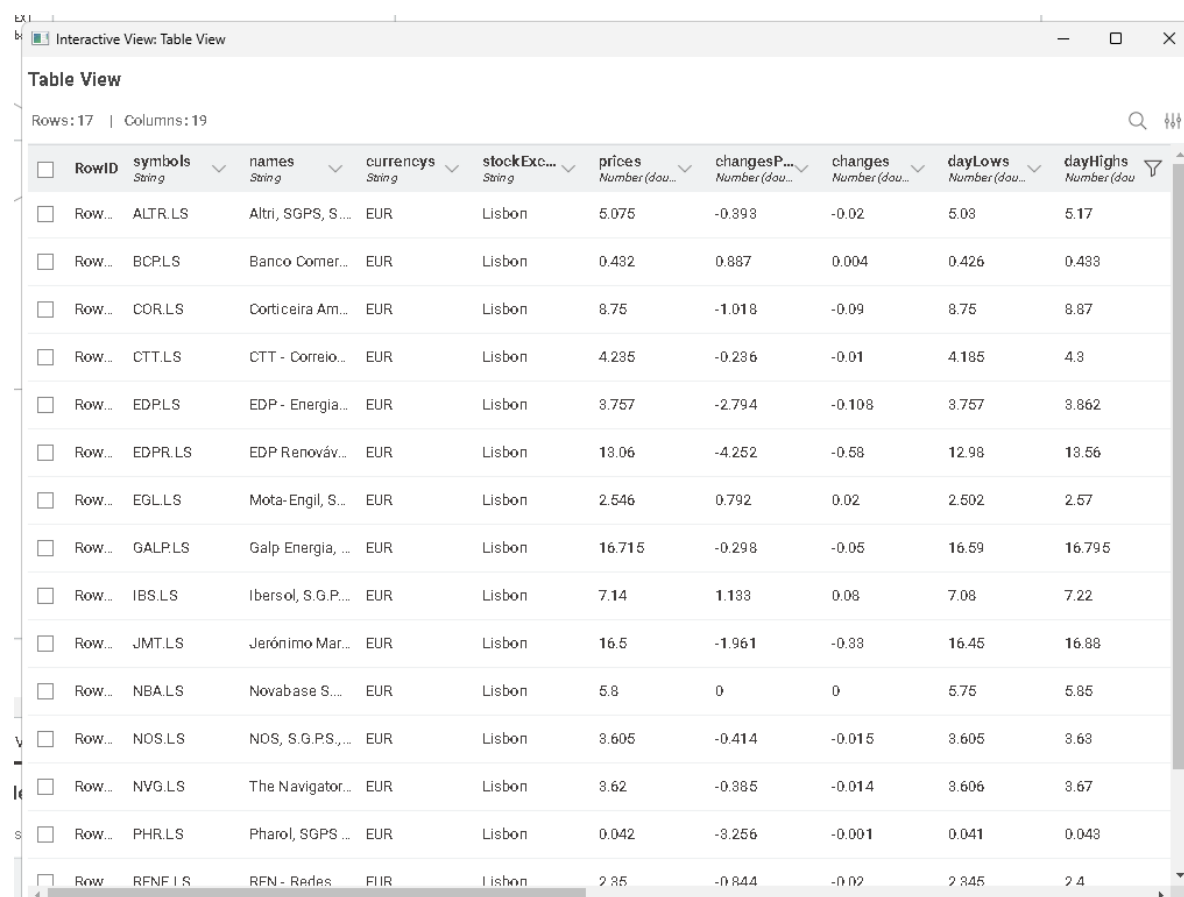


Figura 19 - Job nº7

Resultados

Consegui obter uma table view diretamente no KNIME com dados das ações portuguesas possíveis para compra e com todas informações necessárias de cada uma.



