Título: Implantação de ELT como alternativa ao ETL para transformação e análise de dados em grandes empresas

Capítulo 1 – Introdução (Pesquisar mais métodos de extração para fim de comparação)

* 1. Problema
  2. Objetivos

1.2.1 Objetivo geral

1.2.2 Objetivos específicos

Capítulo 2 – Referencial Teórico

Transformação de dados

Big Data

Data Lakes

Data Warehouses

ETL

ELT

Capítulo 3 - ?? (Comparativo)

4.1 ETL x ELT

4.1.1 Vantagens ETL

4.1.2 Desvantagens ETL

4.1.3 Vantagens ELT

4.1.4 Desvantagens ELT

4.2 Tempo de carregamento

4.3 Tempo de transformação

4.4 Tempo de manutenção

4.5 Complexidade de implementação

4.6 Limitação de dados

4.7 Suporte para data warehouses

INTRODUÇÃO

Nos últimos anos, temos observado uma explosão na quantidade de dados gerados em diversas áreas e setores. Esse grande volume de informações, conhecido como Big Data, tem potencial para fornecer insights valiosos e melhorar a tomada de decisão em organizações. No entanto, para que os dados sejam úteis, é preciso que eles sejam coletados, armazenados e processados de maneira adequada.

Para lidar com esses desafios, surgiram tecnologias como Data Warehouses, Data Lakes e processos de ETL (Extração, Transformação e Carga) ou ELT (Extração, Carga e Transformação), que permitem capturar, armazenar e processar grandes volumes de dados para análises.

Os Data Warehouses são sistemas de armazenamento projetados para suportar a análise de grandes quantidades de dados estruturados. Já os Data Lakes são sistemas mais flexíveis e escaláveis, que permitem armazenar dados de diferentes fontes e formatos, sem a necessidade de transformá-los em um formato estruturado antes de serem armazenados. Os processos de ETL e ELT, por sua vez, são fundamentais para carregar e transformar dados nos Data Warehouses e Data Lakes.

Segundo Rautenberg e Carmo (2019) “Big Data se trata de um conjunto de dados impeditivo de captura, armazenamento, gerenciamento e análise por parte de ferramentas computacionais tradicionais, requer formas inovadoras de processamento de grandes volumes de dados heterogêneos[...]”.

O ETL é um processo tradicional de transformação de dados formado por três etapas: a extração, a transformação e o carregamento de dados (BLASI, 2020). A principal vantagem do ETL, é que ele permite a criação de uma Data Warehouse unificado, que pode ser usado para relatórios e análises.

O ELT pode ser considerado a modernização do processo de ETL. Ao contrário do ETL, o ELT é um processo mais ágil para o carregamento e o processamento de dados (BLASI, 2020). A sua principal vantagem é que ele permite um processamento de dados mais rápido, pois esses dados são carregados no sistema de destino antes de serem transformados.

Bansal e Kagemann (2015), eles decrevem: “[...] A transformação inclui limpeza, racionalização e complementação dos registros, o processo de limpeza removerá erros e padronizará as informações e a complementação implicará no acréscimo de dados[...]”.

A escolha entre a ETL e ELT para a transformação dos dados, pode ter um impacto significativo no sucesso de um projeto de integração de dados. Compreender os fundamentos teóricos e as aplicações práticas dessas duas abordagens é essencial para tomar uma decisão informada.

PROBLEMA

As empresas estão enfrentando desafios cada vez maiores na gestão e análise de grandes volumes de dados devido ao aumento exponencial da coleta de dados. As soluções tradicionais de ETL e Data Warehouse foram amplamente utilizadas, mas estão começando a enfrentar limitações em termos de escalabilidade, complexidade, custos e flexibilidade. Como alternativa, a implantação de uma solução ELT e o uso de um Data Lake podem oferecer vantagens significativas para as empresas que precisam gerenciar e analisar grandes quantidades de dados não estruturados. No entanto, ainda há poucos estudos que exploram os benefícios e desafios dessas abordagens alternativas.

