INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO CÂMPUS CAMPINAS

DEIVE AUDIERES LEAL

**AVALIAÇÃO DE ESTRATÉGIAS PARA A CRIAÇÃO DE PIPELINE DE DADOS VOLTADO PARA CIÊNCIAS HUMANAS**

CAMPINAS

2021

DEIVE AUDIERES LEAL

**AVALIAÇÃO DE ESTRATÉGIAS PARA A CRIAÇÃO DE PIPELINE DE DADOS VOLTADO PARA CIÊNCIAS HUMANAS**

Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas.

Orientador: Prof. Dr. Samuel Botter Martins.

CAMPINAS

2021

**Dados Internacionais de Catalogação na Publicação**

Solicitar a ficha catalográfica pelo sistema Pergamum, (Meu Pergamum, solicitações, ficha catalográfica) após as correções sugeridas pela banca. Apresenta-se no verso da página de rosto

Deive Audieres Leal

**Avaliação de estratégias para a criação de pipeline de dados voltado para ciências humanas**

Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo Câmpus Campinas.

Aprovado pela banca examinadora em: \_\_\_\_\_ de \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ de \_\_\_\_\_.

BANCA EXAMINADORA

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Dr. Samuel Botter Martins (orientador)

IFSP Câmpus Campinas

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof.a Me xxxxxxxxxxxxxxxx

Universidade Estadual de Campinas

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Me xxxxxxxxxxxxxxxx

IFSP Câmpus XXXX

*Dedico este trabalho aos meus familiares,*

*colegas de classe, professores e servidores do Instituto*

*que colaboraram em minha jornada formativa.*

**AGRADECIMENTOS**

Agradeço primeiramente a Deus, pelo dom da vida e pela oportunidade de concluir mais uma etapa de minha experiência acadêmica. Agradeço a todos os professores e servidores do IFSP Campus Campinas, que contribuíram direta e indiretamente para a conclusão desse trabalho. Agradeço também à minha família, que deu todo o apoio necessário para que eu chegasse até aqui. Agradeço ao meu orientador que me auxiliou a solucionar as dificuldades encontradas no caminho.

*"O que prevemos raramente ocorre;*

*o que menos esperamos*

*geralmente acontece".*

*Benjamin Disraeli*

**RESUMO**

O resumo é a apresentação concisa dos pontos relevantes de um documento. O resumo de um trabalho acadêmico deve ser do tipo informativo, para que o leitor conheça a finalidade, metodologia, resultados e conclusões do documento, de tal forma que este possa dispensar a consulta ao original. O resumo deve ressaltar o objetivo, o método, os resultados e as conclusões. É composto por uma sequência de frases concisas, afirmativas e não pode ter enumeração de tópicos, recomendando-se parágrafo único. A primeira frase deve ser significativa, explicando o tema principal do documento. Segue-se indicando a categoria do trabalho, como memória, estudo de caso, análise de situação dentre outros. Deve-se usar o verbo na voz ativa e na terceira pessoa do singular e evitar o uso de símbolos e contrações; fórmulas, equações e diagramas apenas quando absolutamente necessário, seguidos de sua definição na primeira vez em que é mencionado. Quando o resumo não é inserido no próprio documento, deve ser precedido de sua referência. Sua extensão é de 150 a 500 palavras. As palavras-chave devem figurar logo abaixo do resumo, antecedidas da expressão Palavras-chave: e separadas entre si por ponto e finalizada por ponto.

**Palavras-chave**: Xxxxxxx. Xxxxxxx. Xxxxxxx. Xxxxxxx. Xxxxxxx

**ABSTRACT**

Versão do resumo em língua inglesa.