RowID	symbols	names	currencies	stockExc...	prices	changesP...	changes	dayLows	dayHighs
Row...	ALTR.LS	Altri, SGPS, S...	EUR	Lisbon	5.075	-0.393	-0.02	5.03	5.17
Row...	BCPLS	Banco Comer...	EUR	Lisbon	0.432	0.887	0.004	0.426	0.433
Row...	COR.LS	Corticeira Am...	EUR	Lisbon	8.75	-1.018	-0.09	8.75	8.87
Row...	CTT.LS	CTT - Correio...	EUR	Lisbon	4.235	-0.236	-0.01	4.185	4.3
Row...	EDPLS	EDP - Energia...	EUR	Lisbon	3.757	-2.794	-0.108	3.757	3.862
Row...	EDPR.LS	EDP Renováv...	EUR	Lisbon	13.06	-4.252	-0.58	12.98	13.56
Row...	EGL.LS	Mota-Engil, S...	EUR	Lisbon	2.546	0.792	0.02	2.502	2.57
Row...	GALP.LS	Galp Energia, ...	EUR	Lisbon	16.715	-0.298	-0.05	16.59	16.795
Row...	IBS.LS	Ibersol, S.G.P...	EUR	Lisbon	7.14	1.133	0.08	7.08	7.22
Row...	JMT.LS	Jerónimo Mar...	EUR	Lisbon	16.5	-1.961	-0.33	16.45	16.88
Row...	NBA.LS	Novabase S...	EUR	Lisbon	5.8	0	0	5.75	5.85
Row...	NOS.LS	NOS, S.G.P.S.,...	EUR	Lisbon	3.605	-0.414	-0.015	3.605	3.63
Row...	NVG.LS	The Navigator...	EUR	Lisbon	3.62	-0.385	-0.014	3.606	3.67
Row...	PHR.LS	Pharol, SGPS ...	EUR	Lisbon	0.042	-3.256	-0.001	0.041	0.043
Row...	RFNF.LS	RFN - Redes	EUR	Lisbon	2.35	-0.844	-0.02	2.345	2.4

Figura 20 - Resultados - TableView

Consegui obter dois gráficos distintos sobre o ganho/perda das ações portuguesas durante o dia de hoje.

