2022 Segunda Conferência Internacional sobre Tecnologias Avançadas em Controle Inteligente, Meio Ambiente, Computação e Comunicação Engenharia ICATIECE

ETL, ELT e ETL reverso: um estudo de caso de negócios

Bharat Singhal
UPES
Dehradun, Índia
singhalbharat00@gmail.com

Alok Aggarwal
UPES
Dehradun, Índia
alok.aggarwal@ddn.upes.ac.in

Resumo **- Atualmente, a maioria das organizações depende fortemente de seu data warehouse para tomar decisões de nível empresarial. O Data Warehouse extrai dados de várias fontes heterogêneas e, portanto, ao configurar um data warehouse, existem três maneiras de processar dados: ELT (Extrair, Carregar e Transformar), ETL (Extrair, Transformar e Carregar) e ETL reverso. Pode ser um desafio selecionar a melhor abordagem ao decidir como implementar um data warehouse porque tem a ver com custos, procedimentos, desempenho e melhoria contínua da empresa. Neste artigo, discutiremos as três abordagens e seus casos de uso.**

Palavras-chave: ETL, ELT, Reverse-ETL

I. INTRODUÇÃO

Um data warehouse é um repositório de dados extraídos de várias fontes e organizados para relatórios e análises.
Os dados em um data warehouse geralmente estão em um formato desnormalizado, o que significa que não são normalizados no mesmo nível que os dados nos sistemas de origem [1]. Isso torna os dados no data warehouse mais fáceis de consultar e analisar.

Existem alguns métodos diferentes de processamento de dados em um data warehouse. O primeiro é chamado de ETL, ou extrair, transformar e carregar [2]. Este é o processo de extrair dados dos sistemas de origem, transformá-los em um formato adequado para o data warehouse e carregá-los no data warehouse. O segundo método é chamado ELT, ou extrair, carregar e transformar [3]. Este é o processo de extrair dados dos sistemas de origem, carregá-los no data warehouse e transformá-los.
Muitas outras abordagens foram desenvolvidas usando MANET, WSN e várias abordagens de ML [6]-[27].

1.1 ETL

Ao trabalhar com bancos de dados, é essencial planejar e projetar adequadamente as informações para que possam ser empilhadas em estruturas para armazenamento de informações. ETL é uma ferramenta de programação única que combina três recursos distintos, mas cruciais, para facilitar a preparação de dados e a administração do banco de dados. As funções de cada um dos três procedimentos serão visíveis abaixo.

Extract: As informações de um banco de dados de origem são analisadas e a melhor parte delas é extraída durante esse processo. O objetivo de essa progressão é extrair o máximo de informações possível da estrutura de origem, usando o mínimo de recursos possível. O processo de concentração deve ser planejado de forma que não afete adversamente o framework fonte em termos de execução ou tempo de reação [4].

Transform: A informação é peneirada e purificada durante este procedimento, que também prepara os dados removidos combinando-os com outros dados ou utilizando tabelas de consulta ou ferramentas de administração para devolvê-los ao seu estado original. A aprovação de registros, a rejeição de informações (se forem consideradas indignas) e a mistura de informações fazem parte da etapa de mudança. Um dos métodos mais comuns para a transformação da mudança é organizar, separar, limpar as duplicatas, institucionalizar, interpretar e olhar para cima ou verificar a consisté

Carga: Uma das etapas do processo é empilhar as informações no centro de distribuição de informações. Os dados subseqüentes, como os dados extraídos e modificados, são compilados pela capacidade de heap de maneira semelhante à de um cofre de dados objetivo. Vários dispositivos fazem interface com os formulários de extração, alteração e empilhamento para cada registro da fonte, enquanto outros incorporam fisicamente cada registro como outra coluna na tabela do banco de dados objetivo usando a explicação de incorporação SQL.