OBJETIVO GERAL

Comparar os métodos de extração de dados ETL e ELT e os sistemas de armazenamento Data Warehouse e Data Lake.

OBJETIVOS ESPECÍFICOS

- Examinar as principais diferenças entre ETL e Data Warehouse em relação ao ELT e Data Lake e como elas afetam o desempenho, a escalabilidade e a flexibilidade da integração de dados.

- Identificar os fatores que influenciam a escolha entre ETL e Data Warehouse em relação ao ELT e Data Lake, como volume, estrutura e complexidade dos dados.

REFERENCIAL TEÓRICO

Os fundamentos teóricos do ETL remontam ao conceito de armazenamento de dados, que enfatiza a importância da integração de dados para fins de tomada de decisão (Inmon, 2008). O ETL é baseado na teoria de que a integração dos dados deve ser executada antes que os dados sejam carregados em um sistema de destino para garantir a consistência e a precisão dos dados. Em contraste, o ELT, é baseado na teoria de que os dados devem ser carregados em um sistema destino primeiro e depois transformados conforme necessário. É importante destacar que os processos de ETL e ELT, apesar de diferentes, solucionam o mesmo problema BLASI(2020).

BIG DATA

O conceito de Big Data pode ser condensado com sendo um grande volume dados, sendo eles estruturados ou não estruturados, provenientes de diversas fontes, que devem ser gerenciados e analisados, o maior desafio é transformar todo o volume de dados em informação, gerando assim conhecimento e valor para as organizações (JUNIOR et al., 2016).

Arrais (2022) descreve:

“[...] Temos sensores em todos os lugares, crescimento da utilização de dispositivos conectados à rede, aumento do poder de armazenamento com um arquivamento quase infinito, nuvens de processadores com evoluções em todos os recursos computacionais. Todas essas habilidades em constante evolução resultam em uma geração massiva de dados, que vem mudando a ciência, a medicina, os negócios e a tecnologia. Todas essas informações, com grande volume e variedade, necessitam de formas inovadoras para ingestão, transformação, armazenamento e análise, pensando na riqueza que esses dados podem conter para a tomada de decisões.

Big Data é um conjunto de dados que devido a quantidade e a variedade não podem ser manipulados com as ferramentas computacionais tradicionais. Temos, portanto, um crescimento exponencial de dados, heterogêneos, oriundos de diferentes fontes, de forma distribuída e descentralizada. [...] ”.

Os 5 V s

O conceito de Big Data, já foi dividido em 3 Vs (volume, velocidade e variedade), com o passar do tempo os conceitos foram atualizados e hoje já pode-se encontrar referências que abordam o conceito com 5 Vs.



Fonte: <https://www.analyticsvidhya.com/blog/2021/05/what-is-big-data-introduction-uses-and-applications/> . Acesso em: 24 mar. 23.

Neste trabalho são utilizados os conceitos descritos por Arrais (2022):

- Volume: Big Data é formado por um conjunto de dados tão grande que é necessário ferramentas e estruturas especializadas para seu armazenamento, processamento e análise. Esse grande volume é gerado pelas indústrias, saúde, Internet das Coisas e demais sistemas de forma exponencial. Não existe um limite fixo para que uma quantidade de dados seja considerada como big data, mas normalmente são dados em grande escala que oferecem desafios para armazenamento, gerenciamento e processamento, necessitando de ferramentas não tradicionais.

- Velocidade: este conceito está relacionado com a rapidez em que os dados são gerados, o que influencia diretamente no crescimento exponencial dos dados e no volume alto para armazenamento. Algumas aplicações precisam analisar esses dados em tempo real, e, portanto, é necessário ter ferramentas especializadas para essa ingestão e análise com alta velocidade para decisões mais assertivas.

- Variedade: refere-se à variedade de dados que são gerados (dados estruturados, não estruturados e semiestruturados). Temos dados em formato de textos, imagens, áudios, vídeos, oriundos de sensores, de operações; e os sistemas precisam ser flexíveis para atender todos esses formatos e suas especificidades para que possa ser possível o tratamento e aquisição de informações a partir dos mesmos.

