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

ESCOLA SUPERIOR DE TECNOLOGIA IPCA

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

Lista de Figuras

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.

Adicionalmente, exploraram-se diversas funcionalidades importantes de ETL, como a utilização de expressões regulares para tratamento de dados, operações de joins e filtros, bem como a integração de dados de diferentes origens (JSON e CSV) para posterior visualização e análise.

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.

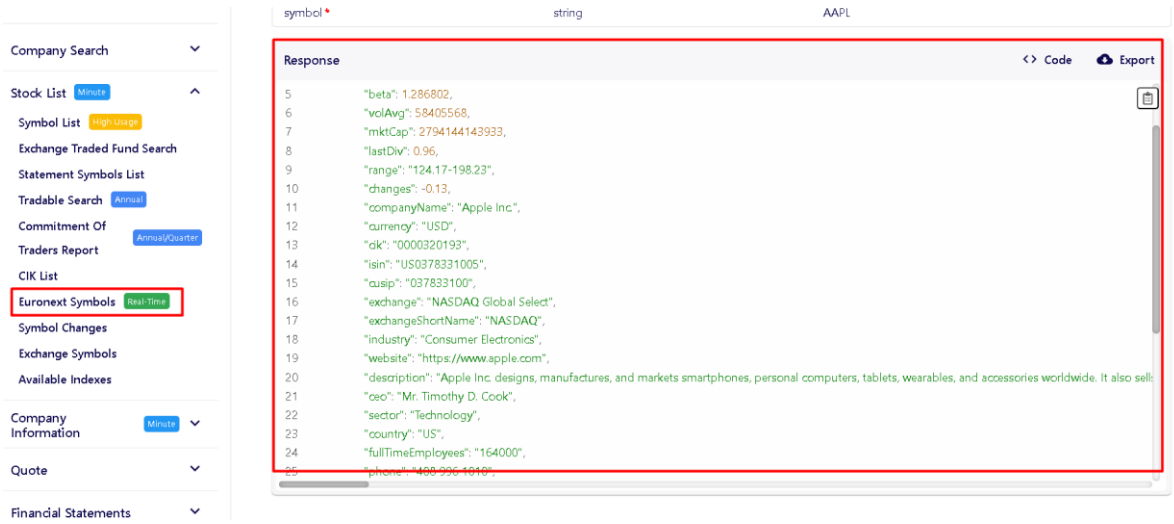
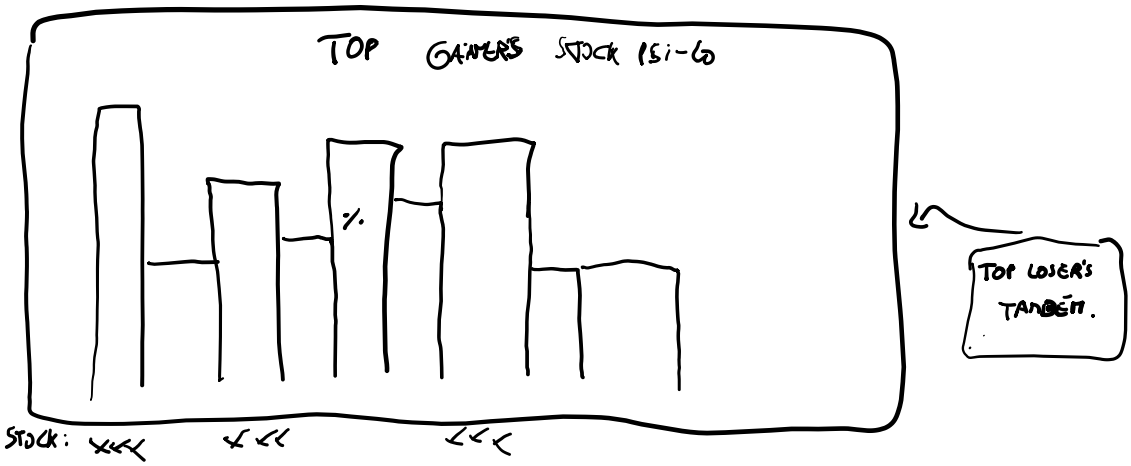
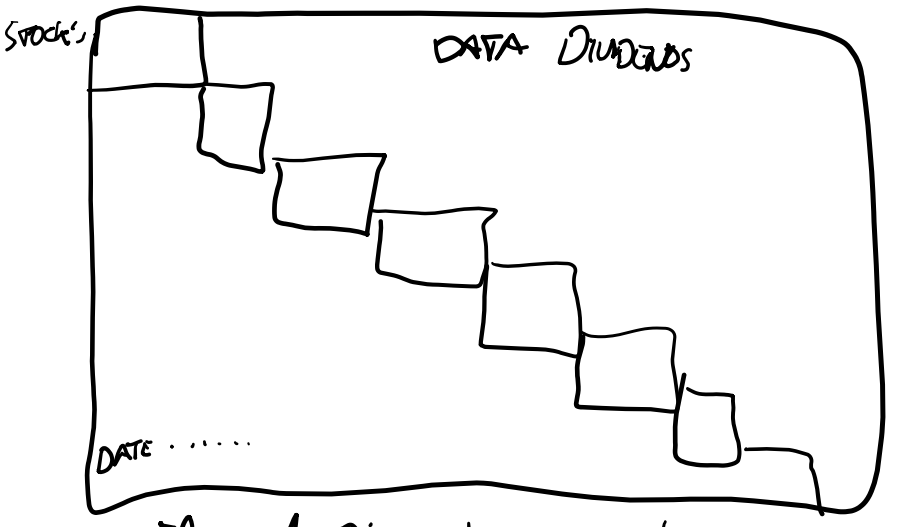


