ISI

Integração de Sistemas de Informação

Trabalho Prático 1

Ano letivo 2024/25

Licenciatura em Engenharia de Sistemas Informáticos

Escola Superior de Tecnologia

Instituto Politécnico do Cávado e do Ave

Índice

Enquadramento

Este trabalho será realizado no âmbito da disciplina de Integração de Sistemas de Informação (ISI) da Escola Superior de Tecnologia do IPCA. A integração de sistemas de informação é uma área fundamental na atualidade, uma vez que a gestão eficiente de dados é crucial para a competitividade das empresas. Este projeto será desenvolvido por, Pedro Silva, que têm como objetivo explorar as boas práticas e ferramentas de ETL (Extract, Transform, Load) em aplicações low code.

O foco deste trabalho é a aplicação de processos de ETL para melhorar a análise e o uso dos dados existentes numa empresa de produção. Através da agregação, limpeza e armazenamento adequados dos dados, pretendemos demonstrar como a informação pode ser transformada em conhecimento valioso para a tomada de decisões. Este estudo não só irá abordar questões técnicas relacionadas com a integração de dados de diferentes formatos (CSV, Excel, webserver, json e XML), mas também discutirá a importância da anonimização dos dados para garantir a conformidade e a segurança das informações processadas.

Problema

A empresa de produção de bobines enfrenta um dos desafios mais comuns no mundo atual: lidar com a grande quantidade de dados gerados durante o processo produtivo. Embora esses dados possam ser uma mais-valia para a empresa, na realidade, frequentemente permanecem parados e sem utilidade prática. Isso ocorre devido à falta de processos adequados para a extração, tratamento, e análise dos dados.

Os dados são registados em diferentes formatos: cada máquina de produção mantém um ficheiro Excel para o registo da produção, enquanto o embalamento é documentado em ficheiros CSV, armazenados nas pastas dos operadores de embalamento. Essa dispersão de dados torna difícil obter uma visão consolidada da produção e do desempenho dos operadores. Além disso, a empresa precisa ser capaz de identificar rapidamente anomalias, como taras de peso superiores a um determinado limite, para garantir a qualidade do produto, e em caso de problema produtivo, precisa de ter a informação se ainda tem produtos dentro das suas instalações.

É imperativo que esses dados sejam agregados, limpos e armazenados de forma que possam ser utilizados para análises que apoiem a tomada de decisões. O objetivo final é transformar esses dados parados em informações úteis que possam ajudar a definir indicadores-chave de desempenho (KPIs) e otimizar processos, melhorando assim a eficiência e a rentabilidade da empresa.

Estrategia utilizada

A estratégia para esta integração envolve todos os passos ETL, engloba extração de ficheiros Excel, CSV, JSON, XML GET de informação na WEB e também em base de dados. Para atingir este objetivo o KNIME foi a ferramenta escolhida, é um programa open source, intuitivo e compatível com vários sistemas operativos.

Extração / Extract

É essencial entender os dados que nos apresentam, apenas após entender qual o seu propósito é podemos fazer a arquitetura de todo o projeto. Foram utilizados vários nodes para a extração, tais como, CSV Reader, XML Reader, GET Request, entre outros.

Transformação / Transform

Foram necessárias várias operações aos dados extraídos. Foram feitas filtragens, agregações e limpeza. Desde conversões de datas, manipulações de texto com recurso a expressões regulares ou anonimização de dados, foram explorados o maior número de recursos que o KNIME tem para oferecer.

Carga / Load

Por fim, foi feito um dashboard que permite a visualização gráfica de toda a informação, mas também foi explorada a possibilidade de enviar dados para um serviço de google sheets e envio de email.

Foi este o fluxo que permitiu uma integração eficiente de dados que inicialmente estavam dispersos para um local central, de fácil visualização facilitando a tomada de decisões.

Transformações

As transformações foram fundamentais para garantir que os dados provenientes de vairos ficheiros fossem adquados para a carga e consequente detalhe necessário para a tomada de decisão.

Devem sempre cumprir boas práticas, para ter um fluxo eficiente.