- Veracidade: este conceito se refere ao aspecto de confiabilidade dos dados. Para extrair valor, gerando informações de qualidade, é necessário que os dados sejam limpos e que ruídos (dados incorretos e/ou faltantes sejam identificados e/ou eliminados).

- Valor: refere-se à utilidade que os dados possuem para a finalidade pretendida, ou seja, para atender alguma necessidade ou resolver um problema. Esse valor está associado diretamente à veracidade e precisão dessas informações, podendo também em algumas aplicações depender da velocidade de processamento para tomada de decisões.

**Data WareHouse**

Data Warehouse é um repositório que armazena dados estruturados e semiestruturados para relatórios e análises. Assim como o Data Lake o Data Warehouse pode armazenar grandes quantidades de informações e ajuda as empresas na tomada de decisão.

Diagrama

Descrição gerada automaticamente

Fonte: <https://www.astera.com/pt/tipo/blog/defini%C3%A7%C3%A3o-de-data-warehouse/>. Acesso em : 04 abr. 23.

**Benefícios do Data Warehouse**

De acordo com Microsoft (2023):

* Consolidação de dados de várias fontes em uma única fonte de verdade.
* Armazenamento e análise de dados históricos de longo prazo abrangendo meses e anos.
* Limpando e transformando dados para que eles sejam precisos, consistentes e padronizados em estrutura e forma.
* Reduzindo os tempos de consulta ao coletar dados e processar análises, o que melhora o desempenho geral em todos os sistemas.
* Carregando dados com eficiência sem precisar lidar com os custos de implantação ou infraestrutura.
* Protegendo dados para que eles sejam privados, protegidos e seguros.
* Preparação de dados para análise por meio de mineração de dados, ferramentas de visualização e outras formas de análise avançada.

**Arquitetura do Data Warehouse**

Um data warehouse, pode ser considerado mais do que apenas um repositório de dados. Ele é um sistema altamente estruturado e cuidadosamente arquitetado composto por várias camadas que interage com seus dados e entre si de maneira diferente, essas camadas incluem Microsoft (2023):

Camada inferior :

Com o processo de extração, transformação e carregamento chamado ELT, responsável também pelos dados armazenados e otimizados onde o tempo de consulta é mais rápido tem melhor desempenho.

Camada intermediária:

Com o processamento analítico online chamado como servidor OLAP, que acessam grandes volumes de dados do Data Warehouse onde os resultados são extremamente rápidos.

Camada superior:

É onde apresenta visualmente os dados processados e também é a camada mais utilizada pelos analistas que têm acesso e pode usar para a necessidade de relatórios e BI de autoatendimento.

**As principais características de um Data Warehouse** (ROCK CONTENT, 2021)

* Gestão dos dados
* Integração dos dados
* Não volatilidade dos dados

**Benefícios do Data Warehouse** (ROCK CONTENT, 2021)

* Agilidade nas consultas
* Dados de qualidade
* Segurança
* Visão histórica
* Escalabilidade
* Autonomia para os funcionários

Data Lake

Data Lake é um repositório que armazena todos os dados estruturados, não estruturados e semiestruturados.

Ajuda muitas empresas na tomada de decisão, as empresas que implementaram os data lakes superaram 9% a performance de empresas semelhantes no crescimento orgânico da receita (AWS, 2023).

Círculo

Descrição gerada automaticamente com confiança média

Fonte: <https://aws.amazon.com/pt/big-data/datalakes-and-analytics/what-is-a-data-lake/>, Acesso em 03 abr. 23.

**Os elementos essenciais de uma solução de Data Lakes e análise.**

As empresas que estão criando Data Lakes e uma plataforma de análise precisam de vários recursos importante como:

* Movimentação de dados

Os Data Lakes permitem importar qualquer quantidade de dados em tempo real economizando tempo na definição de estrutura de dados, esquemas e transformação (AWS, 2023).

