Instituto Politécnico do Cávado e do Ave
Escola Superior Tecnologia

António Carlos Fernandes Araújo 20746

Relatório de Integração de Sistemas de Informação

24 de outubro de 2024

Índice

Índice de Figuras

1. Enquadramento

Este projeto foi desenvolvido no âmbito da unidade curricular de Integração de Sistemas de Informação do curso de Licenciatura em Engenharia de Sistemas Informáticos.

O projeto centraliza-se na extração, transformação e carregamento (ETL) de dados provenientes de um ficheiro no formato ICS, amplamente utilizado para a partilha e sincronização de eventos de calendário.

O trabalho desenvolvido procurou demonstrar o domínio das operações de ETL aplicadas a dados de calendário, transformando dados originalmente complexos e dispersos numa estrutura uniforme e pronta para análise e visualização. Através de várias etapas do processo ETL, foram aplicadas técnicas de normalização, filtragem e manipulação de dados, com vista a proporcionar resultados adequados a múltiplos fins, como armazenamento em bases de dados, envio de notificações por e-mail e criação de visualizações dinâmicas que suportem a tomada de decisão.

2. Problema

O objetivo deste projeto é desenvolver um sistema eficaz para a gestão e análise de dados de eventos provenientes de ficheiros de calendário em formato ICS. Este tipo de ficheiro é amplamente utilizado para partilhar e sincronizar eventos entre diferentes plataformas e dispositivos. No entanto, a leitura e análise dos dados nele contidos podem ser complexas devido à estrutura não uniformizada e à variedade de informações que comporta.

Pretende-se, com este trabalho, demonstrar a aplicação prática de um processo ETL que permita normalizar e estruturar os dados dos eventos de calendário para facilitar a sua manipulação, visualização e integração em sistemas de informação distintos. A solução visa resolver a necessidade de transformar dados brutos e complexos em informação utilizável, automatizando a preparação dos dados para que possam ser facilmente consultados, analisados e integrados.

Com isto, o projeto pretende fornecer uma base para a criação de dashboards, o envio de notificações automatizadas por e-mail e o armazenamento eficiente dos eventos em bases de dados relacionais, tornando o processo mais eficiente e útil para a tomada de decisão e para a partilha de informações organizadas.

3. Estratégia Utilizada

Para alcançar os objetivos definidos, foram combinadas diversas tecnologias e operadores em etapas sequenciais, de modo a realizar a extração, transformação e carregamento dos dados (ETL) de forma eficiente e estruturada. Dado que o KNIME não possui um leitor direto para ficheiros ICS, foi necessário recorrer a um processo inicial em Python, antes de avançar com a manipulação dos dados no KNIME.

Extração do Calendário e Conversão de Ficheiro: Inicialmente, utilizou-se Python no ambiente Visual Studio Code para aceder à API do Google Calendar e extrair os eventos, gerando um ficheiro no formato ICS. De seguida, um segundo script em Python converteu o ficheiro ICS para CSV, facilitando a integração dos dados no KNIME.

Importação de Dados no KNIME: Com o ficheiro convertido para CSV, utilizaram-se operadores de leitura no KNIME para importar os dados e iniciar o processo de transformação.

Expressões Regulares: Durante a fase de transformação, aplicaram-se expressões regulares (Regex) para normalização e limpeza dos dados. Esta abordagem permitiu padronizar formatos, corrigir inconsistências e filtrar informações irrelevantes, preparando os dados para as etapas seguintes.

Operações de Dados: Foram aplicados diversos operadores para operações específicas sobre os dados, incluindo concatenações de campos, cálculos de tempos, substituições, filtros e validações. Estas operações organizaram e simplificaram a estrutura dos dados, facilitando o uso em relatórios e visualizações.

Armazenamento em Base de Dados: Para garantir a persistência dos dados, os eventos foram armazenados numa base de dados SQL. Os operadores de inserção e atualização criaram um repositório de dados acessível para consultas e análises futuras.

Visualização e Geração de Relatórios: Para uma análise visual dos dados, incorporaram-se operadores de visualização, permitindo a criação de dashboards e relatórios com métricas relevantes para a distribuição temporal dos eventos.