Diagramas

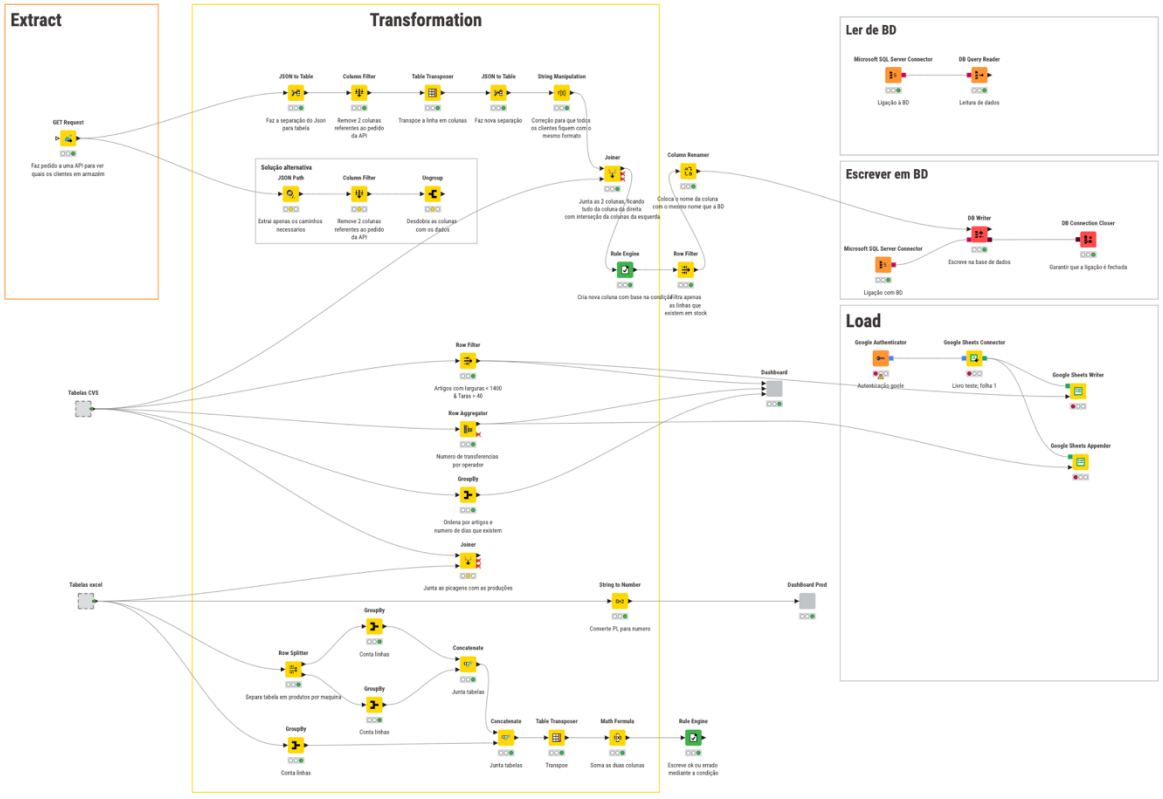


Figura 1 Imagem global

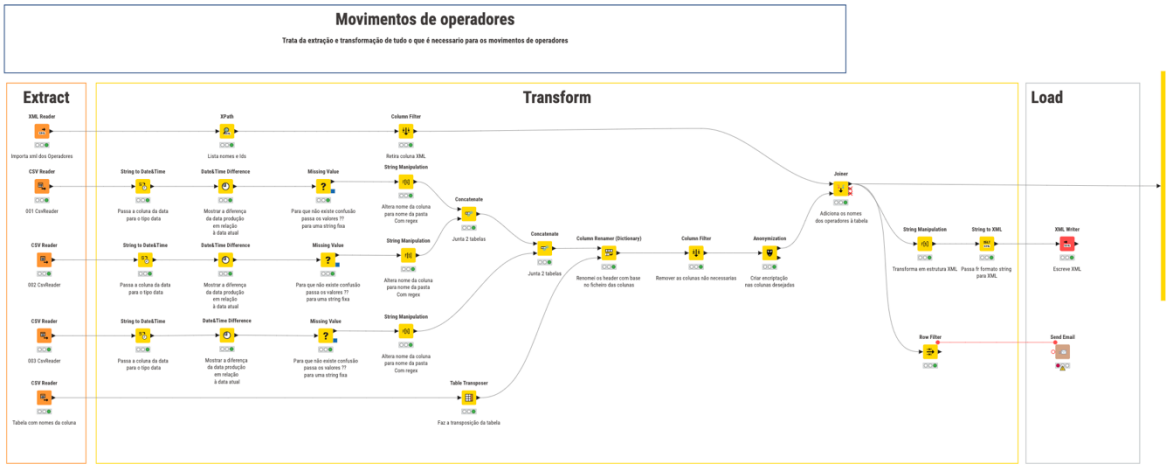


Figura 2 Tabelas CSV

Figura 3 Tabelas excel

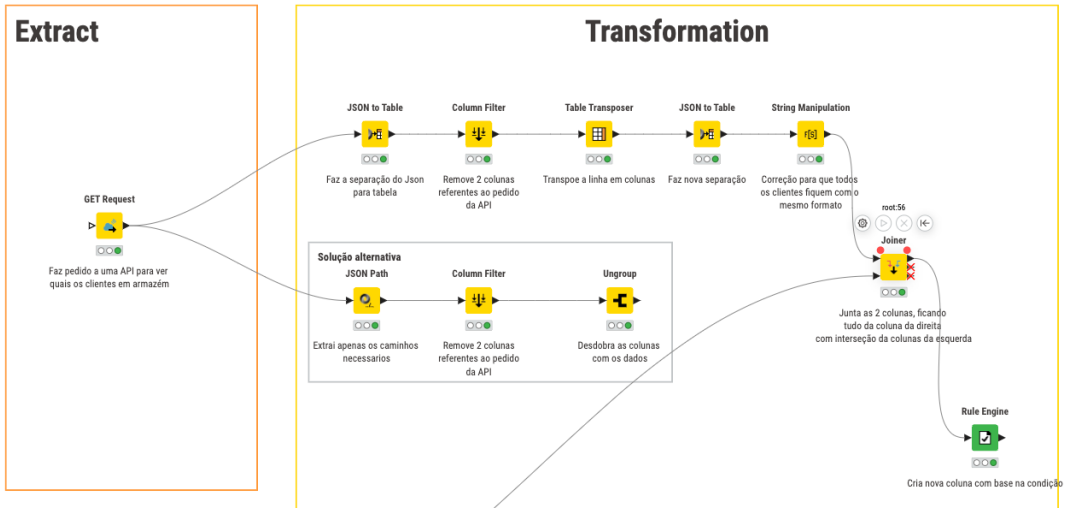


Figura 4 - Get Request

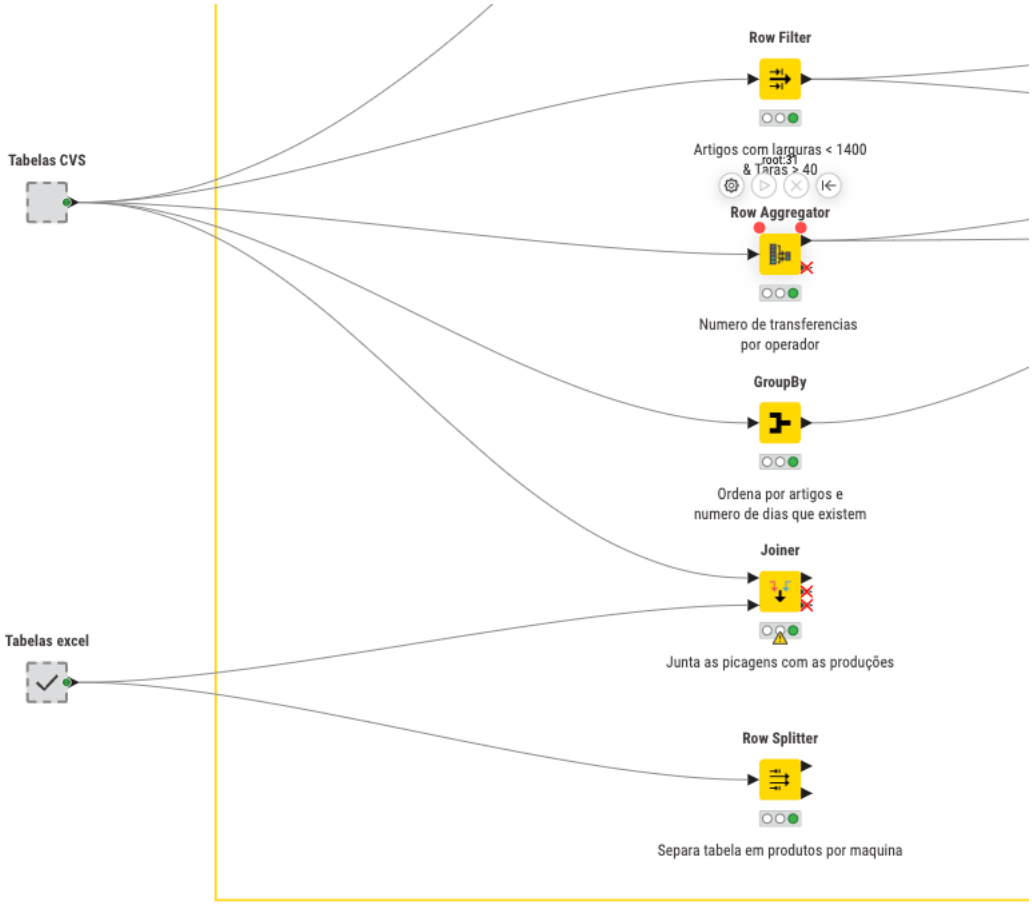


