Instituto Politécnico do Cávado e do Ave

Escola Superior de Tecnologia

Licenciatura em Engenharia de Sistemas Informáticos

Trabalho Prático (TP01)

Disciplina: Integração de Sistemas de Informação

Ano letivo 2024/2025

Aluno:

Vitor Sá Nº 20484

Barcelos

outubro de 2024

Revisão

Número da Revisão	Data da Revisão
1.1	24-10-2024
1.2	01-11-2024
1.3	05-11-2024

Índice

Conteúdo

Índice de figuras

Acrónimos e Abreviaturas

Acrónimo/Abreviatura	Significado	Descrição
ETL	Extract, Transform, Load	Processo de extração, transformação e carregamento de dados, comum em sistemas de integração para mover e modificar dados entre fontes e destinos distintos.
JSON	JavaScript Object Notation	Formato leve de intercâmbio de dados, amplamente utilizado para estruturar dados de forma legível para humanos e máquinas.
IP	Internet Protocol	Identificador numérico atribuído a cada dispositivo conectado a uma rede, que permite a comunicação na internet.
HTTP	HyperText Transfer Protocol	Protocolo usado para a transferência de dados na internet, fundamental para o funcionamento de páginas web.
XML	eXtensible Markup Language	Linguagem de marcação usada para definir documentos com uma estrutura que pode ser facilmente interpretada por máquinas e humanos, frequentemente usada em relatórios.
PDI	Pentaho Data Integration	Ferramenta de ETL da Pentaho que permite a extração, transformação e carregamento de dados de diversas fontes para vários formatos.
URL	Uniform Resource Locator	Endereço que especifica a localização de um recurso na internet, como uma página web.
TP	Trabalho Prático	Designação comum para trabalhos de natureza prática realizados por alunos no contexto de disciplinas académicas.
DB	Database (Base de Dados)	Coleção estruturada de dados que pode ser acedida e gerida por sistemas específicos.
KPI	Key Performance Indicator	Indicador chave de desempenho, métrica utilizada para avaliar o sucesso de uma organização ou de um sistema em atingir objetivos específicos.

Introdução

No contexto da era digital, o volume de dados gerados pelas interações online cresce de forma exponencial, criando a necessidade de ferramentas e processos robustos para lidar com essa informação de maneira eficiente.

A integração e o tratamento desses dados são fundamentais para que as organizações possam extrair insights valiosos, monitorizando atividades e tomar decisões baseadas em dados concretos. No âmbito da disciplina de Integração de Sistemas de Informação, o presente trabalho prático (TP01) visa aplicar conhecimentos sobre processos de ETL (Extract, Transform, Load), essenciais para a manipulação e transformação de grandes volumes de dados.

Neste projeto, o foco recai sobre a análise de logs de acesso a um servidor web, utilizando a ferramenta Pentaho Data Integration (PDI). Os logs são fornecidos em formato JSON, e o objetivo é transformar esses dados brutos em relatórios estruturados, disponibilizando as informações de forma mais acessível e facilitar a monotorização de atividades no servidor.

1. Problema

Este relatório explora o desafio de analisar grandes volumes de dados provenientes de logs de acesso a servidores WEB. Esses logs contêm informações valiosas, como endereços IP, endereços acedidos, códigos de status HTTP e tempos de resposta, que são essenciais para validar a atividade dos utilizadores, avaliar o desempenho do sistema e identificar possíveis ameaças à segurança.

No entanto, o volume massivo de dados gerado torna o processo manual de análise impraticável, pois demoraria muito tempo e estaria sujeito a erros.

Para resolver esse problema, é necessário um processo automatizado de ETL (Extract, Transform, Load) que permita extrair os dados relevantes desses logs em formato JSON, transformá-los de maneira a padronizar e calcular métricas importantes, e finalmente gerar relatórios acessíveis em formatos como Excel e XML. Esses relatórios facilitam a análise do tráfego, a identificação de padrões de uso e a deteção de problemas de desempenho, permitindo que as decisões sobre o sistema sejam mais informadas e eficazes.

Por fim a informação é também carregada num dashboard online para que seja possível aceder a todos esses dados remotamente.

A automação desse processo não só aumenta a eficiência e precisão, mas também garante que a supervisão do servidor web ocorra de forma contínua e consistente, contribuindo para a melhoria e segurança do sistema.

2. Estratégia utilizada

Para o desenvolvimento deste projeto ETL, foi elaborada uma estratégia que assegura uma manipulação eficaz dos dados de logs e uma geração de relatórios informativos e consistentes.