Envio de Notificações por E-mail: Como parte do processo de automação, foram configurados operadores para o envio de notificações por e-mail sobre eventos futuros. Esta etapa garante que os utilizadores recebem alertas com informações importantes de forma conveniente.

4. Transformações

Na primeira etapa, foram aplicadas as seguintes transformações para estruturar os dados:

Criação de Colunas: Adicionadas colunas como Evento_Title, Start_Date, End_Date, Location e Status para organizar as informações essenciais dos eventos, utilizando Regex.

Conversão de Datas: Ajustadas as colunas Start_Date e End_Date para formatos consistentes e compatíveis com a base de dados.

Filtragem, Tratamento e Limpeza: Aplicados filtros para remover colunas irrelevantes, eliminar eventos cancelados, tratar strings e limpar entradas numéricas no título dos eventos.

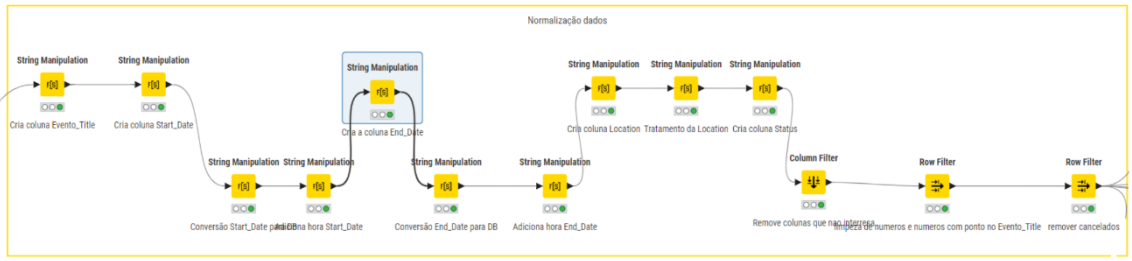

Figura 1 - Normalização de dados

Nesta outra etapa, já com a normalização concluída, começa com a conversão das datas, originalmente em string, para o tipo Date&Time, facilitando operações de dashboard do KNIME. De seguida, calcula-se a duração de cada evento ao medir a diferença entre Start_Date e End_Date, criando uma nova coluna que mostra essa diferença em dias. Com as datas já convertidas, extrai-se o ano de Start_Date, gerando uma coluna Event_Year que permite agrupar e analisar os eventos por ano. A seguir, os dados são agrupados por ano, com uma contagem do número de eventos para cada ano específico, resultando numa tabela que mostra a distribuição dos eventos ao longo do tempo. Para finalizar, é criado um gráfico de linha, com Event_Year no eixo x e o número de eventos no eixo y, oferecendo uma visualização clara da tendência de eventos ao longo dos anos.

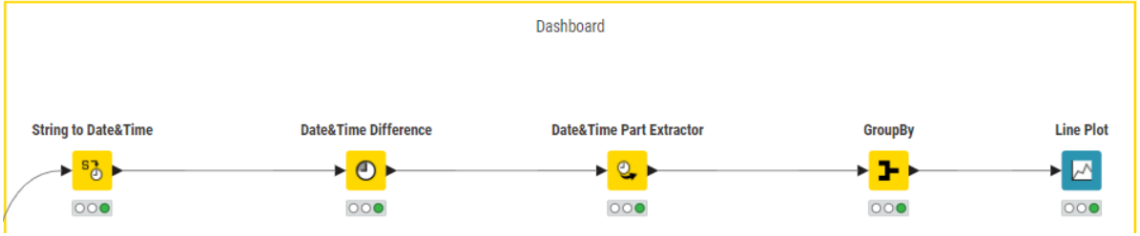

Figura 2 - Criação Dashboard

Nesta etapa, foi estabelecida uma conexão com a base de dados e inseridos os dados processados:

DB Connector: Configurada a conexão com a base de dados SQL, utilizando os parâmetros necessários.

DB Insert: Inseridos os dados normalizados na base de dados, preenchendo as tabelas com as informações de eventos, garantindo que estejam acessíveis para consultas futuras.