Figura 5 - Transformações para DashBoard

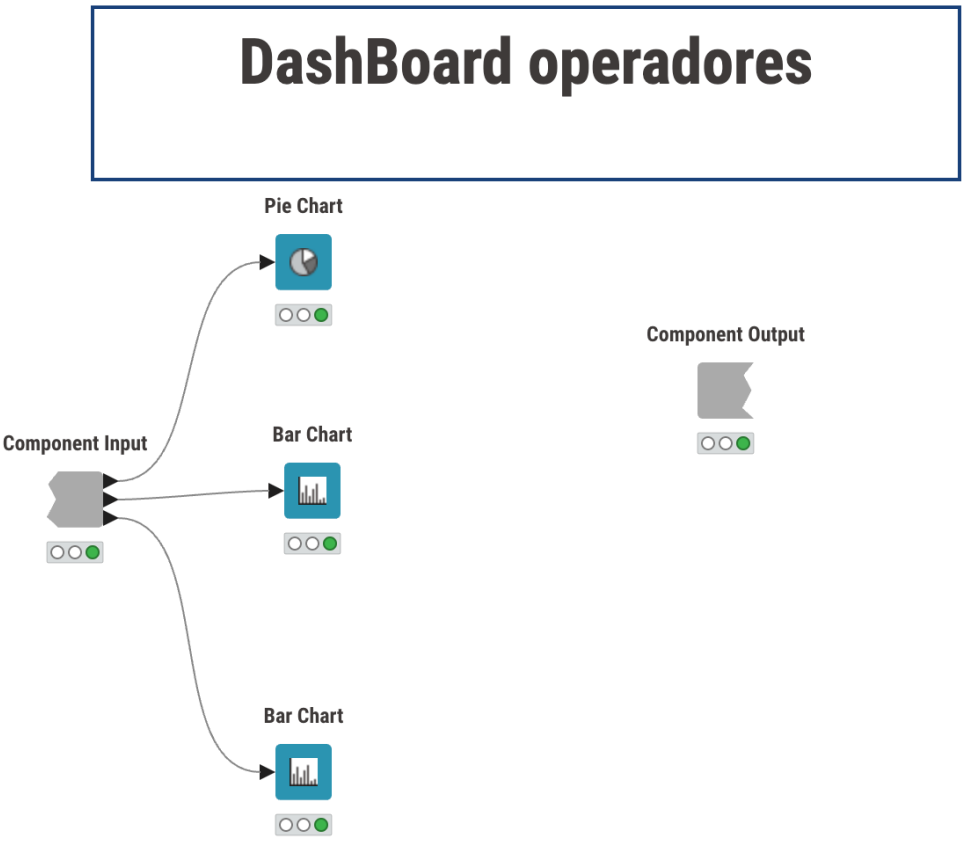


Figura 6 - DashBoard

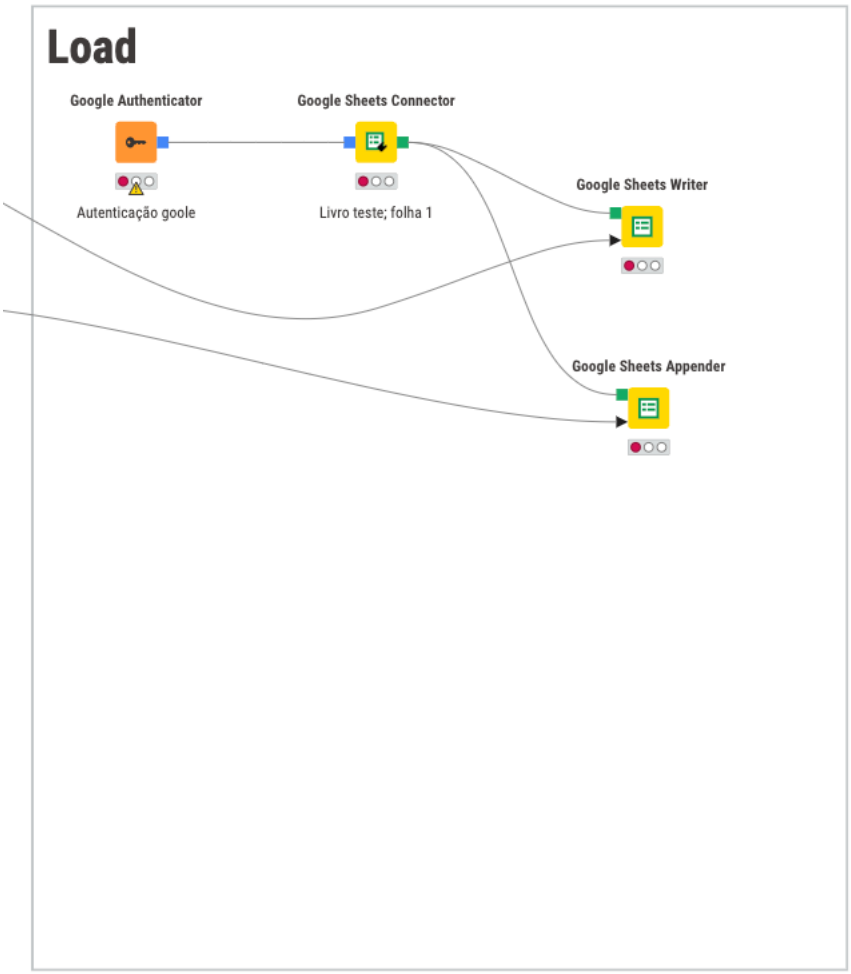


Figura 7 - Google sheets

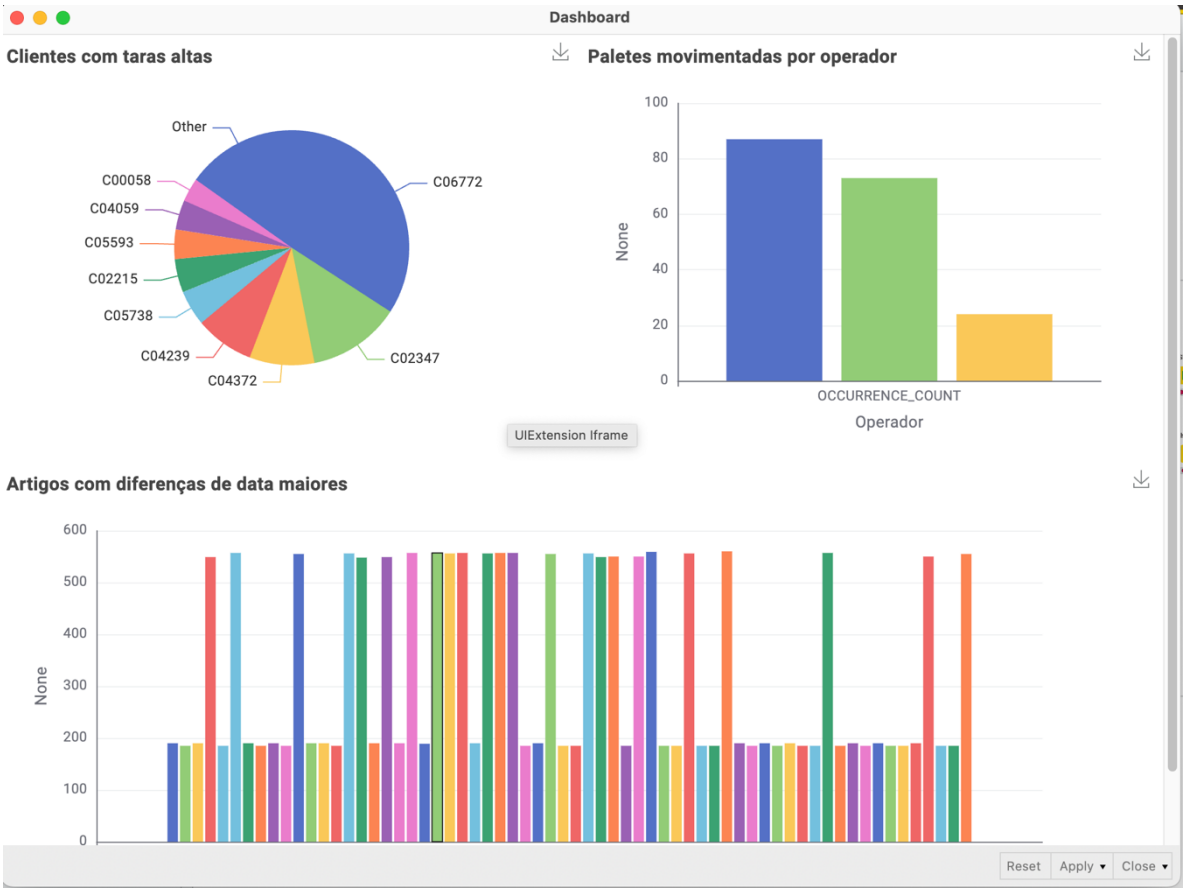


Figura 8 - Resultado DashBoard

Explicação

Figura 1

É a imagem global, onde conseguimos o início e o final do fluxo de dados.

Figura 2 – Tabelas CSV

Aqui temos três tipos de extração de dados, seguem numerados para facilitar as suas transformações.

1. Extração de vários ficheiros CSV na mesma pasta.
2. Extração de apenas um ficheiro CSV
3. Extração de um ficheiro XML

No caso 1,a transformação tem um pormenor importante de salientar, pois será importante para a transformação, a extração incluiu a ativação da opção “Path

column”, esta opção permite que seja adicionada uma nova coluna com o caminho da origem de dados.