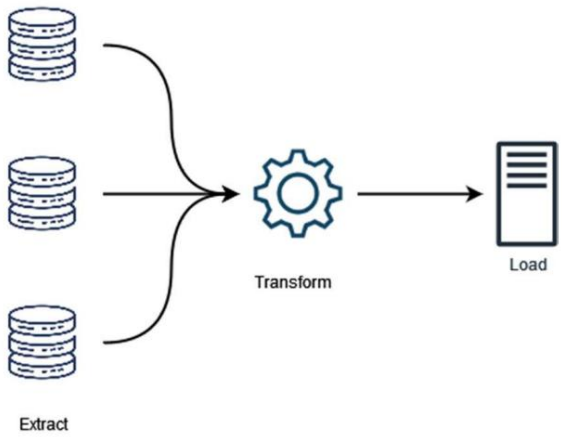

Fig. 1. Diagrama ETL

978-1-6654-9396-3/22/$31,00©2022 IEEE

1.2 ELT O

método Extrair, Carregar e Transformar (ELT) combina a codificação manual e o método ETL em um arranjo semelhante.
Semelhante à abordagem ETL, os dados são separados. Depois de remover seus dados, você começa rapidamente o estágio de empilhamento, que envolve a combinação de todas as fontes de dados em um único arquivo integrado [4]. Quando os dados são separados da fonte e colocados nas tabelas de organização, eles são uma duplicata grosseira.
Isso significa que os nomes dos segmentos permanecem os mesmos do banco de dados de origem e que você não adiciona novos campos de dados ou informações. No entanto, você pode canalizar linhas e seções desnecessárias ao remover informações para evitar o desperdício de recursos com informações desnecessárias. As estruturas agora seriam capazes de suportar grande capacidade e números flexíveis, graças aos atuais avanços da estrutura baseada em nuvem.

Como resultado, manter o controle de todas as informações brutas extraídas requer um conjunto de dados grande e em expansão, bem como uma preparação rápida. A estrutura do centro de distribuição de informações objetivas atualmente abriga as informações que foram extraídas de várias fontes. Usando drivers SQL locais, as alterações e justificativas de negócios são conectadas, que usam fluxos de informações para mover dados da fonte para as tabelas organizadoras.

Isso economiza dinheiro e reduz a quantidade de trabalho extra exigido pelo nível central ETL.

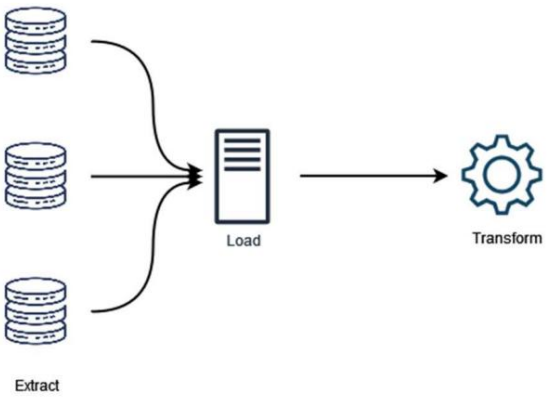

Fig. 2. Diagrama ELT

1.3 ETL reverso O

oposto de ETL/ELT é ETL reverso. O ETL reverso torna o data warehouse a origem em oposição ao destino. Para permitir a ação, os dados são extraídos do warehouse, processados para atender às necessidades de formatação de dados no destino e inseridos em um aplicativo para que possam ser usados por marketing, vendas, suporte e outras equipes nas ferramentas que usam.

Ao reinserir os dados nos sistemas de negócios, um ETL reverso "operacionaliza" os dados em uma empresa [5].

A esse respeito, um ETL reverso é utilizado em conjunto com outros pipelines de dados, em vez de substituir os pipelines ETL ou ELT.