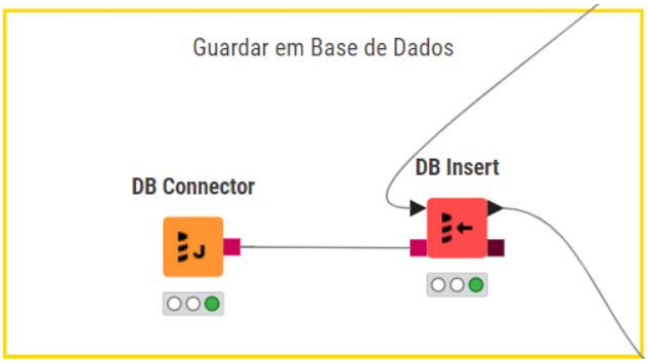

Figura 3 - Inserir em Base de Dados

Esta etapa identifica e guarda eventos futuros, além de enviar notificações por e-mail:

String to Date&Time: Os dados de data foram convertidos de texto para o formato Date&Time, essencial para possibilitar o filtro de eventos com base em intervalos temporais.

Date&Time-based Row Filter: Com os dados já em formato Date&Time, aplicou-se um filtro para selecionar apenas os eventos que ocorrerão no futuro. Este passo é fundamental para separar eventos relevantes e próximos dos que já ocorreram.

Excel Writer: Os eventos futuros filtrados foram exportados para um ficheiro Excel.

Send Email: Configurada uma notificação automática por e-mail, anexando o ficheiro Excel com os eventos futuros e anexando a imagem do dashboard gerado na etapa anteriormente.

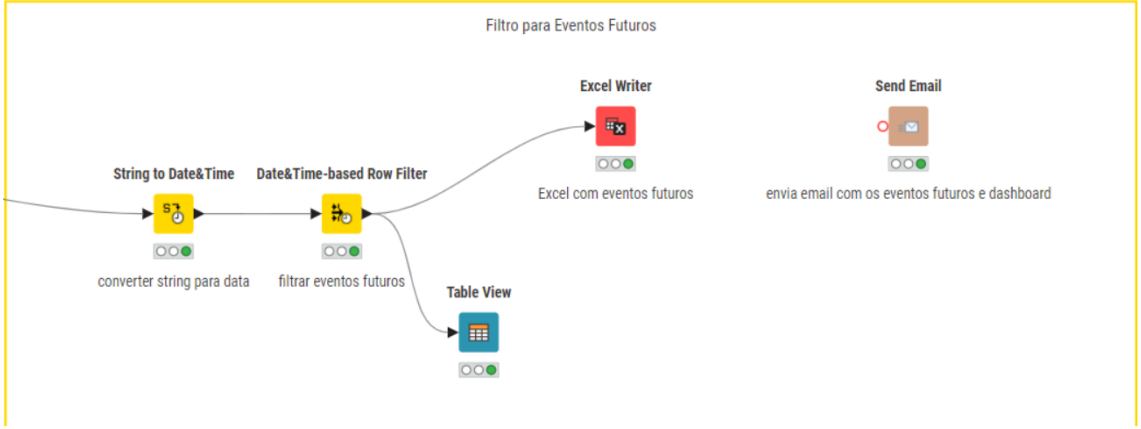

Figura 4 - Filtro de eventos futuros

Nesta etapa, aplicou-se um Row Filter com um filtro em Regex para visualizar exclusivamente os eventos que contêm a palavra "Teste" em alguma parte do seu nome.