Figura 1 - Financial Company Profile API

Prototipagem





Fluxo de Dividendos para saber fluxograma de pagamentos.

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.

- **Visualização**: Geração de gráficos simples ou iterativos para identificar as ações que mais ganharam e perderam no dia etc.

Vídeo demonstrativo

Para complementar a explicação teórica do processo de ETL desenvolvido neste projeto, foi criado um vídeo demonstrativo que ilustra, de forma visual e prática, a implementação e execução dos workflows no KNIME. O vídeo mostra 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.

Figura 2 - 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:

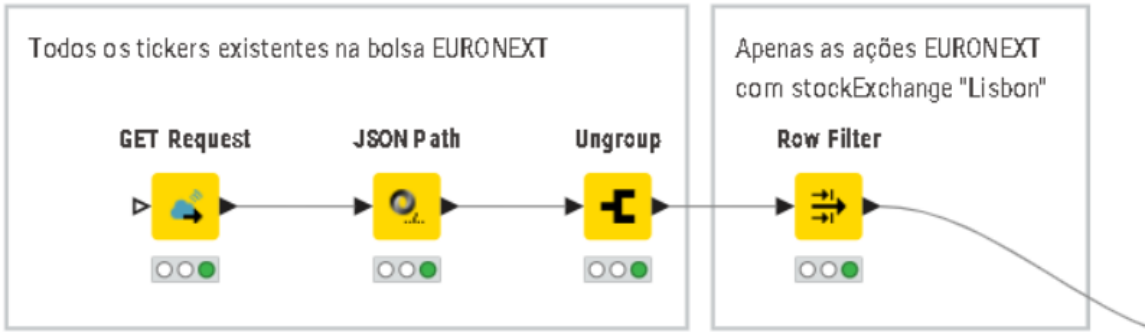


Figura 3 - Only Stocks Lisbon

Figura 4 - Details All Stock EURONEXT

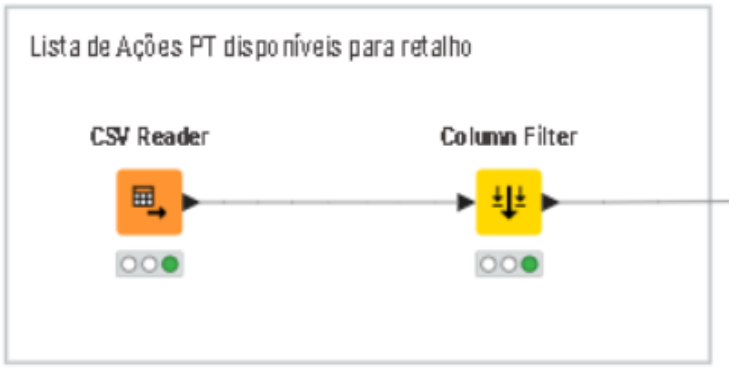