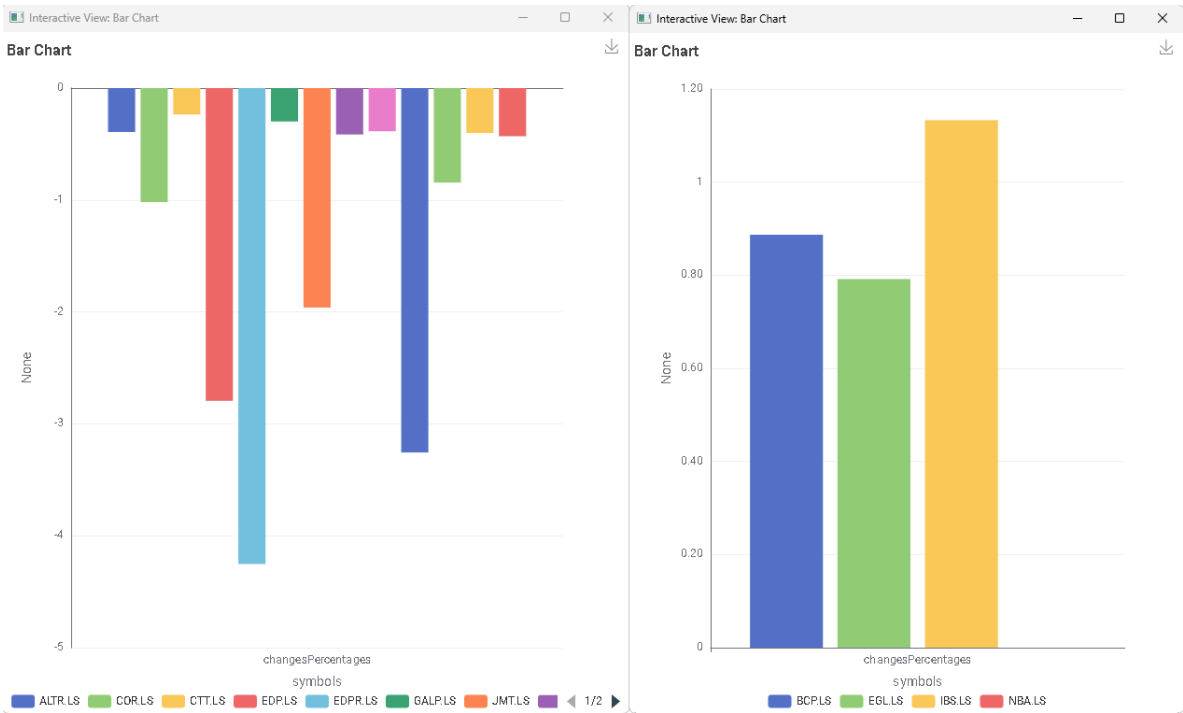


Figura 21 - Resultados - Ganho/Perda Stocks PSI20

Consegui obter à semelhança da tableview um ficheiro google sheets para poder ser partilhado com mais utilizadores.

Stocks PSI-20																	Partilha	
Ficheiro Editar Ver Inserir Formatar Dados Ferramentas Extensões Ajuda																		
Q Menu																		
100% 100%																		

Figura 22 - Resultados – GoogleSheet

Consegui criar uma forma de avisar um utilizador via email acerca do googlesheet já ter sido atualizado.

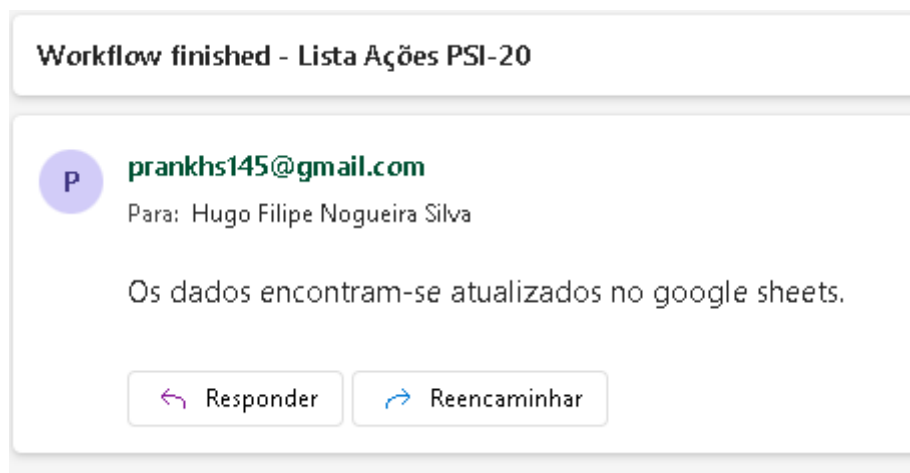


Figura 23 - Resultados – EmailSender

Consegui utilizar dados de um CSV com o histórico passado de cada ação e visualizá-los num gráfico iterativo com o cálculo da moving average de cada ação. E também um gráfico iterativo com a comparação entre duas ações distintas.