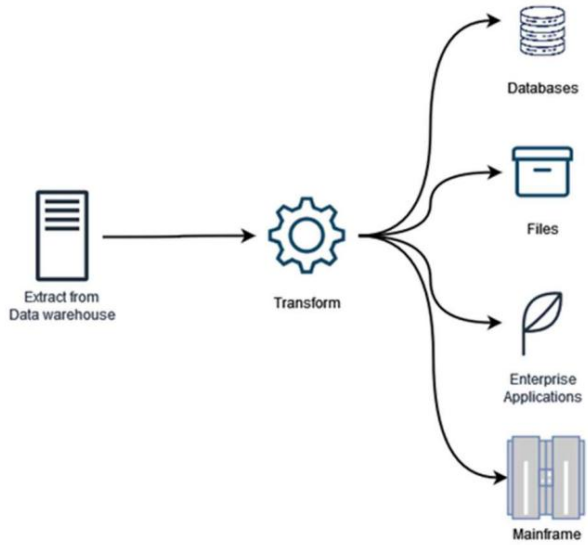

Fig. 3. Diagrama ETL reverso

II. CENÁRIOS DE NEGÓCIOS PARA ETL & ELT

A etapa mais difícil dos processos ETL e ELT é a transformação de dados. Os processos ETL são mais adequados para operações em lote com suporte transacional para fontes de dados monolíticas, muitas vezes legadas. Apesar de sua natureza OLTP (Online Transaction Processing), esses bancos de dados não inserem ou atualizam dados rapidamente. A maioria das modificações de dados é feita manualmente ou em lotes.
Normalmente, armazenamentos de dados não relacionais, como bancos de dados NoSQL, ou armazenamento de blob ou sistemas de arquivos Hadoop, são usados para ingerir dessas fontes de dados altamente dinâmicas e sem esquema. Para a carga de trabalho analítica, eles são transformados e processados em formato tabular.

Além disso, as decisões de negócios baseadas em dados devem ser tomadas em um tempo muito limitado. A frequência varia de semanal ou diária para a geração desses relatórios tradicionais de inteligência de negócios. Decisões baseadas em dados ao vivo são quase totalmente esperadas ao lidar com grandes volumes de dados de streaming.
Uma dessas ilustrações é uma transmissão ao vivo continuamente atualizada de dados do mercado de ações.

Em tais circunstâncias, os aplicativos ETL desaceleram rapidamente. Os processos de transformação param e as filas de entrada de dados ficam transbordando, o que faz com que o usuário espere por um período de tempo irracional. Uma abordagem alternativa para processamento de dados conhecida como ELT (Extração, Carregamento e Transformação) é utilizada para resolver esse problema.

A localização da transformação de dados é a principal distinção entre ETL e ELT. O ELT, ao contrário do ETL, não muda nada durante o trânsito. O banco de dados de back-end manipula a transformação. Isso indica que os dados são enviados diretamente para o data warehouse de destino dos sistemas de origem em uma área de preparação. Dentro do banco de dados, a lógica de negócios da transformação entra em ação. Depois de trabalhar com os dados brutos preparados, o processo de transformação copia os dados processados para uma área separada.

Em vez do idioma nativo do ETL, todo o processamento ocorre no idioma nativo do data warehouse do destino, portanto, nenhum trabalho pesado é necessário do pipeline. A ascensão e a adoção generalizada de data warehouses na nuvem são fatores adicionais que contribuem para a ascensão do ELT. Existem muitas opções para armazenamento de dados hospedado em nuvem gerenciado, onde tudo é tratado pelo provedor de nuvem, desde o dimensionamento do sistema até o gerenciamento do armazenamento ou hardware e qualquer requisito de instalação de software. O Amazon Redshift, que é um data warehouse comum hospedado na nuvem, pode ser provisionado em minutos. Snowflake, um data warehouse em nuvem, permite armazenamento de dados ilimitado e separação completa de computação e armazenamento.

Os data warehouses em nuvem são executados em máquinas poderosas com muita RAM, vários processadores de núcleo e armazenamento rápido em disco SSD. Mesmo se um ou mais nós falharem, os clusters de vários nós permanecerão online. Com base na configuração de carga, o número de nós pode aumentar ou diminuir dinamicamente. Estes também são conhecidos como sistemas com armazenamentos colunares para processamento massivamente paralelo (MPP). O mecanismo de armazenamento colunar possibilita a recuperação rápida de dados, e o recurso MPP permite que as consultas sejam processadas em paralelo em todos os núcleos da CPU de cada nó.

TABELA 1. DIFERENÇA ENTRE ETL e ELT

Parâmetro	ETL	ELT
Uso ideal	Dados estruturados, sistemas legados e bancos de dados relacionais; transformando os dados antes de carregá-los no Data Warehouse.	Carregamentos de dados mais rápidos e oportunos, dados estruturados e não estruturados e grande conjunto de dados; transformando dados conforme a necessidade
Privacidade	As informações de identificação pessoal podem ser eliminadas na etapa de transformação de pré-carga.	São necessárias grandes salvaguardas para a privacidade, uma vez que os dados são diretamente carregado.
Transformações Secundárias	servidor executa as transformações necessárias. A pré-limpeza e a transformação pesada de computação são ideais.	o Maior velocidade e eficiência são alcançadas, pois o banco de dados realiza transformações para carregar e transformar simultaneamente.
Manutenção	Alta manutenção devido à presença de vários servidores de processamento.	Carga de manutenção reduzida por causa de menos sistema.
Despesas	Problemas monetários devido a servidores separados	Menos sobrecarga monetária devido a pilhas de dados simplificadas.
Compatibilidade com Data Lake	A compatibilidade do data lake é não existe com ETL	Compatibilidade com data lake está lá com ELT.
Saída de dados	A saída é estruturada.	Saída pode ser Estruturada, não estruturada ou semiestruturada.
Quantidade de conjuntos de dados	de dados de pequenos e volume moderado.	Conjuntos de dados de grande volume.