Figura 5 - Yahoo Finance CSV Stock Lisbon Retail

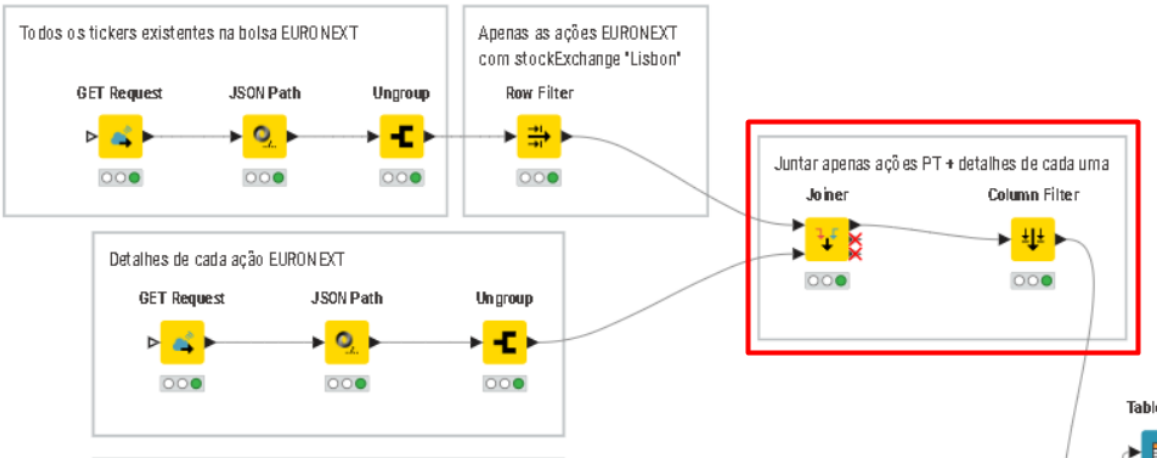


Figura 6 - JOIN Stocks Lisbon with details

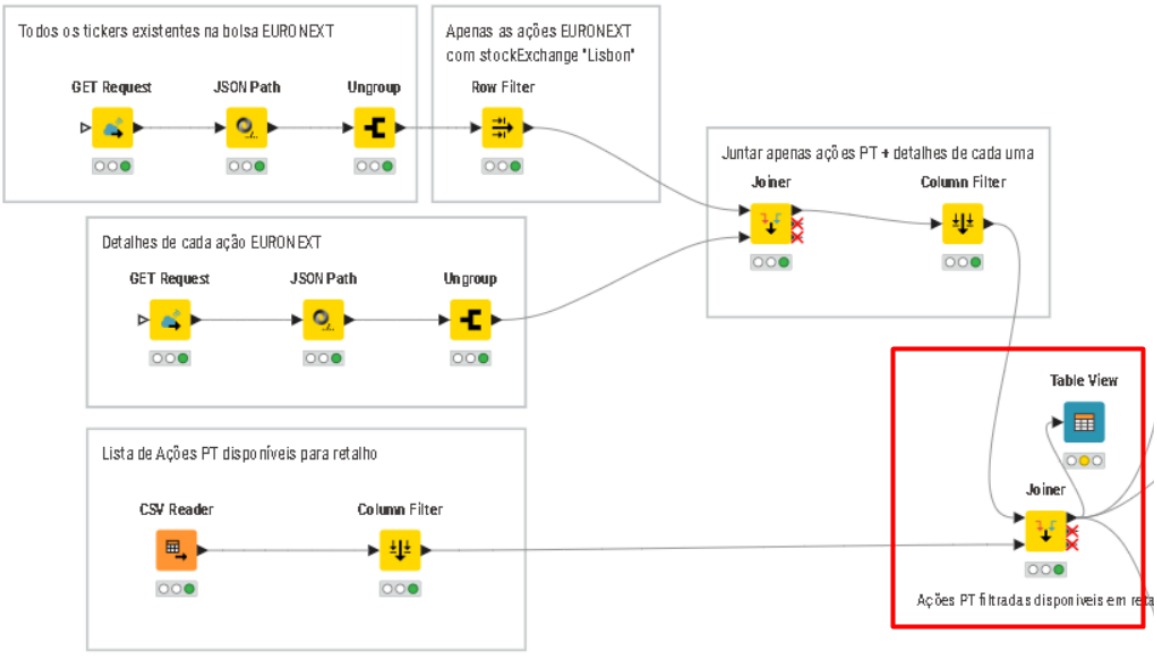


Figura 7 - JOIN Stocks Lisbon Retails with details

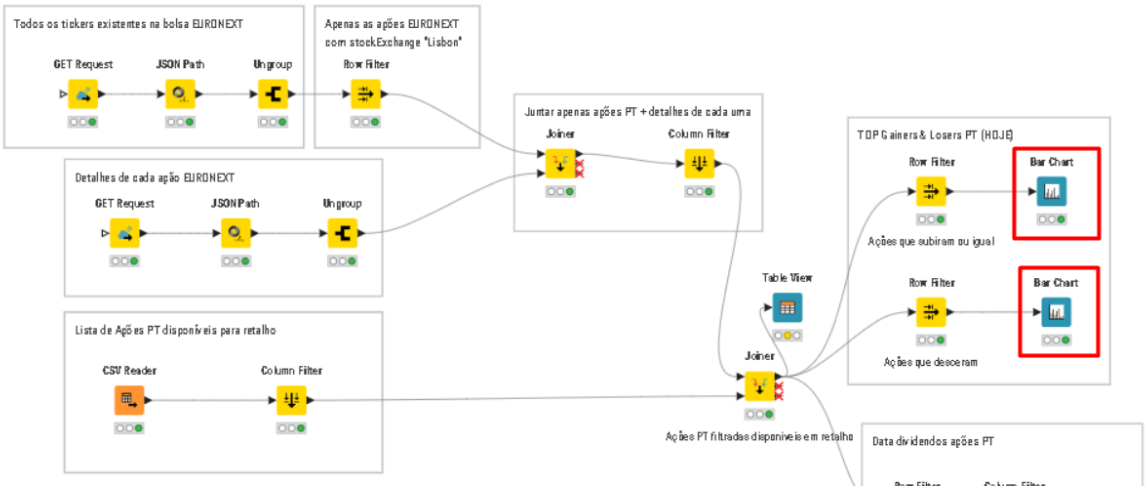


Figura 8 - Reports with Charts

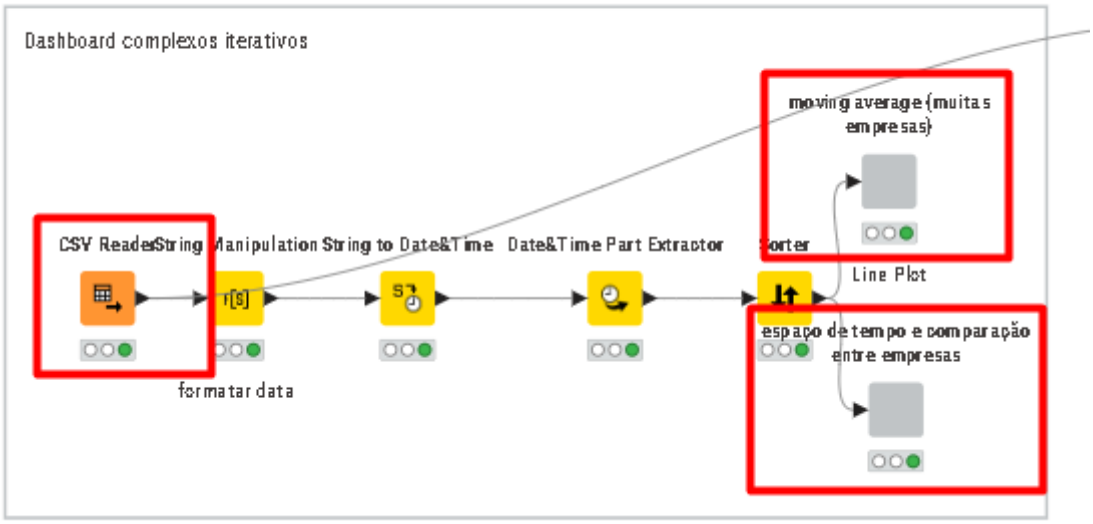
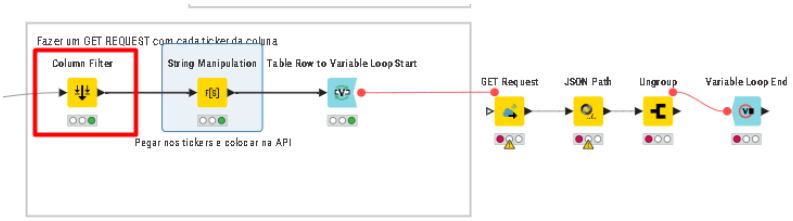


