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

Índice

Conteúdo

Índice de figuras

Introdução

No contexto atual, a integração de dados desempenha um papel crucial na otimização de processos empresariais e na tomada de decisões informadas. A disciplina de Integração de Sistemas de Informação proporciona conhecimentos valiosos para enfrentar esses desafios. Este relatório apresenta o primeiro trabalho prático da disciplina, com foco no desenvolvimento de um processo ETL (Extract, Transform, Load) para a manipulação de dados de logs de acessos a um servidor web.

Para este trabalho, foi utilizada a ferramenta **Pentaho Data Integration (PDI)**, que permitiu a manipulação dos dados de entrada, representados por ficheiros JSON, e a geração de múltiplos formatos de ficheiros de saída, como Excel e XML. O objetivo central é demonstrar como processos de ETL podem ser aplicados para transformar e analisar dados de logs, gerando relatórios que contribuem para o monitoramento e a melhoria de desempenho de sistemas web.

1. Problema

A análise de logs de acesso a servidores web é essencial para monitorizar a atividade dos utilizadores, identificar possíveis problemas de desempenho e garantir a segurança. No entanto, a elevada quantidade de dados torna o processo manual inviável e propenso a erros. Com isso, torna-se necessário implementar um processo automatizado de ETL que capture e transforme as informações mais relevantes para a análise.

O trabalho consiste em extrair dados relevantes de logs JSON, transformar e processar esses dados com operações de filtragem, agregação e normalização, e gerar relatórios nos formatos desejados. Em resumo, o projeto visa:

- Extrair informações de logs JSON, identificando campos essenciais para a análise, como IP, URL, código de estado HTTP e tempo de resposta.
- Transformar esses dados para padronizar formatos e calcular métricas de interesse.
- Gerar relatórios que auxiliam na análise do tráfego e na identificação de padrões de uso e possíveis gargalos de desempenho.

2. Estratégia utilizada

O desenvolvimento do projeto ETL foi orientado pelos seguintes passos:

- **Identificação dos dados**: Compreender os dados de log disponíveis e definir quais campos são essenciais para o objetivo do projeto (por exemplo, ip_address, timestamp, request_method, status_code, response_time, etc.).
- **Definição de objetivos**: Estabelecer as estatísticas desejadas, como URLs mais visitados, códigos de estado mais frequentes, e tempo médio de resposta.
- **Limpeza e normalização**: Garantir que os dados de log estão em um formato padronizado, realizando tratamentos de campos nulos ou inválidos.
- **Transformação de dados**: Aplicar operações para agrupar e calcular as métricas de interesse, como contagem de acessos por URL, cálculo de média e mediana do tempo de resposta.
- **Geração de relatórios**: Configurar a saída dos dados processados em ficheiros Excel e XML, permitindo uma análise mais detalhada e organizada.
- **Documentação e validação**: Registar e testar cada etapa do processo para garantir que os dados manipulados atendam aos objetivos definidos e estão em conformidade com o esperado.

3. Transformações

No projeto, o **Pentaho Data Integration (PDI)** foi utilizado para implementar diversas transformações essenciais para o processamento e análise dos logs de servidor. As transformações principais incluem:

- **Filtro de Linhas**: Seleciona registos que atendem a critérios específicos (como códigos de estado 404 ou acessos a URLs específicas).
- **Agrupamento de Dados (Group By)**: Agrupa os dados por campos específicos, como referrer, para calcular o número de acessos por URL e a média do tempo de resposta.
- **Ordenação de Dados**: Organiza os dados por relevância (ex.: URLs mais acedidos) para facilitar a análise.
- **Conversão de Formatos**: Gera relatórios em diferentes formatos, como XML e Excel, para atender a necessidades específicas de visualização.
- **Exportação de Logs**: Adiciona logs de execução e erro para monitoramento do processo.

Transformações Específicas do Projeto

Transformação Principal: (Figura 1)

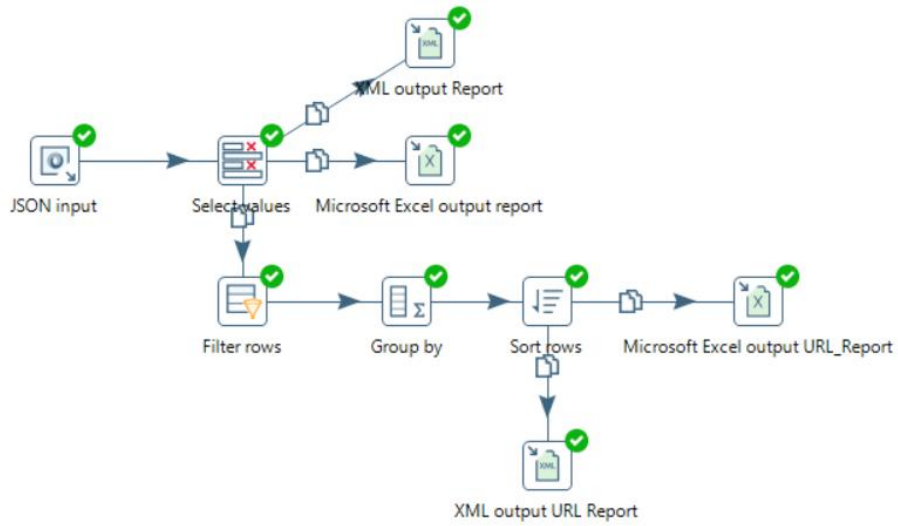

Figura 1 – Transformation_Reports

- **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.
    - **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.