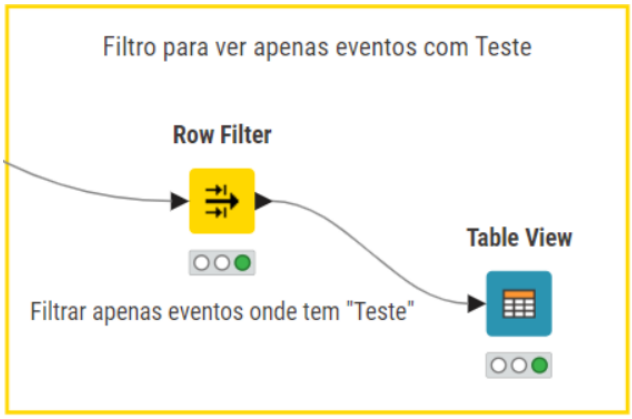

Figura 5 - Filtro para identificar eventos específicos

Nesta etapa, realizou-se o processo de normalização com um ficheiro XML como entrada, explorando a funcionalidade de XPath para extrair e manipular dados. O objetivo foi replicar as transformações feitas anteriormente, demonstrando que é possível alcançar o mesmo resultado com dados em formato XML.

XML Reader: Leitura do ficheiro XML contendo os dados de eventos.

XPath: Utilização de expressões XPath para extrair informações específicas do XML, facilitando a conversão para colunas utilizáveis no KNIME.

Column Renamer, Filter e Resorter: Renomeação, filtragem e reorganização das colunas extraídas, tornando os dados consistentes com o formato das etapas anteriores.

String Manipulation: Realizadas manipulações idênticas às etapas anteriores, como a conversão e ajuste das colunas Start_Date e End_Date, assegurando que os dados de tempo estejam no formato adequado.

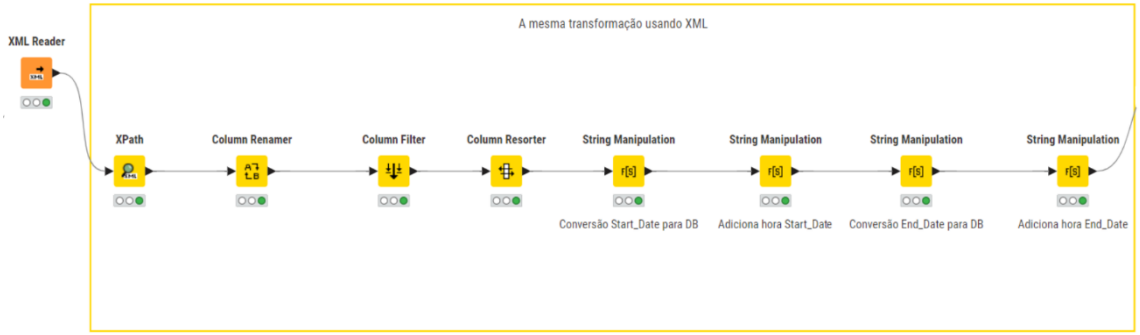

Figura 6 - Normalização com XML

5. Conclusão

Este projeto permitiu aplicar e consolidar os conhecimentos adquiridos na unidade curricular de Integração de Sistemas de Informação, nomeadamente na implementação de processos de ETL com recurso a diversas ferramentas e metodologias. A partir da extração de dados de um ficheiro ICS e utilizando Python para a conversão inicial dos dados para CSV, foi possível implementar um workflow no KNIME que realiza a normalização, transformação e armazenamento dos dados de eventos de calendário.

Ao longo do desenvolvimento, foi dada especial atenção à limpeza e estruturação dos dados, explorando expressões regulares e operações de manipulação para garantir uma representação uniforme e confiável dos eventos. Adicionalmente, foram implementados processos de automação, como o envio de notificações de eventos futuros por e-mail e a criação de dashboards para visualização dos resultados, facilitando a análise e a tomada de decisão.

Este trabalho demonstrou a importância da integração de diferentes tecnologias e abordagens para lidar com dados complexos e dispersos, reforçando a capacidade de adaptar soluções a cenários específicos. No futuro, seria interessante explorar novas funcionalidades, como a integração com outras APIs, a aplicação de métodos de análise preditiva e o desenvolvimento de relatórios mais dinâmicos e interativos. Assim, o projeto constitui uma base sólida para futuras explorações e inovações no âmbito da gestão e análise de dados em sistemas de informação.

Bibliografia

KNIME AG. (2024). KNIME. https://www.knime.com/getting-started-guide

Material de apoio fornecido pelo docente Luís Ferreira