Figura 9 - Global Stocks to Iterative Dashboards with moving average



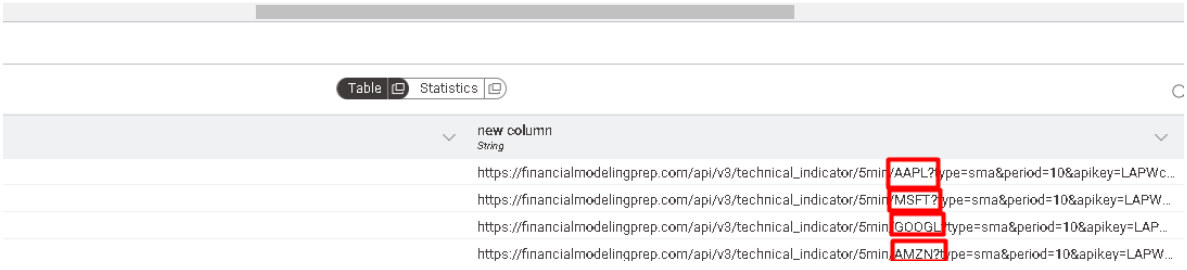


Figura 10 - Global Stocks with symbol to API Get Request

Explicação dos diagramas:

- **Extração de Dados**: Foi utilizado o nó GET Request para acessar a API de ações do PSI-20 e criados ficheiros Excel com uma listagem de todas as informações de cada ação do PSI-20 com dados de uma API externa e de um CSV.

- **Filtros e Joins**: Aplicou-se filtros com base na coluna "Lisbon" porque a API devolve todas as ações existentes na bolsa europeia (EURONEXT) e foi feita a combinação com os dados de ações detalhadas usando um Inner Join para conectar as tabelas por símbolo de ação.

- **Transformações de Dados:** Limpeza e normalização de dados utilizando funções de agregação e manipulação (concatenar, *calcular variações*, etc). Calcular variações não chegou a ser utilizada, mas foi explorada essa ferramenta para calculo do change price diário de cada ação.

- **Criação de dados falsos e limpeza:** Foi criado um fluxo com dados fictícios para testar a limpeza dos dados (figura nº9).

Jobs

Explicação

Job nº1:
Juntar todas as ações portuguesas extraídas da bolsa europeia e com detalhes de cada uma.

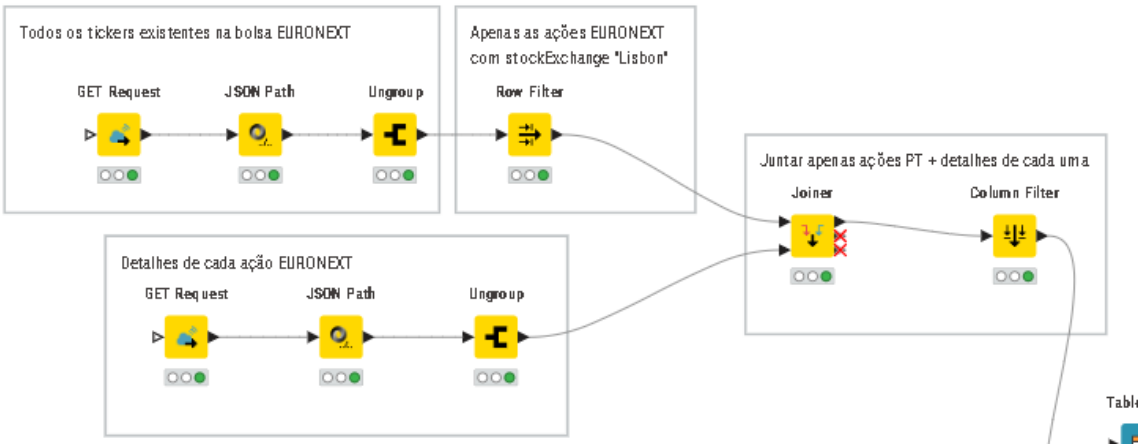


Figura 11 - Job nº1

Job nº2:
Juntar todas as ações portuguesas com detalhe anteriormente conseguidas no Job nº1 mas agora filtrando apenas pelas ações disponíveis em retalho.