III. CENÁRIOS DE NEGÓCIOS PARA ETL REVERSO

Em geral, marketing, vendas e suporte trabalham mais juntos quando se trata de ETL reverso. O ETL reverso auxilia na personalização de e-mail para marketing. Várias empresas usam boletins informativos para alcançar consumidores existentes e potenciais, normalmente desenvolvendo fluxos de email orientados por dados que podem variar em complexidade. O ETL reverso permite que as empresas importem dados de uso do produto para o Salesforce for Sales e os combinem para fornecer uma visão abrangente do uso do produto. Eles podem ver, por exemplo, quando alguém se inscreve, faz uma atividade específica ou gasta uma quantia específica de dinheiro. Para suporte, os agentes são mais capazes de ajudar os clientes tendo uma visão completa deles.

TABELA 2. DIFERENÇA ENTRE ETL/ELT e REVERSE ETL

Parâmetro ETL/ELT		ETL reverso
Sincronização Modo	Pode haver extração de dados completa ou incremental. Podemos usar a operação UPSERT para mesclar dados em um banco de dados ou data warehouse	O CDC é difícil de aplicar no ETL reverso, pois o warehouse normalmente não fornece colunas “updated_at” do log de transações. ou Temos que acompanhar o que precisa ser atualizado e o que precisa ser criado.
Transformações de dados	Vamos do específico ao geral. Extraímos dados de diferentes fontes específicas para depois integrá-los em um destino comum.	Vamos do geral ao específico, tendo que adequar cada API de aplicação de negócio.
Qualidade dos dados	Menos qualidade Sobrecarga de dados, pois o destino é banco de dados/armazém de dados	dados altos qualidade sobrecarga, pois são necessários mais validação e conhecimento do destino.
Falhas e reexecução do trabalho	Trabalhos ETL/ELT são idempotentes, o que significa que não importa com que frequência você os execute, eles devem produzir os mesmos resultados.	Os trabalhos de ETL reverso não são idempotentes, pois a reexecução pode resultar em efeitos colaterais indesejados, pois dependem da lógica de negócios do destino.

VI. CONCLUSÃO

ETL é um paradigma clássico. Funciona com infraestruturas convencionais de data center, que já estão sendo substituídas por tecnologias de nuvem. Como a infraestrutura já existente ou implantações específicas são muito mais inclinadas ao ETL, as grandes empresas ainda preferem esse método.

O ELT faz uso eficaz das tecnologias de nuvem atuais, tornando-o o futuro do armazenamento de dados. Ele fornece informações importantes que podem ajudar as empresas a tomar as decisões de negócios corretas e possibilita que as empresas analisem grandes conjuntos de dados com menos manutenção. À medida que as ferramentas nativas de integração de dados para soluções Hadoop e NoSQL continuam avançando, o escopo do ELT pode eventualmente se expandir.

O surgimento de uma pilha de dados de nova geração destaca uma tendência importante: as empresas devem incorporar recursos de dados dentro das equipes nas divisões de negócios, em vez de mantê-los em silos centralizados (armazéns de dados). As soluções ETL reversas que lidam com a operacionalização de dados, ou seja, fecham o ciclo analítico operacional, são, portanto, parte do futuro da pilha de dados moderna.

REFERÊNCIAS

[1] Kumar A. e outros. "Simulação e Análise de Protocolos de Autenticação para Internet Móvel das Coisas (MIoT)," *PDGC,* 2014, pp. 423-428.

[2] P. Gupta et al., "Algoritmo de programação baseado em confiança e confiabilidade para nuvem IaaS," *Lect. Notas em EE.* vol. 150, 2013, pp. 603-607.

[3] Govil Kapil et al., *"Técnica de seleção de cabeça de cluster para otimização da conservação de energia em MANET", PDGC,* 2014, pp. 39-42.

[4] *A. Kumar et al.,* "Primitivos criptográficos leves para redes ad hoc móveis", *RTCNDSSCCIS,* vol. 335, 2012, pp. 240-251.

