Instituto Superior de Engenharia

Politécnico de Coimbra

Integração de Dados

CTeSP Tecnologias e Programação de Sistemas de Informação (Cantanhede)

Professor: João Leal

joao.leal@isec.pt

UNIgreen The green European University

PRR Plano de Recuperação e Resiliência

REPÚBLICA PORTUGUESA

Financiado pela União Europeia NextGenerationEU

Integração de dados em sistemas distribuídos

Instituto Superior de Engenharia
Politécnico de Coimbra

- Existência de múltiplas fontes e formatos (CSV, Excel, NoSQL, SQL, APIs)

- Problemas comuns:
  - Dados redundantes
  - Diferentes padrões de codificação (UTF-8, ISO-8859-1)
  - Diferentes unidades de medida (kg vs g, € vs $)
  - Chaves primárias diferentes

UNIgreen The green European University · PRR Plano de Recuperação e Resiliência · República Portuguesa · Financiado pela União Europeia NextGenerationEU

Integração de dados em sistemas distribuídos

Técnicas de resolução:

- Mapeamento de esquemas
- Unificação de modelos de dados
- Normalização e padronização

Integração de dados em tempo real

- A integração de dados em tempo real representa um paradigma fundamentalmente diferente da integração por lotes (*batch processing*).

- Enquanto o processamento por lotes lida com grandes volumes de dados em intervalos de tempo definidos (ex: diário ou noturno), a Integração de Dados em Tempo Real visa capturar, processar e entregar dados no momento em que são gerados, ou com uma latência mínima, geralmente medida em milissegundos ou segundos.

Integração de dados em tempo real

- A necessidade de integração em tempo real surge em cenários onde a informação tem um valor temporal crítico.

- Exemplos comuns incluem a deteção de fraude em transações financeiras, a monitorização de sistemas de IoT (*Internet of Things*) para alertas imediatos, e a personalização de experiências web baseada no comportamento atual do utilizador.

Integração de dados em tempo real

Tipo de Processamento	Latência Típica	Descrição	Casos de Uso
Batch (Lotes)	Horas a Dias	Processamento de grandes volumes de dados em períodos programados.	Relatórios mensais, migrações de dados.
Near Real-Time	Segundos a Minutos	Os dados são processados rapidamente, mas com um pequeno atraso tolerável.	Atualização de dashboards operacionais.
Real-Time	Milissegundos	Processamento imediato dos dados à medida que são gerados.	Deteção de fraude, controlo de processos industriais.

Tecnologias e Arquiteturas

- Para alcançar a baixa latência exigida pela integração em tempo real, são necessárias tecnologias e padrões de arquitetura específicos.

Change Data Capture (CDC)

- O **Change Data Capture (CDC)** é uma técnica essencial que identifica e rastreia as alterações (inserções, atualizações e eliminações) nos dados de uma base de dados de origem, e as entrega a um sistema de destino em tempo real.

- O CDC permite que os sistemas de destino se mantenham sincronizados com a origem sem a necessidade de realizar leituras completas e dispendiosas da base de dados.

Change Data Capture (CDC)

Os métodos mais comuns de implementação de CDC incluem:

Baseado em Logs (Log-Based CDC):

- Considerado o método mais eficiente e menos intrusivo.

- O sistema lê diretamente os logs de transações da base de dados (ex: Binlog do MySQL, Redo Logs do Oracle) para capturar as alterações antes que estas sejam aplicadas.

Change Data Capture (CDC)

Baseado em Triggers:

- Utiliza *triggers* (gatilhos) na base de dados para registar as alterações numa tabela de auditoria separada.

- Embora simples, pode introduzir sobrecarga (*overhead*) no desempenho da base de dados de origem.

Change Data Capture (CDC)

Baseado em Colunas de Timestamp:

- Envolve a consulta periódica de tabelas que possuem colunas de data/hora de última modificação.

- É o método menos "real-time" e mais propenso a falhas de captura de alterações.

Message Queues e Stream Processing

- A integração em tempo real é frequentemente implementada através de plataformas de Message Queues ou Event Streaming, que atuam como intermediários de alta velocidade para o fluxo de dados.

Message Queues e Stream Processing

RabbitMQ:

- É um *message broker* tradicional, otimizado para o encaminhamento de mensagens individuais e para a gestão de filas de tarefas (*task queues*).

- É ideal para cenários de comunicação ponto-a-ponto ou para garantir a entrega de tarefas.

Message Queues e Stream Processing

Apache Kafka

- É uma plataforma de *event streaming* distribuída, concebida para lidar com um elevado volume de dados contínuos e para o processamento em tempo real em grande escala.

- Kafka é otimizado para a persistência de eventos e para a leitura por múltiplos consumidores, sendo a escolha preferencial para a maioria dos pipelines de Big Data em tempo real.