* Armazene e catalogue dados com segurança

Os Datas Lakes permitem armazenar dados relacionais, como  banco de dados operacionais e dados de aplicações de linha de negócios, dados não relacionais também como aplicativos móveis, dispositivos IoT e mídias sociais. Tem capacidade de atender dados por meio de crawling, catalogação e indexação de dados. Os dados tem que ser protegidos para garantir que os ativos dados sejam protegidos (AWS, 2023).

* Análises

Os Data Lakes permitem executar análises sem a necessidade de mover os dados para um sistema de análise separado. Permitindo ter várias funções dentro de uma empresa, como cientistas de dados, desenvolvedores de dados e analistas de negócios, acessando dados de ferramentas e frameworks analíticos. Incluindo frameworks de código aberto, como Apache Hadoop, Presto, Apache Spark e também ofertas comerciais  de fornecedores de data WareHouse e inteligência empresarial (AWS, 2023).

* Machine Learning

O Data Lake permite que a empresa gere diferentes tipos de situações, incluindo relatórios de dados históricos e machine learning, criando modelos para prever resultados sugerindo uma série de ações prescritas para chegar no resultado ideal (AWS, 2023).

**Alguns exemplos onde os Data Lakes agregam valores:**

* Melhores interações com o cliente

O Data Lake ajuda na combinação de dados do cliente em uma plataforma CRM com análise de mídia social, uma plataforma de marketing para capacitar a empresa a entender o grupo de clientes mais lucrativos e a causa da perda de cliente, as promoções e recompensas que aumentam a fidelidade (AWS, 2023).

* Aumento da eficiência operacional

O Data Lake facilita o armazenamento e execução de análise em dados de IoT gerados por máquina para descobrir soluções para reduzir custos operacionais e aumentar a qualidade.

A Internet das Coisas (IoT) apresenta várias maneiras de coletar dados sobre o processo, como fabricação de dados em tempo real com dispositivos conectados à internet (AWS, 2023).

Os desafios do Data Lakes são os dados brutos que são armazenados sem supervisão do conteúdo.

O Data Lakes torna os dados utilizáveis, precisa de um mecanismo definido para catálogo e proteger os dados, sem isso os dados não podem ser encontrados ou serem confiáveis. Para atender as necessidades do público mais amplo o data lakes tem que ter governança, consistência semântica e controle de acesso (AWS, 2023).

O Data Lakes é ideal para ser implantado na nuvem, pois oferece uma performance, escalabilidade, confiabilidade, disponibilidade, mecanismo analítico e enormes economias. Os principais motivos que os clientes perceberam a nuvem como vantagem para data lakes são: segurança, implantação mais rápida, disponibilidade, atualização de recursos, funcionalidades, elasticidade, cobertura geográfica e custo vinculados à utilização real (AWS, 2023).

**A arquitetura do Data Lakes:**

O armazenamento e o processamento de dados trabalham juntos para criar uma arquitetura em camadas coesivamente, que é informada por Big Data e executada sobre o Data Lakes. A arquitetura pode formar uma estrutura operacional de um data lakehouse. Cada empresa tem uma configuração própria e exclusiva, porém a maioria das arquiteturas de data lakehouse tem o seguinte (AWS, 2023):

* Gerenciamento e orquestração de recursos
* Conectores para fácil acesso
* Análise confiável
* Classificação de dados
* Extrair, carregar, transformar(ELT) processos
* Segurança e suporte
* Governança e administração

**ETL (Extract, Transform, Load)**

É um pipeline de dados usados para coletar dados de várias fontes, após a coleta vem a transformação dos dados de acordo com as regras de negócios e o carregamento dos dados em um armazenamento de dados de destino.

A transformação de dados que ocorre geralmente envolve várias operações, como filtragem, classificação, agregação, junção de dados, limpeza de dados, desduplicação e validação de dados. As três fases do ETL são executadas em paralelo para economizar tempo Microsoft (2023).

**Processo de ETL** Aws (2023)

1. Extração de dados relevante do banco de dados de origem
2. Transformação dos dados para que sejam mais adequados a análises.
3. Carregamento dos dados no banco de dados de destino.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Fonte: <https://learn.microsoft.com/en-us/azure/architecture/data-guide/relational-data/etl>, Acesso em 11 abr. 23.