O processo foi dividido em etapas sequenciais, onde cada fase contribui para transformar os dados brutos de logs em informações úteis para análise. Abaixo está a descrição detalhada de cada etapa da estratégia utilizada:

1. Identificação dos Dados: O primeiro passo consistiu em compreender o conteúdo dos logs e identificar os campos de dados essenciais para a análise. Nesta fase, foram definidos campos como ip_address, timestamp, request_method, status_code e response_time, que representam informações críticas para o monitoramento de acessos ao servidor. Esta identificação permitiu delimitar quais dados seriam extraídos e transformados, evitando o processamento desnecessário de informações irrelevantes.
2. Definição de Objetivos: Após a identificação dos dados, foram estabelecidos os principais objetivos analíticos do projeto. Estes objetivos incluíam métricas como URLs mais visitados, códigos de estado HTTP mais frequentes (para identificação de possíveis erros de servidor ou cliente) e o tempo médio de resposta do servidor. Esses objetivos foram fundamentais para orientar as transformações e a geração de estatísticas que trariam valor para o monitoramento e melhoria do sistema web.
3. Transformação de Dados: Com os dados padronizados, foram aplicadas operações de transformação para gerar as métricas e estatísticas desejadas. Isso incluiu o agrupamento de acessos por URL para identificar os mais visitados, cálculo da média e mediana do tempo de resposta para análise de desempenho e contagem de ocorrências de cada código de status HTTP para monitoramento de erros. A utilização do MongoDB como base de dados foi essencial para realizar agregações rápidas e consultas eficientes sobre os dados de log.
4. Geração de Relatórios: Uma vez calculadas as métricas, os dados processados foram exportados em formatos de fácil visualização e análise, como Excel e XML. Esses relatórios foram estruturados para facilitar a compreensão e auxiliar na análise de tráfego e desempenho do servidor, tornando as informações acessíveis para os responsáveis pela administração do sistema.
5. Documentação e Validação: Durante todo o processo, cada etapa foi documentada e validada para garantir que os dados manipulados estivessem alinhados aos objetivos do projeto e que os resultados fossem consistentes com as expectativas. A validação incluiu

testes para verificar a integridade dos dados transformados e a precisão das métricas geradas, assegurando a qualidade e confiabilidade do processo ETL.

A adoção dessa estratégia garantiu um fluxo de trabalho eficiente e a obtenção de relatórios que são cruciais para o monitoramento e otimização do sistema web.

O uso do MongoDB como base de dados proporcionou flexibilidade e escalabilidade, permitindo uma gestão eficiente dos dados em formato JSON e facilitando a integração com as ferramentas de transformação e geração de relatórios.

3. Transformações

Transformações, no contexto de um projeto de ETL (Extract, Transform, Load), são operações aplicadas aos dados para convertê-los de seu formato original em um formato mais útil ou estruturado. Essas transformações são essenciais para padronizar, enriquecer e organizar dados brutos, facilitando sua análise e utilização em relatórios ou sistemas de decisão.

No Pentaho Data Integration (PDI), cada transformação é configurada como um passo específico que define como os dados devem ser manipulados antes de serem carregados para o destino final.

No projeto de análise de logs, algumas transformações principais incluíram:

- Filtro de Linhas: Esta transformação permite selecionar apenas os registros que atendem a critérios específicos, como acessos com códigos de erro (por exemplo, 404) ou a URLs específicas. Esse filtro ajuda a focar a análise em aspetos relevantes do tráfego do servidor.
- Agrupamento de Dados (Group By): Agrupa os dados com base em um ou mais campos, como referrer ou URL, permitindo calcular estatísticas, como o número de acessos a uma URL específica ou a média do tempo de resposta. Esse tipo de transformação é útil para gerar métricas agregadas sobre o comportamento dos utilizadores.
- Ordenação de Dados: Organiza os registros com base em critérios de relevância, como as URLs mais acedidas ou o tempo de resposta. Ordenar dados ajuda a identificar tendências e padrões, facilitando a análise e visualização.
- Conversão de Formatos: Transforma os dados para relatórios em diferentes formatos (como XML e Excel) de acordo com as necessidades de visualização. Esse passo assegura que os relatórios sejam acessíveis para diversos fins, sejam eles técnicos ou executivos.

Essas transformações compõem a lógica do fluxo ETL no Pentaho, permitindo que os dados sejam preparados de forma automatizada para análise, relatórios e monitoramento, resultando em uma visão estruturada e prática das informações de logs.