Arquiteturas Lambda e Kappa

- Estas arquiteturas definem a forma como os dados são processados para garantir tanto a precisão histórica quanto a velocidade em tempo real.

Arquiteturas Lambda e Kappa

Arquitetura Lambda:

- Combina duas camadas de processamento: uma camada Batch (para precisão e dados históricos) e uma camada Speed (para baixa latência e dados em tempo real).
- Os resultados de ambas as camadas são combinados numa camada Serving para apresentar uma visão completa.
- A sua principal desvantagem é a complexidade de manter e desenvolver código em duas camadas separadas.

Arquiteturas Lambda e Kappa

Arquitetura Kappa:

- Simplifica a Lambda ao eliminar a camada Batch.
- Todo o processamento é feito através de uma única camada Stream (processamento de fluxo).
- Os dados históricos são tratados simplesmente reprocessando o log de eventos desde o início.
- Esta arquitetura é mais simples de manter e mais económica, sendo a tendência atual para muitos sistemas de Big Data.

O Desafio do Big Data na Integração

O conceito de Big Data é frequentemente definido pelos seus 3 Vs originais, que descrevem as características que tornam os dados difíceis de processar com métodos tradicionais :

- ***Volume***
- ***Velocidade (Velocity)***
- ***Variedade (Variety)***

O Desafio do Big Data na Integração

- ***Volume:*** A quantidade massiva de dados gerados (terabytes, petabytes, exabytes). A integração deve ser capaz de escalar horizontalmente para lidar com este volume.

- ***Velocidade (Velocity):*** A rapidez com que os dados são gerados, recolhidos e processados (*abordado anteriormente - Real-Time*).

O Desafio do Big Data na Integração

- ***Variedade (Variety):*** A diversidade de formatos e tipos de dados, que vão desde o estruturado (tabelas relacionais) ao semi-estruturado (JSON, XML) e ao não-estruturado (texto livre, imagens, logs).

A Variedade é o principal desafio, pois a integração tradicional baseada em esquemas fixos (Schema-on-Write) não é adequada para dados que não se encaixam facilmente em linhas e colunas.

Tecnologias e Plataformas

- Para lidar com a variedade e o volume do Big Data, surgiram novas tecnologias e arquiteturas de armazenamento.
  - *Bases de Dados NoSQL*
  - *Data Lakes e Data Warehouses Modernos*

Bases de Dados NoSQL

- As bases de dados NoSQL (Not Only SQL) são concebidas para lidar com grandes volumes de dados e esquemas flexíveis, sendo essenciais na integração de dados variados.

- Data Lakes e Data Warehouses Modernos
- A integração de Big Data requer uma infraestrutura de armazenamento que possa receber dados em qualquer formato.

Data Lakes e Data Warehouses Modernos

- A integração de Big Data requer uma infraestrutura de armazenamento que possa receber dados em qualquer formato.

- **Data Warehouse (DW):** Armazena dados estruturados que foram limpos, transformados e modelados (Schema-on-Write) para fins de Business Intelligence (BI) e relatórios predefinidos. É otimizado para consultas analíticas rápidas.

Tecnologias e Plataformas

- **Data Lake (DL):** Armazena dados brutos em qualquer formato (estruturado, semi-estruturado, não-estruturado) e permite que o esquema seja aplicado no momento da leitura (Schema-on-Read). É ideal para ciência de dados, machine learning e análises exploratórias.

- **Data Lakehouse:** Uma arquitetura híbrida que combina a flexibilidade e o baixo custo do Data Lake com as estruturas de gestão de dados e desempenho do Data Warehouse, utilizando formatos abertos como Parquet ou Delta Lake .

Tecnologias e Plataformas

- O Data Lake atua frequentemente como o ponto de entrada (*staging area*) para todos os dados, incluindo os não-estruturados, antes de serem integrados e transformados para o Data Warehouse ou para outras aplicações.

Processamento de Dados Não-Estruturados

- A integração de dados semi-estruturados (ex: JSON, XML) e não-estruturados (ex: logs, texto livre) exige etapas adicionais de **Parsing** e **Normalização** para que possam ser utilizados em análises ou carregados em bases de dados relacionais.

Processamento de Dados Não-Estruturados

Parsing (Análise Sintática):

- É o processo de analisar o formato dos dados para extrair informações significativas.
- Por exemplo, num ficheiro JSON, o ***parsing*** identifica as chaves e os valores correspondentes.
- Em logs de servidor, o ***parsing*** separa os campos (*timestamp*, IP, método HTTP, URL) com base em delimitadores ou padrões (*regex*).