**Como funciona a extração de dados ?**

A extração de dados, copiam ou extraem dados brutos de várias fontes e armazenam na área de preparação onde os dados extraídos são armazenados temporariamente, isso significa que o conteúdo armazenado é excluído após a conclusão da extração dos dados. A área de preparação pode reter um arquivo de dados para fins de solução de problemas Aws (2023).

A extração de dados acontece de uma das três maneiras :

* Notificação de atualização
* Extração gradual
* Extração completa

**Como funciona a transformação de dados ?**

A transformação de dados consolida e transforma os dados brutos na área de preparação onde vão ser preparados para o data Warehouse de destino.

A transformação de dados pode envolver os seguintes tipos de alteração de dados:

* Transformação de dados básica

  - Limpeza de dados

- Eliminação de duplicação de dados

- Revisão de formato de dados

* Transformação de dados avançada

  - Derivação

- Junção

- Separação

- Resumo

- Encriptação

**Como funciona o carregamento de dados ?**

O carregamento de dados, movem os dados transformados da área de preparação para o data warehouse de destino Aws (2023).

O carregamento de dados tem dois métodos:

* Carregamento completo
* Carregamento Incremental

 - Carregamento incremental por transmissão

 - Carregamento incremental em lotes

**ELT ( Extract, Load, Transform)**

É um pipeline que a transformação ocorre no armazenamento de dados de destino , sem ter a necessidade de usar um mecanismo de transformação separado. Os recursos de processamento do armazenamento de dados de destino são usados para transformar os dados Microsoft (2023).

 O processo de ELT se tornou popular com a adoção da infraestrutura em nuvem, que oferece aos bancos de dados de destino o poder de processamento necessário para as transformações AWS (2023).

Tela de celular com texto preto sobre fundo branco

Descrição gerada automaticamente com confiança média

Fonte: <https://learn.microsoft.com/en-us/azure/architecture/data-guide/relational-data/etl>, Acesso em 11 abr. 23.

**Processo de ELT** Blasi (2020)

1. Extração: faz extração de dados brutos de diversas fontes para a integração em um repositório de dados único.
2. Carregamento: o carregamento dos dados coletados em um data Warehouse ou repositório de dados.
3. Transformação: Na transformação dos dados brutos em dados modelados dentro de um data Warehouse para a aplicação de business intelligence, análise de dados e advanced analytics.

REFERÊNCIAS

BANSAL, Srividya K. Towards a Semantic Extract-Transform-Load (ETL) Framework for Big Data Integration. 2014 IEEE International Congress on Big Data, [s. l.], 2014.

Inmon, W.H., Strauss, D. & Neushloss, G. (2008). DW 2.0: The Architecture for the Next Generation of Data Warehousing. Morgan Kaufmann.

PAUNCZ, Alex. **ETL vs. ELT: Which is Right for Your Data Warehouse?** 2021. Disponível em: https://www.cdata.com/blog/20210706-etl-vs-elt?kw=&cpn=2023385644&utm\_source=google&utm\_medium=cpc&utm\_campaign=CData\_-\_Search\_-\_Branding\_-\_DSA&utm\_content=Branding\_-\_DSA&utm\_term=|&kw=&cpn=2023385644&gclid=Cj0KCQiAxbefBhDfARIsAL4XLRq3JP\_lo9i\_9EkvWfp2MfeKGGo6tQZZ8Vq7O0HjCcs0lZVzuOso-zIaAirnEALw\_wcB. Acesso em: 14 mar. 2023.

BLASI, Isabela. **ETL X ELT: qual a diferença?** 2020. Disponível em: https://blog.indicium.tech/etl-vs-elt-diferencas/?utm\_source=Google&utm\_medium=cpc&utm\_term=&utm\_campaign=19229929630&utm\_content=&gclid=Cj0KCQiAxbefBhDfARIsAL4XLRpHfjmOB5-JAR-YWRjnhhC1TvnUQNWYbNgtbLE8Rzm5dxDS9LglxFwaAtQCEALw\_wcB. Acesso em: 14 mar. 2023.