A sua transformação passou pela conversão de uma coluna com dados de texto para datas, de seguida foi criada uma coluna com a diferença de datas, entre a coluna existente e uma data inserida. Depois foram analisadas e corrigidas todas as string que não tinham valor, para um valor fixo, de forma que ficassem completamente identificados. De seguida foi utilizada uma das mais poderosas ferramentas, na minha opinião, do KNIME, o node **String Manipulation**, permite vários recursos para a manipulação de strings, com várias funções integradas como por exemplo a contagem de caracteres, verificar o comprimento de strings, transformação em maiúsculas, entre muitas outras, neste caso escolhi a função regexReplace, que permite a integração de uma expressão regular para substituir um determinado valor. Na própria aplicação tem uma explicação de como utilizar a função:
“

Applies regex to string and replaces str if regex matches.

Examples:

regexReplace("abc", "[a-zA-Z]{3}", "cba") = "cba"
regexReplace("aBc", "[a-zA-Z]{3}", "AbC") = "AbC"
regexReplace("abcd", "[a-zA-Z]{3}", "ABC") = "ABCd"

”

Neste caso pretendia identificar o número de operador, aproveitando a coluna, acrescentada na extração do ficheiro.

A expressão regular que me garantiu uma boa extração foi a seguinte

“

regexReplace($Operador$, ".*/([^/\\\\]+)[/\\\\][^/\\\\]+$", "$1")

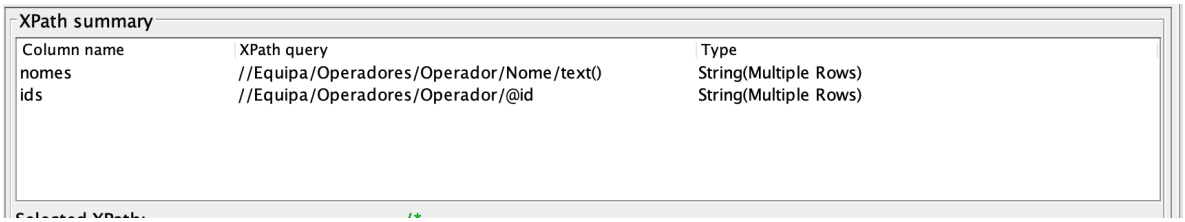
“

Depois fui concatenando os ficheiros de todos os operadores de forma ter tudo na mesma tabela.

A extração 2, como os dados não tem colunas criei este ficheiro com a informação das colunas do tipo de extração 1.

Com estes dois tipos de dados extraídos precisei apenas de fazer a transposição da tabela e utilizar o node **Column Renamer**, que permite a rescrever o nome das colunas identificando a tabela pretendida. De seguida fiz a exclusão de colunas que são necessárias, e por fim, um dos passos mais interessantes e, em muitos casos, essenciais, a anonimização de dados. Após algumas pesquisas encontrei, na KNIME Community Hub, este node que permite fazer encriptação utilizando, por exemplo, salting de uma coluna já existente na tabela, permite também ter consistência nos dados, muda o seu valor, mas mantem a sua característica, por exemplo, um artigo quando anonimizado, é substituído sempre pelo mesmo valor anonimizado (isto se não tiver valor de salting ou mantendo sempre o mesmo em cada linha da tabela).

XML foi criado, juntamente com o seu DTD, contem informação sobre os operadores, numero e nome. Esse é a extração 3, que utilizai o **XML reader**, quando é utilizado, todo o XML fica na mesma linha da tabela, portanto, é necessário utilizar outra ferramenta para conseguir passar a informação desejada em várias colunas e linhas, para isso utilizei o **XPath**, com recurso à expressões **XPath** criei duas colunas para os nomes e identificação dos operadores.



De seguida fiz a juncão dos dados 1 e 2, já concatenados, com os dados 3. O objetivo é criar uma coluna, com o nome do operador mediante a sua identificação. Para isso utilizei o node **Joiner**, que recebe input de duas tabelas faz a sua junção. Após a junção pela pelos ids dos operadores crio um ficheiro XML para guardar no disco, e ao mesmo tempo, faço um filtro para encontrar algum operador não registado, ou algum tipo de erro, caso existam é enviado um email a alertar. Simultaneamente fecho o metanode.

Figura 3 – Tabelas de excel

Foi feita a extração em excel para de seguida foi feita a remoção de colunas que não são necessárias.

De seguida foi efetuada a anonimização aos dados pretendidos, seguido de uma separação de linhas com datas inválidas e datas corretas, as linhas com datas válidas seguem o fluxo normal, as restantes ficam identificadas para futuras correções e/ou validações do cliente.

Depois é feita a passagem de string para data e a utilização do node string manipulation com recurso a uma expressão regular para conseguir retirar o nome da máquina.

Por fim a concatenação e fecho do metanode.

Figura 4 - Get Request

O pedido à API lista todos os clientes existentes em armazém, de seguida cruzar os dados para saber se os clientes da de qualquer uma das tabelas existem ou não em armazém.

Neste diagrama demonstro 2 caminhos possíveis para chegar ao mesmo resultado.