**Keywords**: Xxxxxxx. Xxxxxxx. Xxxxxxx. Xxxxxxx. Xxxxxxx.

**LISTA DE FIGURAS**

Figura 1 – Título............................................................................................................... 15

Figura 2 – Título............................................................................................................... 26

Figura 3 – Título............................................................................................................... 33

**LISTA DE TABELAS**

Tabela 1 – Título............................................................................................................... 15

Tabela 2 – Título............................................................................................................... 26

Tabela 3 – Título............................................................................................................... 30

Tabela 4 – Título............................................................................................................... 32

**LISTA DE ABREVIATURAS**

abr. abril

adapt. adaptação

bibl. biblioteca

cap. capítulo

**LISTA DE SIGLAS**

ELT *Extract, Load, Transform*

ETL *Extract, Transform, Load*

OLAP *Online Analytical Processing*

**LISTA DE SÍMBOLOS**

cm centímetro

km quilômetros

m2 metro quadrado

**SUMÁRIO**

[1 INTRODUÇÃO 15](#__RefHeading___Toc3718_1625992954)

[2 OBJETIVOS 16](#__RefHeading___Toc3720_1625992954)

[2.1 Objetivo Geral 16](#__RefHeading___Toc3722_1625992954)

[2.2 Objetivos Específicos 16](#__RefHeading___Toc3724_1625992954)

[3 FUNDAMENTAÇÃO TEÓRICA 17](#__RefHeading___Toc3726_1625992954)

[3.1 Engenharia de Dados 17](#__RefHeading___Toc4329_1625992954)

[3.2. ETL e ELT 17](#__RefHeading___Toc4331_1625992954)

[3.3. Data Warehouse 19](#__RefHeading___Toc4338_1625992954)

[4 METODOLOGIA 21](#__RefHeading___Toc3730_1625992954)

[5 DESENVOLVIMENTO 22](#__RefHeading___Toc3732_1625992954)

[6 CONSIDERAÇÕES FINAIS 23](#__RefHeading___Toc3734_1625992954)

[REFERÊNCIAS 24](#__RefHeading___Toc3736_1625992954)

# 1 INTRODUÇÃO

Com o aumento acelerado de produção de dados e o amplo uso dentro dos mais diversificados cenários e objetivos, a engenharia de dados se destaca no sentido de manter e criar uma base sólida de dados, que então poderão ser consumidos, seja por cientistas de dados ou engenheiros de aprendizado de máquina ou analistas de dados. O engenheiro de dados tem entre suas atribuições a construção e manutenção de arquiteturas direcionadas a dados, tais como banco de dados e sistemas de processamento em grande escala. Dentro da gama de possibilidades, a ETL 一 do inglês, Extract, Transform, Load 一 ou extração, transformação e carregamento é uma das etapas que merece destaque.

Nessa etapa, a extração corresponde ao agrupamento das mais diversas fontes de dados. Nela os dados são recuperados em sua forma crua, seja em bancos relacionais, planilhas, arquivos, ou outras formas de armazenamento de dados. Por sua vez, a transformação visa tornar os dados aptos para a utilização. Assim, são utilizadas técnicas de limpeza, classificação 一 categóricos e contínuos 一, e normalização sobre os dados. Por fim, o carregamento dos dados dá um local de acesso ao consumidor dos dados (por exemplo, o cientista de dados). Exemplos de locais de acesso são API – Application Programming Interface – e data warehouse.

Um data warehouse é um banco de dados estruturado voltado para consultas avançadas, com grande quantidade de dados históricos e direcionado para suportar análises. Nele são centralizados uma grande quantidade de dados advindos de diferentes fontes que passaram pelo processo de ETL.

Sendo uma etapa tão importante e que tem uma massa de dados que pode chegar à casa de terabytes, foram surgindo diferentes ferramentas que auxiliassem no processo de ETL. Assim, temos desde ferramentas open source a ferramentas proprietárias, mas todas voltadas para maximizar e melhorar as informações que são angariadas. Visando o uso de ferramentas open source para a execução do ETL, há uma ausência de estudos que comparem as diferentes soluções existentes, seus prós e contras, bem como quais cenários cada ferramenta é indicada.