FÁTIMA, Nida. **ETL vs. ELT: Qual é a diferença?** 2020. Disponível em: https://www.astera.com/pt/type/blog/etl-vs-elt-whats-the-difference/. Acesso em: 14 mar. 2023.

MAGNUM, Lucas. **Engenharia de Dados — EL, ETL e ELT**: abordagens de extração de dados de forma simplificada. Abordagens de extração de dados de forma simplificada. 2021. Disponível em: https://lucasmagnum.medium.com/engenharia-de-dados-el-etl-e-elt-b42142058c87. Acesso em: 14 mar. 2023.

RAUTENBERG, Sandro; CARMO, Paulo Ricardo Viviurka do. Big data e ciência de dados. **Brazilian Journal Of Information Science**: research trends, [S.L.], v. 13, n. 1, p. 56-67, 29 mar. 2019. Faculdade de Filosofia e Ciências. <http://dx.doi.org/10.36311/1981-1640.2019.v13n1.06.p56>.

GARCIA, Marco. **Big Data**: o que é, conceito e definição. O que é, conceito e definição. 2022. Disponível em: https://cetax.com.br/big-data/. Acesso em: 22 mar. 2023.

JUNIOR, Jose Carlos Da Silva Freitas; MAÇADA, Antonio Carlos Gastaud; OLIVEIRA, Mirian; BRINKHUES, Rafael Alfonso. BIG DATA E GESTÃO DO CONHECIMENTO: DEFINIÇÕES E DIRECIONAMENTOS DE PESQUISA. **REVISTA ALCANCE**, [*S. l.*], p. 04-22, 1 nov. 2016.

ARRAIS, Karolayne Fernandes. **CONSTRUÇÃO DE UM PIPELINE DE DADOS UTILIZANDO SERVIÇOS DA NUVEM**. 2022. 91 f. TCC (Graduação) - Curso de Bacharelado em Engenharia de Computação, Centro de Ciências Exatas e de Tecnologia, Universidade Federal de São Carlos, São Carlos - Sp, 2022.

BAHGA, A.; MADISETTI, V. **Big Data Science & Analytics:** A Hands-On Approach. Published by Arshdeep Bahga & Vijay Madisetti. 2019.

MICROSOFT. **O que é um data warehouse?**: saiba o que é data warehouse, os benefícios de usar um, as práticas recomendadas a serem consideradas durante a fase de design e quais ferramentas incorporar quando finalmente for a hora de criar. Disponível em: https://azure.microsoft.com/pt-br/resources/cloud-computing-dictionary/what-is-a-data-warehouse/#get-started. Acesso em: 04 abr. 2023.

ROCK CONTENT. **O que é um Data Warehouse e quais são as suas principais características?** 2021. Disponível em: https://rockcontent.com/br/blog/data-warehouse/. Acesso em: 05 abr. 2023.

MICROSOFT. **O que é Data Lake**: veja como os data lakes diferem de data warehouses e data lakehouses. descubra como criar uma base escalonável para todas as suas análises com o azure. Disponível em: https://azure.microsoft.com/pt-br/resources/cloud-computing-dictionary/what-is-a-data-lake/#what-is-a-data-lake. Acesso em: 03 abr. 2023

AWS. **O que é um data lake?**: armazene todos os seus dados em um repositório centralizado em qualquer escala. Armazene todos os seus dados em um repositório centralizado em qualquer escala. Disponível em: https://aws.amazon.com/pt/big-data/datalakes-and-analytics/what-is-a-data-lake/. Acesso em: 04 abr. 2023.

MICROSOFT. **Extract, transform, and load (ETL)**. Disponível em: https://learn.microsoft.com/en-us/azure/architecture/data-guide/relational-data/etl. Acesso em: 11 abr. 2023.

AWS. **O que é ETL?** Disponível em: https://aws.amazon.com/pt/what-is/etl/. Acesso em: 11 abr. 2023.