Transformação de Logs:

- **Descrição**: Uma transformação adicional captura informações sobre erros e registos de execução. Isso ajuda a monitorizar o andamento do processo ETL e identificar possíveis problemas.
- **Configuração**:
    - **Write to Log**: Regista o status de cada passo e eventuais erros.
    - **Text File Output**: Armazena logs detalhados em ficheiros de texto.

4. Jobs

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

Job_CreateFolders: (Figura 2)

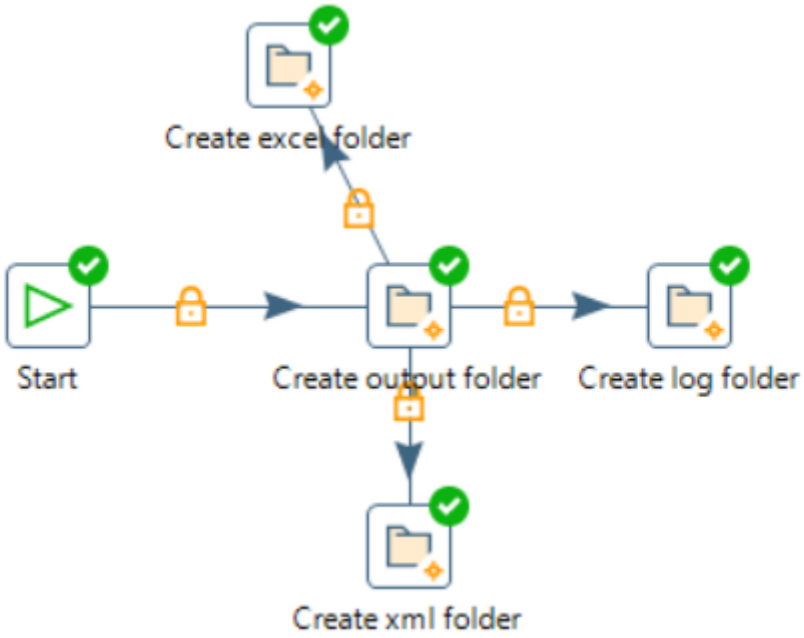

Figura 2 - Job_CreateFolders

- **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.

Job_Main: (Figura 3)

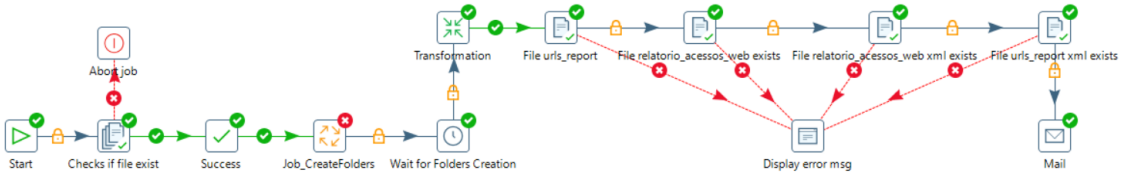

Figura 3 - Job_Main

- **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.

- **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.

5. Vídeo de Demonstração

6. Conclusão e trabalhos futuros

O projeto demonstrou com sucesso como é possível utilizar o Pentaho Data Integration para automatizar a análise de logs de servidor web. Foram extraídas informações valiosas sobre o tráfego, desempenho e padrões de uso dos URLs acedidos, o que auxilia na gestão e optimização dos serviços de servidor.

Para futuros trabalhos, as seguintes melhorias são recomendadas:

- **Aprimoramento das transformações**: Refinar as operações de ETL para optimizar o desempenho, especialmente em grandes volumes de dados.
- **Anonimização de dados**: Implementar métodos avançados de anonimização para garantir conformidade com regulamentações de privacidade, como o RGPD.
- **Integração de fontes externas**: Enriquecer a análise de logs integrando dados de outras fontes, como informações de localização baseadas no IP.
- **Visualização de dados**: Explorar ferramentas de visualização para apresentar os resultados de maneira mais acessível.
- **Automatização completa**: Implementar uma maior automação no envio de relatórios e monitoramento do processo ETL, minimizando a necessidade de intervenção manual.

7. Referências

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