A resposta do webserver é dada, como na maior parte dos pedidos, num ficheiro json, para fazer o tratamento ao ficheiro json, utilizei o node Json to table, que faz a divisão do primeiro nível do ficheiro, de seguida removi as colunas que não eram necessárias. Como o node Json to table, faz a divisão do json em várias colunas, utilizei o node Transpose para ter os dados organizados em linhas. De seguida utilizei novamente o node Json to table para separar os dados em 2 colunas. Como os dados não da coluna de cliente não são iguais fiz uma manipulação de string para colocar o prefixo “C0” antes de linha na coluna de cliente.

Por outro lado, encontrei outra solução possível, após o get request utilizei o json path para de imediato os dados do json, depois de retirar as colunas não necessárias, foi utilizado o node Ungroup para desagrupar a informação que estava na mesma linha.

Depois com o Joiner juntei os dados provenientes do metanode Tabelas CSV, e utilizei o node Rule Engine para fazer a verificação se o cliente existe ou não em armazém.

No node Rule Engine, podem ser utilizados operadores lógicos e até condições, o seu uso é similar a outras linguagens.

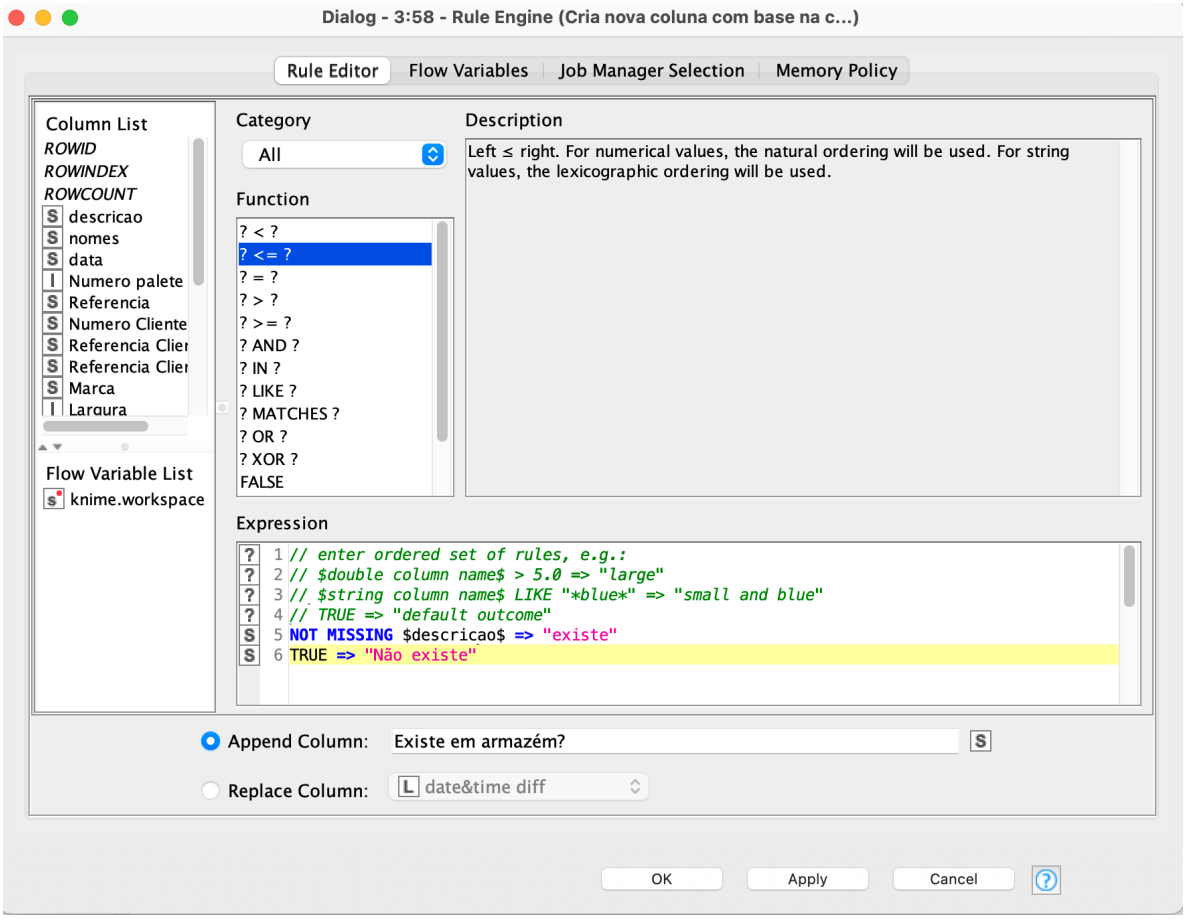


Figura 9 - Rule Engine

Figura 5 – Tranformações para DashBoard

Foram feitas algumas transformações, para serem mostradas no dashboard, simples de alguns exemplos de condições e filtragens.