Processamento de Dados Não-Estruturados

Normalização:

- Após a extração, os dados precisam de ser normalizados, ou seja, **transformados num formato consistente e padronizado**. Isto pode incluir:
  - *Aplanar (Flattening)*
  - *Tipagem*
  - *Padronização*

Data Governance e Qualidade de Dados na Integração Avançada

- À medida que a integração de dados se torna mais complexa (tempo real, Big Data, variedade de fontes), a necessidade de gerir e garantir a fiabilidade dos dados integrados torna-se crítica.

- A *Data Governance* (Governança de Dados) e a Qualidade de Dados são os pilares que sustentam a confiança e o valor dos sistemas de integração.

Data Governance e Qualidade de Dados na Integração Avançada

- A *Data Governance* (Governança de Dados) é o conjunto de políticas, normas, padrões e práticas que orientam, monitoram e avaliam a gestão e o uso dos dados, para assegurar que sejam utilizados de forma consistente, segura e em conformidade com as regulamentações (ex: RGPD).

O Papel da *Data Governance* na Integração de Dados

A *Data Governance* é vital para a integração, pois define:

- **Propriedade dos Dados:** Quem é responsável pela qualidade e integridade dos dados em cada sistema de origem (*Data Owners*).

- **Padrões de Integração:** Normas para nomenclatura, formatos de dados e modelos de dados canónicos a serem usados nos pipelines de integração.

O Papel da *Data Governance* na Integração de Dados

- **Segurança e Acesso:** Políticas que garantem que apenas utilizadores e sistemas autorizados possam aceder e modificar os dados integrados, especialmente em ambientes de *Big Data* e *cloud*.

- **Conformidade:** Assegurar que os dados integrados cumprem os requisitos legais e regulamentares.

Qualidade de Dados (Data Quality)

- A Qualidade de Dados refere-se à adequação dos dados para o uso pretendido.

- Dados de baixa qualidade podem levar a decisões de negócio erradas, ineficiências operacionais e perda de confiança.

Qualidade de Dados (Data Quality)

Instituto Superior de Engenharia
Politécnico de Coimbra

Dimensão	Descrição	Exemplo de Falha
Precisão (Accuracy)	O grau em que os dados refletem a realidade.	Um endereço de cliente está incorreto.
Completude (Completeness)	A percentagem de valores não nulos em campos críticos.	O campo "Email" está vazio para 30% dos clientes.
Consistência (Consistency)	Os dados são coerentes entre diferentes sistemas.	O nome do cliente é "João Silva" num sistema e "J. Silva" noutro.
Atualidade (Timeliness)	Os dados estão disponíveis e atualizados no momento necessário.	O inventário de produtos está desatualizado em 24 horas.
Validade (Validity)	Os dados estão em conformidade com o formato, tipo e intervalo de valores definidos.	Um campo de idade contém o valor "abc" ou um número negativo.

Processos de *Data Profiling* e *Data Cleansing*

Data Profiling (*Criação de Perfil de Dados*):

- É a análise sistemática dos dados de origem para descobrir a sua estrutura, qualidade e conteúdo.

- O ***profiling*** identifica padrões, anomalias, valores únicos, e a distribuição de valores, fornecendo uma base para definir as regras de qualidade e transformação

Processos de *Data Profiling* e *Data Cleansing*

Data Cleansing (*Limpeza de Dados*):

- É o processo de detetar e corrigir ou remover registos incorretos, incompletos, imprecisos ou irrelevantes.

- Inclui a padronização de formatos, a correção de erros de digitação e a resolução de inconsistências.

Master Data Management (MDM)

- O **Master Data Management (MDM)** é uma disciplina que se foca na criação e manutenção de uma visão única, consistente e precisa dos dados de negócio mais críticos (*Master Data*) em toda a organização.

- O *Master Data* inclui entidades como Clientes, Produtos, Fornecedores e Localizações.

Master Data Management (MDM)

- **Relação com a Integração de Dados**

A integração de dados é o mecanismo que alimenta o MDM, e o MDM, por sua vez, garante a qualidade e a consistência dos dados integrados.

Estratégias de MDM

- Existem várias abordagens para implementar o MDM, que se refletem na forma como os dados mestres são integrados.

- A implementação de uma estratégia de MDM é um passo avançado na integração de dados, garantindo que as decisões de negócio são baseadas numa fonte de verdade única e fiável.

Estratégias de MDM

Estratégia	Descrição	Vantagem
Registry (Registo)	Cria um índice centralizado dos dados mestres, mas os dados reais permanecem nos sistemas de origem.	Implementação mais rápida e menos disruptiva.
Consolidation (Consolidação)	Os dados mestres são extraídos dos sistemas de origem, limpos, unificados e carregados num repositório central (o Hub MDM).	Fornece uma "visão dourada" única para análise e relatórios.
Coexistence (Coexistência)	Semelhante à Consolidação, mas o Hub MDM também sincroniza as alterações de volta para os sistemas de origem.	Garante que todos os sistemas operativos utilizam a mesma versão dos dados mestres.