[5] Kumar A. et al., "Análise de desempenho de MANET usando curva elíptica cryptosystem," *ICACT,* 2012, pp. 201-206.

[6] Singh T., Srivastava DK e Aggarwal A., "Uma nova abordagem para utilização de CPU em um paradigma multicore usando quicksort paralelo," *CICT,* 2017, pp. 1-6.

[7] Mittal S. et al., "Reconhecimento de situação em ambientes baseados em sensores usando redes conceituais," *IIT,* 2012, pp. 579-584.

[8] SK Gupta et al., "Algoritmo de roteamento para conservação de energia em MANET," *CICN,* 2015, pp. 165-167.

[9] Singh V. et al., "Uma abordagem holística, proativa e inovadora para validação pré, durante e pós-migração do subversion para git," *CMC,* vol. 66, no.3, pp. 2359-2371, 2021.

[10] S. Aggarwal et al., "Método otimizado de controle de potência durante soft handoff na direção de downlink de sistemas WCDMA," *PDGC,* 2014, pp. 433-438.

[11] Rajput IS et al., "Um algoritmo de busca paralela eficiente na rede de interconexão Hypercube," *PDGC,* 2012, pp. 101-106.

[12] Goyal MK et al., "Efeito da mudança na taxa do operador de algoritmo genético na composição de assinaturas para sistema de detecção de intrusão de uso indevido," *PDGC,* 2012, pp. 669-672.

[13] S. Aggarwal et al., "Análise de desempenho do algoritmo de transferência suave usando lógica difusa em sistemas CDMA", *PDGC,* 2012, pp. 586-591.

[14] MK Goyal et al. "Modelo de gerenciamento de confiança baseado em QoS para Cloud IaaS," *PDGC,* 2012, pp. 843-847.

[15] Kumar A., Krishan G. et al., "Projeto e análise de mecanismo de confiança leve para dados secretos usando primitivos criptográficos leves em MANETs", *Jour. de N/w Security,* vol. 18, não. 1, pp. 1-18, 2016.

[16] Bijalwan D. et al., "Reconhecimento automático de texto em cena natural e sua tradução em linguagem definida pelo usuário," *PDGC,* 2014, pp. 324-329.

[17] Aggarwal S. et al., "Sobre desafios e oportunidades na segunda onda da revolução das TIC para os países do sul da Ásia", *PDGC,* 2012, pp. 597-602.

[18] Singh V. et al. "Uma abordagem de transformação digital para arquitetura de microsserviços orientada a eventos residindo em vcs avançados", *CENTCON,* 2021, pp. 100-105.

[19] S. Aggarwal et al., "Tendências no controle de energia durante soft handoff na direção de downlink de redes celulares 3G WCDMA," *PDGC,* 2012, pp. 603-608.

[20] V. Singh et al., "Aspectos de migração baseados em DevOps do sistema de controle de versão legado para VCS distribuído avançado para implantação de microsserviços", *CSITSS,* 2021, pp. 1-5.

[21] Aggarwal S. et al., "Análise de transferência suave e seus efeitos na capacidade de downlink de redes celulares 3G CDMA", *PDGC,* 2012, pp. 1-6.

[22] S. Aggarwal et al., "Tendências no controle de energia durante soft handoff na direção de downlink de redes celulares 3G WCDMA," *Proc. PDGC,* 2012, pp. 603-608.

[23] Singh V. et al. "Uma abordagem de transformação digital para arquitetura de microsserviços orientada a eventos residindo em vcs avançados", *CENTCON,* 2021, pp. 100-105.

[24] Aggarwal S. et al., "Sobre desafios e oportunidades na segunda onda da revolução das TIC para os países do sul da Ásia", *PDGC,* 2012, pp. 597-602.

[25] S. Aggarwal et al., "Análise de desempenho do algoritmo de transferência suave usando lógica difusa em sistemas CDMA", *PDGC,* 2012, pp. 586-591.

[26] Rajput Iswar Singh et al., "Um algoritmo de busca paralelo eficiente em Rede de Interconexão Hipercubo," *PDGC,* 2012, pp. 101-106.

[27] Goyal MK et al., "Efeito da mudança na taxa do operador de algoritmo genético na composição de assinaturas para sistema de detecção de intrusão de uso indevido," *PDGC,* 2012, pp. 669-672.