Figura 24 - Resultados - MA Ação

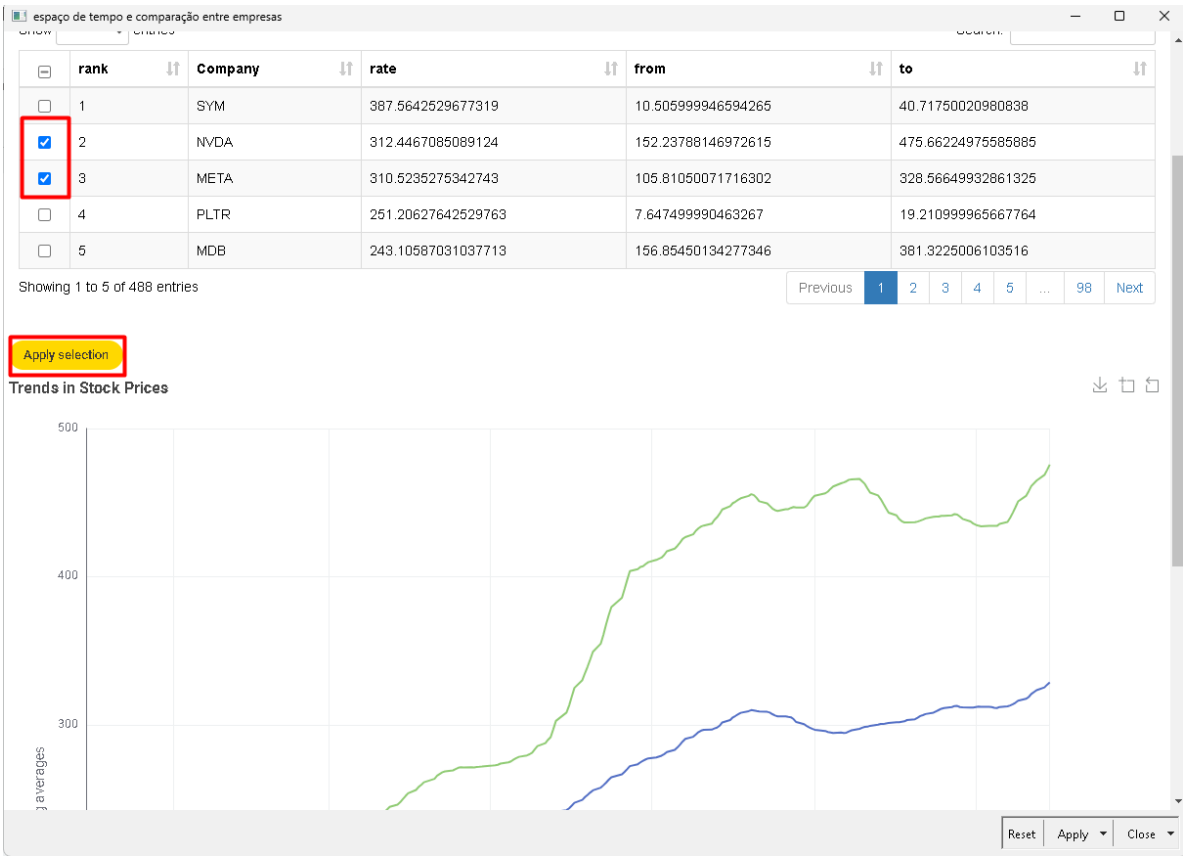


Figura 25 - Resultados - MA Comparativo entre Ações

Por último foi possível utilizar os gráficos acima mencionados, porém com dados atuais realizando várias chamadas a API's distintas de cada ação.