Figura 6 - DashBoard

O KNIME dá a possibilidade de criar componentes, desde que sejam do mesmo tipo, neste caso vários gráficos que podemos juntar num só, criando um dashboard interativo com possibilidade de fazer deploy numa pagina web.

Com os itens preparados do ponto anterior, fiz 3 graficos para conseguir ter uma visão global sobre o problema e a sua possível solução.

Figura 7 – Google sheets

Simultaneamente à inserção dos dados no dashboard, os dados são enviados para um ficheiro no google sheets, primeiro foi utilizado o node Google Authenticator depois foi efetuada a ligação com o serviço externo e por fim a escrita num ficheiro novo e também o append a um ficheiro existente.

Conclusão e trabalhos futuros

Neste trabalho, foi apresentada a integração de dados utilizando a plataforma KNIME para realizar o processo de ETL com ficheiros em vários formatos, como CSV, Excel, JSON e XML vindos de varias fontes, web, base de dados em sql ou no próprio disco.

O uso do KNIME revelou-se uma solução eficaz, intuitiva e robusta para lidar com grandes volumes de dados, realizando transformações complexas, como anonimização de dados e manipulação de strings através de expressões regulares.

A capacidade de gerar dashboards interativos, possibilitando uma análise visual dos dados, foi fundamental para facilitar a interpretação e tomada de decisões. Além disso, o uso de serviços externos, como o Google Sheets e a integração de APIs REST, provou ser uma mais-valia, permitindo a sincronização contínua de dados e uma comunicação eficiente entre sistemas.

Com o processo de ETL devidamente implementado, foi possível garantir que os dados fossem limpos e organizados de forma centralizada, assegurando que toda a informação está atualizada e pronta para ser utilizada na análise e nos processos de decisão.

Mesmo sendo um confesso fã de programação, vejo o imenso potencial nesta ferramenta de low code. Acredito que, para muitos problemas, pode ser uma excelente solução, na sua simplicidade, eficiência, suporte existente e rapidez de execução O KNIME demonstrou que pode facilitar a vida dos programadores e analistas, automatizando processos complexos de forma visual e acessível.

Fico com mais um ferramenta e garanto que a vou utilizar.

Bibliography

Knime. (n.d.). *https://docs.knime.com/*. Retrieved from https://docs.knime.com/: https://docs.knime.com/latest/analytics_platform_best_practices_guide/index.html#what-is-knime

Glossário

ETL (Extract, Transform, Load) - Processo de extração, transformação e carregamento de dados. Envolve a extração de dados, que podem ter varias origens, a transformação para garantir que estão num formato adequado para análise ou integração, e o carregamento desses dados para um sistema, dashboard ou base de dados central.

Anonimização - Processo de remover ou modificar informações que possam identificar diretamente ou indiretamente qualquer tipo de dados, protegendo a privacidade dos dados. No KNIME, é utilizado o node de anonimização para garantir que os dados sensíveis, como nomes ou identificações, sejam protegidos.

CSV (Comma Separated Values) - Formato de ficheiro simples utilizado para armazenar dados tabulares, onde os valores são separados por vírgulas (ou outro delimitador). Amplamente utilizado para exportar e importar dados em diferentes sistemas.

XML(eXtensible Markup Language) - Linguagem de marcação extensível utilizada para descrever dados de forma estruturada, sendo amplamente utilizada para troca de dados entre sistemas. Os ficheiros XML permitem definir tags personalizadas, facilitando a leitura e interpretação por diferentes sistemas.

Metanode - No KNIME, um metanode é um agrupamento de vários nodes num único componente, facilitando a organização e visualização de fluxos de trabalho complexos.

Node - Elemento básico de um fluxo de trabalho no KNIME. Cada node representa uma operação ou uma transformação de dados, como leitura, filtragem, agregação ou visualização.

Salting - Método utilizado para proteger dados e outros dados sensíveis na anonimização ou encriptação. De um grosso modo, consiste em adicionar um valor aleatório (o “sal”) antes de aplicar uma função de hash, tornando mais difícil a obtenção do valor original a partir do hash.

API (Application Programming Interface) - Interface que permite a comunicação entre diferentes sistemas ou aplicações.

XPath (XML Path Language) - Linguagem utilizada para navegar e selecionar dados dentro de documentos XML. No KNIME, o node XPath permite extrair informações específicas de documentos XML, convertendo dados em colunas de uma tabela.

Dashboard - Ferramenta visual interativa utilizada para exibir dados e informações em tempo real, normalmente em gráficos e tabelas. No contexto do KNIME, dashboards podem ser criados para visualização de dados processados, facilitando a análise.

RegEx (Expressão Regular) - Sequência de caracteres que define um padrão de pesquisa. Em KNIME, o node “String Manipulation” permite utilizar expressões regulares para manipular strings, como substituir ou extrair partes de texto com base em padrões.