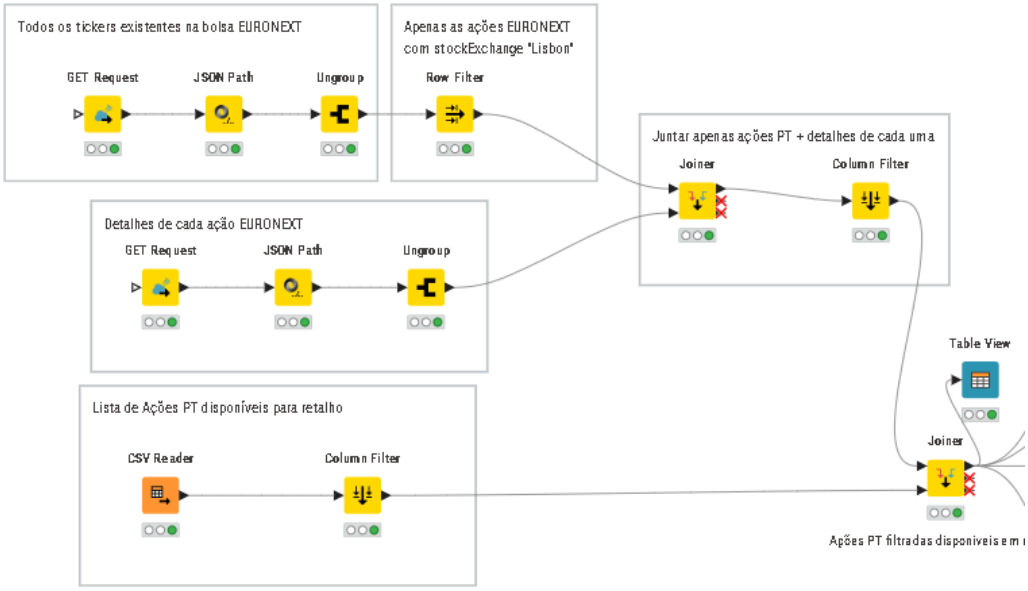


Figura 12 - Job nº2

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.

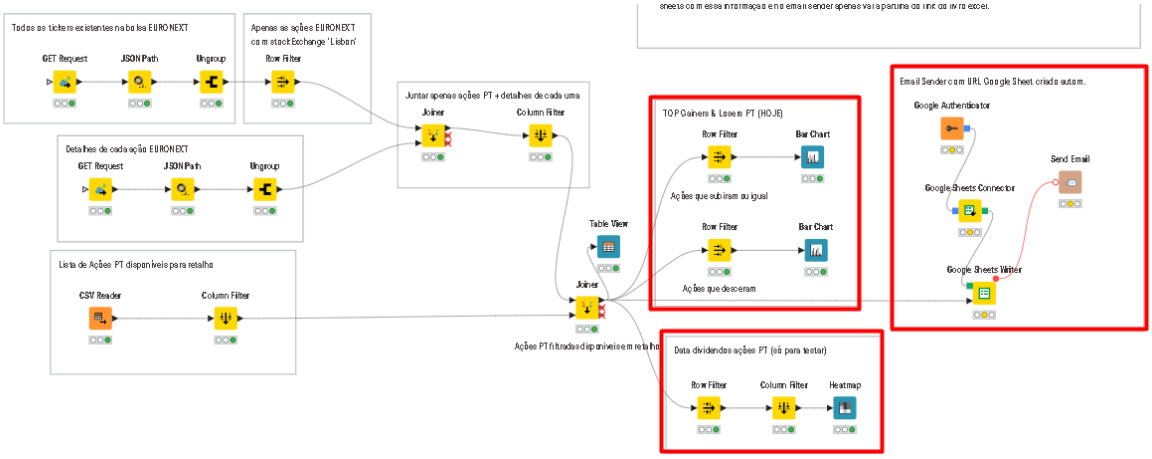


Figura 13 - Job nº3

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.