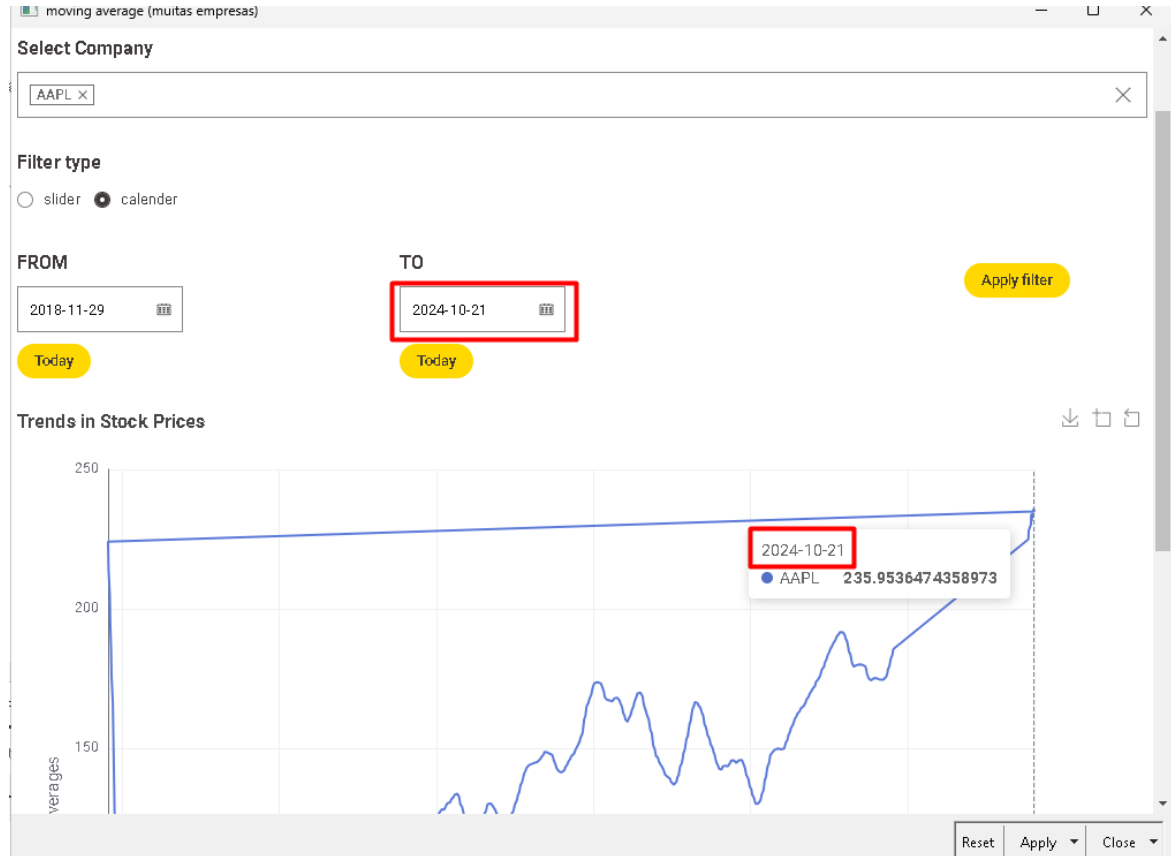


Figura 26 - Resultados - Combinação de API's para gráficos iterativos com dados atuais

Constrangimentos/Dificuldades

- **Loop no GET Request** - O KNIME oferece suporte parcial a streaming de dados, mas não nativamente no sentido de processamento contínuo de fluxo como algumas ferramentas especializadas em streaming (como Apache Kafka, Apache Flink ou Spark Streaming). Isto impede a minha ideia inicial de análise de dados específicos em tempo real conforme oscilação dos mercados financeiros.
- **Passar API's com variáveis de colunas para um GET Request através de flow variables** - Consegui realizar esta operação mas de outra forma, através das flow variables tive dificuldade porque o GET Request após uma chamada à API não modificava para o ticker seguinte.
- **Anexar ficheiro ao email sender** - Consegui guardar o ficheiro excel com caminho relativo, porém, quando é realizado um anexo de um ficheiro a um email é necessário ter caminho absoluto. Desisti da ideia e para experimentar mais temas criei um google sheets com essa informação e no email sender apenas vai a partilha do link do livro excel.

Conclusão e Trabalhos Futuros

Este trabalho prático permitiu aplicar de forma concreta as técnicas de ETL, utilizando a plataforma KNIME para a integração de dados financeiros provenientes de diversas fontes, como APIs públicas e ficheiros CSV. Através da criação de workflows automáticos, foi possível demonstrar a eficácia do processo de extração, transformação e carregamento de dados, produzindo gráficos e dashboards que facilitaram a visualização e análise de ações do mercado financeiro.

Para futuras iterações, várias melhorias podem ser consideradas. Uma delas é a integração de ferramentas de processamento de dados em tempo real, como o Apache Kafka ou Spark Streaming, que vão permitir a análise contínua das oscilações de mercado em vez de depender de dados estáticos passados ou atuais.

Bibliografia

API Financial Modeling Prep: <https://financialmodelingprep.com>

Yahoo Finance: <https://finance.yahoo.com>

Documentação do KNIME: <https://www.knime.com/documentation>