Transformações Específicas do Projeto

Transformação Principal: (Figura 1)

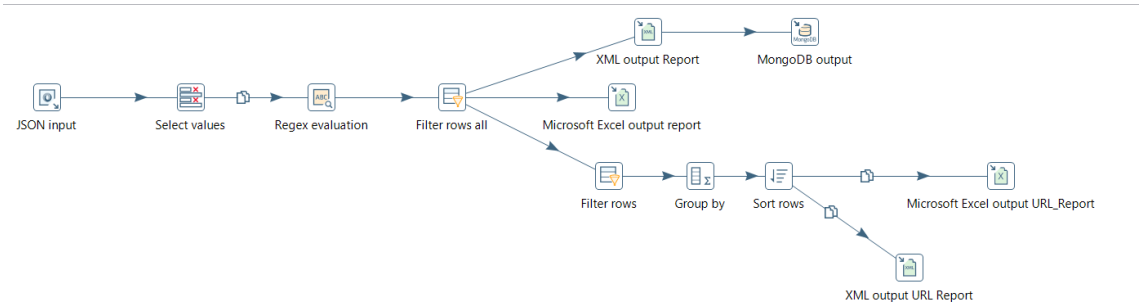

Figura 1 - Transformação Principal

- **Descrição**: A transformação principal lê o ficheiro JSON com logs de servidor, filtra os dados e executa as agregações necessárias, como contagem de acessos por URL (referrer) e cálculo da média do tempo de resposta (response_time).

Configuração: **JSON Input**: Carrega os dados de log no formato JSON.

Regex evaluation: verifica a formatação dos endereços IP,

Filter Rows: Remove valores nulos ou vazios do campo referrer.

Group By: Realiza as agregações para contar acessos e calcular a média do tempo de resposta.

Sort Rows: Ordena os URLs mais acedidos em ordem decrescente.

Excel Output: Salva os resultados em ficheiro Excel para análise.

4. Jobs

O projeto conta com dois Jobs principais: **Job_CreateFolders** e **Job_Main**.

Job_CreateFolders: (Figura 2)

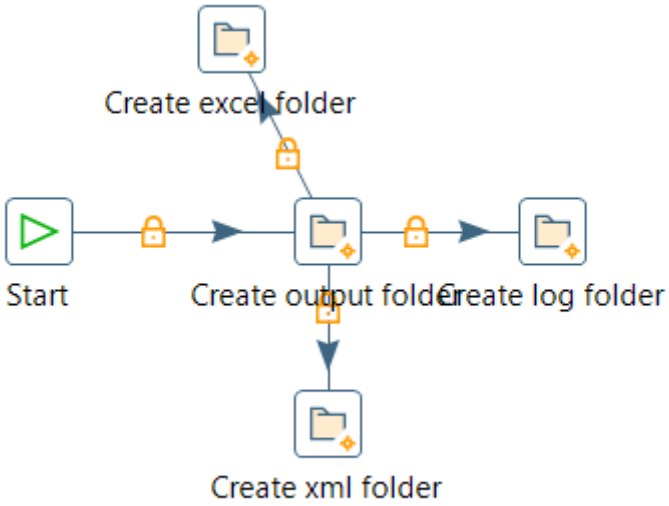

Figura 2 - Job Create Folders

- **Descrição**: Este job cria a estrutura de pastas necessária para organizar as entradas, saídas e logs do projeto. Isso inclui pastas para armazenar os relatórios em XML, Excel e logs.

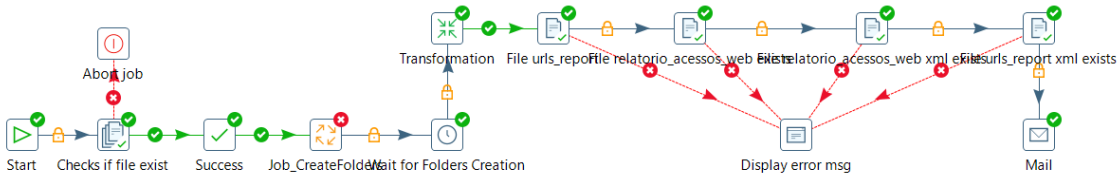

Figura 3 - Job Main

- **Descrição**: O Job principal executa a sequência completa de transformações, verificando se o ficheiro de entrada existe, iniciando as transformações de dados e, ao final, enviando o relatório por e-mail.

Integração de Sistemas de Informação 10

Configuração: **File Exists**: Verifica a presença do ficheiro JSON antes de iniciar as transformações.

Transformations: Executa a sequência de transformações configuradas.

Send Mail: Envia o relatório gerado por e-mail ao destinatário configurado.

- Exportação de Logs: Adiciona informações sobre a execução e possíveis erros do processo de ETL, permitindo monitorizar o funcionamento e a eficácia das transformações e está implementada apenas no corpo email.

5. Dashboard

No projeto, o uso do Metabase para configurar o dashboard permitiu uma análise abrangente e acessível dos dados de logs de servidor, sem a necessidade de configuração manual. Com visualizações automáticas baseadas nas tabelas de dados disponíveis, o dashboard oferece insights fundamentais para o monitoramento contínuo do sistema. Entre as principais visualizações estão o "Timestamp by Day of the Week", que revela padrões de tráfego ao longo da semana, e o "Status Code Over Time", que monitora os códigos de status HTTP, permitindo identificar problemas de desempenho. A segmentação dos "Dados JSON per Server Name" facilita a análise em ambientes com múltiplas instâncias, destacando anomalias em servidores específicos. Com outras visualizações adicionais, o dashboard proporciona uma visão completa e intuitiva do sistema, garantindo uma gestão mais eficiente e proativa.