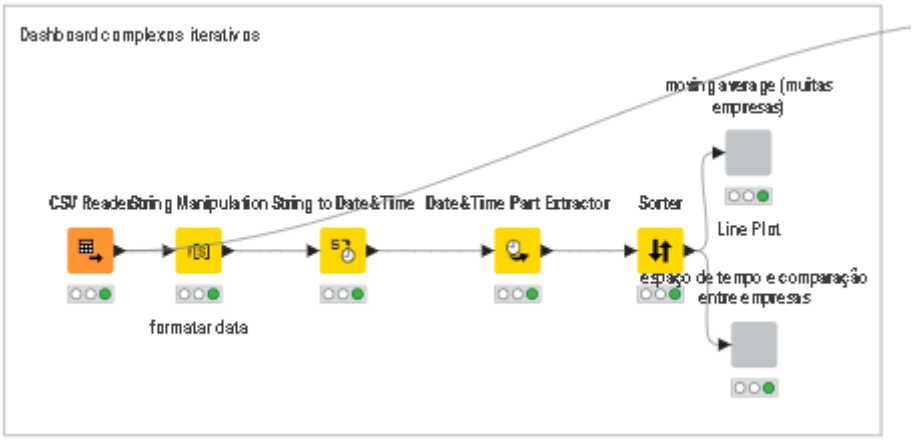


Figura 14 - Job nº4

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.

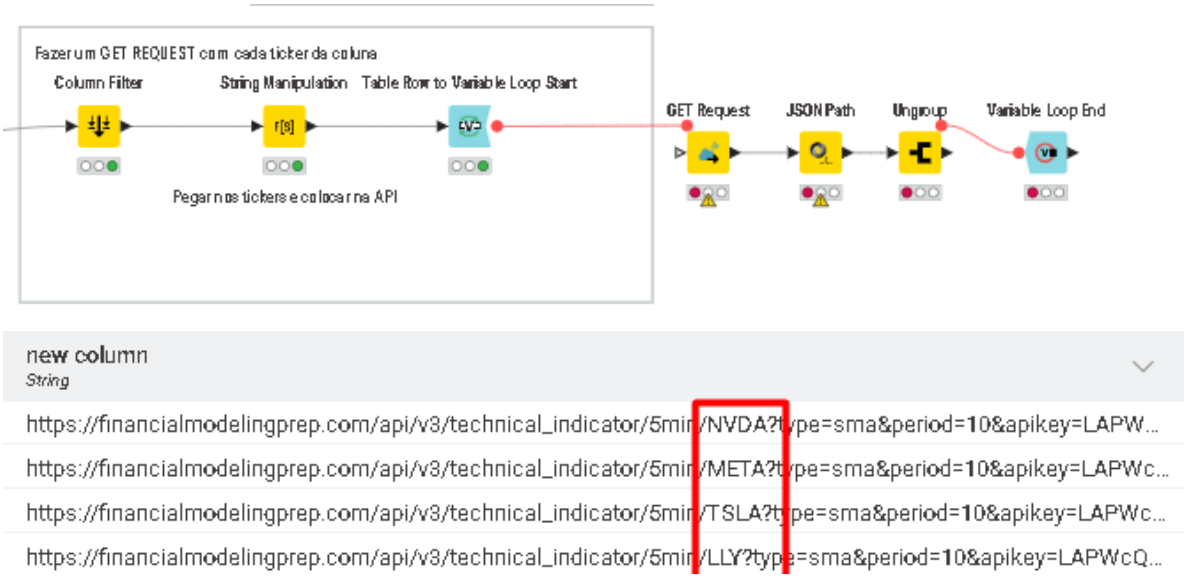


Figura 15 - Job nº5

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 feitas com variáveis de colunas e tentar fazer vários GET Request - Consegui a parte da criação das diferentes API's mas a parte de associar ao GET Request está mapeada porém tem algum erro que desconheço e impede funcionamento.
- 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 permitiriam a análise contínua das oscilações de mercado em vez de depender de dados estáticos. Outro ponto a explorar é o aperfeiçoamento do loop de GET Requests, com o intuito de automatizar completamente o processo de extração de dados atualizados sem falhas.

Bibliografia

API Financial Modeling Prep: *https://financialmodelingprep.com*

Yahoo Finance: *https://finance.yahoo.com*

Documentação do KNIME: *https://www.knime.com/documentation*