# 2 OBJETIVOS

### **2.1 Objetivo Geral**

Desenvolver pipeline de dados via ELT para Big Data com ferramentas de integração de dados com enfoque em ciências humanas.

### **2.2 Objetivos Específicos**

* Desenvolver um processo de ELT aplicável e acessível
* Armazenar em ambiente de *big data*
* Utilizar unicamente ferramentas open source
* Integrar dados para que estejam apropriados para análise ou consumo final

# 3 FUNDAMENTAÇÃO TEÓRICA

### 3.1 Engenharia de Dados

Entendida como uma subárea dentro da ciência de dados, a engenharia de dados é um campo que se consolida como sendo essencial em ambientes direcionados a dados, e que o cenário exige tratamento de grandes volumes de dados brutos e disponibilidade.

O engenheiro de dados trabalha na garantia de que os dados coletados possam ir da fonte ao consumidor através de fluxos automáticos ou não, mas que facilitam o acesso aos dados.

Como tal, espera-se que tenha conhecimento de banco de dados relacionais, não-relacionais, linguagem de programação, em especial SQL, entendam como otimizar e recuperar dados. Seu trabalho é direcionado de acordo com as regras para o qual desenvolve o pipeline, fluxo, de dados.

WHITE (2018), observa que o engenheiro de dados pode se “encaixar” em três funções principais. A primeira, generalista, é responsável desde a aquisição de dados até a análise e visualização. A segunda é centralizada em pipeline, e trabalham juntamente com os cientistas de dados na utilização dos dados coletados. Por fim, há o que é centralizado no banco de dados, sua dedicação é concentrada no banco de dados analítico, trabalham com o desenvolvimento de esquemas, tabelas e com data warehouses.

Tamir et all (2015) dizem que o engenheiro de dados “torna possível pegar uma grande quantidade de dados e traduzir em insights". Para eles, engenharia de dados, “(...) não é simplesmente sobre a manutenção de um repositório para grandes volumes de dados, mas é sobre criar possibilidades para todos, desde desenvolvedores, passando por cientistas de dados e chegando aos executivos”[[1]](#footnote-2).

Entre as atividades desempenhadas pelo engenheiro de dados, destaca-se a extração, transformação e carregamento de dados. Na bibliografia é comum encontrarmos a sigla em inglês ETL, – Extract, Transform, Load –, assim também a usaremos.

### 3.2. ETL e ELT

O processo de ETL é a etapa em que os dados são extraídos da fonte inicial, seja elas um banco ou diversos bancos, com alta ou baixa granularidade, recebem então tratamento, se necessário, e é realizado o carregamento no repositório.

A extração é a primeira fase do ETL. Nela é realizada a extração de diversas fontes de dados. As diferentes fontes têm características próprias que são gerenciadas e só então realiza-se a extração. Deve-se observar a estrutura das fontes para que possam ser ajustadas ao repositório que as receberá. Os dados são recuperados em sua forma crua, seja em bancos relacionais ou não, planilhas, arquivos, e assim por diante. É um processo que, por sua vez, pode ser separado em dois outros, a importação dos dados em si e a atualização dos dados periodicamente.

Por sua vez, a transformação se ocupa em tornar os dados aptos para a utilização. Assim, são utilizadas técnicas de limpeza, classificação 一 categóricos e contínuos 一, e normalização sobre os dados. Assim sendo, tudo o que não condiz com o que é esperado passam por “um processo de exclusão ou tratamento de acordo com as regras de negócio da aplicação de destino, solucionando-as para garantir confiabilidade ao processo de ETL” (Lira Filho, 2013, p. 24). Dados também podem ser convertidos, padronizados e tipificados para se adequarem às regras de negócio observadas pelo engenheiro de dados.

Por fim, a última fase do ETL, é o carregamento dos dados. Nela os dados são gravados em um repositório de destino que permite o acesso ao consumidor. No caso de um data warehouse, os dados são carregados em tabelas dimensão e tabelas fato, que fazem parte do cubo de dados OLAP.