Figura 4 - Dashboard

6. Demonstração

No vídeo, apresentei uma demonstração prática do projeto ETL em funcionamento, ilustrando cada etapa do processo de integração e transformação dos dados de logs do servidor. Durante a demonstração, é possível observar como o Pentaho Data Integration executa as tarefas de extração, transformação e carregamento de dados, transformando registros JSON em relatórios prontos para análise.

A interface do Pentaho permite acompanhar o fluxo das operações, destacando a eficiência e a automatização das tarefas de processamento de dados.

Além disso, incluí um QR code no vídeo para que os interessados possam acessar o projeto diretamente.

Este QR code facilita o acesso a mais informações sobre o projeto, permitindo que os espectadores explorem os detalhes e resultados de forma prática.

Figura 5 - QR Code para video de demonstração

7. Conclusão

O desenvolvimento deste projeto demonstrou a eficácia do uso do Pentaho Data Integration para automatizar e otimizar o processo de análise de logs de servidor web. Através de um fluxo ETL bem estruturado, foi possível extrair e transformar dados relevantes, convertendo logs em informações detalhadas sobre o tráfego, desempenho e padrões de uso dos URLs acedidos. Esses resultados contribuem de forma significativa para análise do sistema, permitindo uma gestão mais informada e proativa dos serviços de servidor.

O projeto não só alcançou os objetivos estabelecidos, mas também destacou áreas com potencial de aprimoramento para futuras iterações. Para melhorar ainda mais o processo e os resultados, são recomendadas as seguintes iniciativas:

- Aprimoramento das Transformações: Otimizar as operações de ETL para lidar de forma ainda mais eficiente com grandes volumes de dados, garantindo um desempenho robusto e escalável. Procurar transformar os dados em novas métricas estatísticas.
- Integração de Fontes Externas: Ampliar o contexto da análise ao incorporar dados externos, como informações de localização baseadas no IP, enriquecendo os insights sobre o comportamento dos utilizadores. Receber também dados através de uma API, assim procurando automatizar o processo.
- Visualização de Dados: Explorar de forma mais profunda a ferramenta Metabase para criar mais formas de demonstração de dados. Procurar outras alternativas para a demonstração de dados, possivelmente uma aplicação local.
- Automatização Completa: Expandir a automação, incluindo o agendamento e envio automático de relatórios, além do monitoramento constante do processo ETL, reduzindo a necessidade de intervenção manual e assegurando a continuidade das operações.

Essas recomendações visam não apenas aperfeiçoar o desempenho do sistema atual, mas também proporcionar uma análise de dados mais abrangente, segura e acessível, alinhada com as melhores práticas de gestão de dados. Em conclusão, este projeto demonstra como a aplicação de processos ETL, juntamente com uma ferramenta poderosa como o Pentaho, pode transformar dados complexos em informações estratégicas para a melhoria contínua dos sistemas web.

8. Bibliografia

- https://pentaho-public.atlassian.net/wiki/spaces/EAI/pages/371558145/Filter+Rows consultado em 27/10/2023
- https://www.google.com/search?q=write+to+log+pentaho+example&rlz=1C1GCEA_enPT1044PT1044&oq=write+to+log+pentaho&gs_lcrp=EgZjaHJvbWUqCggBEAAYExgWGB4yBggAEEUYOTIKCAEQABgTGBYYHjIKCAIQABgTGBYYHjIKCAMQABgTGBYYHjIKCAQQABgTGBYYHjIKCAUQABgTGBYYHjIKCAYQABgTGBYYHjIKCAcQABgTGBYYHjIKCAgQABgTGBYYHtIBCDY5MDVqMGo3qAIAsAIA&sourceid=chrome&ie=UTF-8#fpstate=ive&vld=cid:00758d93,vid:gwI8WPbK3mU,st:0 consultado em 28/10/2023
- https://stackoverflow.com/questions/66694905/how-show-transformation-result-in-write-to-log-step-in-pentaho consultado em 28/10/2023
- https://www.mockaroo.com/ consultado em 27/10/2023
- https://help.hitachivantara.com/Documentation/Pentaho/Data_Integration_and_Analytics/9.4/Products/Pentaho_Data_Integration consultada documentação da Hitachi online no período de 27/10/2023 a 29/10/2023
- https://www.youtube.com/watch?v=1XB6mpXCgi4 consultado em 27/10/2023

- **Stack Overflow - Pentaho Logging Examples**: https://stackoverflow.com
- **Mockaroo (Data Generation)**: https://www.mockaroo.com/
- **Metabase**: https://www.metabase.com/