O cubo OLAP é um tipo de estrutura que expande os limites dos bancos de dados relacionais amplamente utilizados, melhorando a velocidade para a análise dos dados. Eles têm a capacidade de mostrar e executar cálculos de grandes quantidades de dados ao mesmo tempo que fornecem ao usuário meios de trabalhar tais dados. Assim sendo, as possibilidades de utilização analítica aumentam consideravelmente através da concentração, divisão e inclusão de dados de acordo com a demanda e os questionamentos que o uso dos dados podem vir a dar suporte para responder.

Com relação ao ETL há uma variação, o ELT, onde inverte-se a carga dos dados com a transformação. Nesse modelo, os dados são extraídos da fonte primária e carregados em uma base que aglutinará os diversos tipos de dados. Com os dados armazenados e acessíveis o sistema poderá então realizar as transformações e normalizações necessárias de modo mais otimizado, pois os dados estarão no mesmo ambiente.

Em nosso estudo, nos ocuparemos de criar um pipeline de dados voltado para o primeiro tipo, ETL, não sendo então tratadas diretamente maneiras de trabalhar em um processo de ELT de dados, embora em muitas oportunidades poderão ser aplicadas as mesmas tratativas.

O data warehouse será então o destino final dos dados processados pelo ETL, torna-se então importante entender o que é esse repositório a que nos referimos.

### 3.3. Data Warehouse

Surgindo de um conceito acadêmico da década de 1980, o Data Warehouse , DW, vem suprir a necessidade de organizar e obter dados de bases transacionais, que realizam apenas a transação de inserir, atualizar e excluir nos bancos de dados relacionais, como registros de vendas de mercadorias, e favorecer a análise de dados de negócio ou criação de relatórios de gestão.

O data warehouse é compreendido como parte de um planejamento para a organização eficiente dos dados e direcionado para dar suporte às estratégias empresariais. Os dados armazenados são mantidos por um longo período, permitindo que as análises desses dados, previamente consolidados, agrupados e indexados, possam ser estendidas na dimensão tempo. Proporciona o gerenciamento de um grande volume de dados que estariam, a princípio, dispersos em múltiplas fontes, sendo então centralizados, e a análise de dados coletados através de sistemas transacionais.

Por sua vez, complementando no que se relaciona a análise de dados, o OLAP, Online Analytical Processing, favorece a visualização e exploração dos dados. Tais tecnologias são voltadas para sustentar consultas sofisticadas e incomuns, assim como combinar informações por meio de análises comparativas e visualizações customizadas através do tempo. Tais tecnologias possibilitam observações claras de dados multidimensionais, através de compreensões múltiplas e variadas.

Os DW são modelados pensando que o sistema está em constante crescimento e que os dados devem estar integrados, são orientados por assunto e posições históricas. Os dados não são atualizados ou excluídos no DW, mas apenas inseridos iterativamente. Os DW podem ser ainda divididos em data mart, que são subconjuntos voltados para um tema, assunto, específico

As abordagens normalizada e dimensional são as mais usadas para o armazenamento. Na perspectiva normalizada as tabelas são agrupadas por “áreas temáticas que refletem categorias de dados gerais” (RASLAN e CALAZANS, 2014, p.34). Já na perspectiva dimensional os dados são divididos em fatos, os dados numéricos da transação, e em dimensões, informações de referência que contextualizam os fatos.

# 4 METODOLOGIA

# 5 DESENVOLVIMENTO

# 6 CONSIDERAÇÕES FINAIS

Parte do texto que apresenta os resultados correspondentes aos objetivos ou hipóteses levantados na introdução. O mesmo se aplica se o resultado é a proposta de um produto ou processo. Descreve de forma resumida o que se aprendeu sobre o tema, até mesmo propostas para outros trabalhos referentes ao assunto. Deve estar coerente com o desenvolvimento e relacionado à introdução. Pode ainda estabelecer relações com outros fatos referentes à mesma matéria. Em trabalhos acadêmicos, o termo Conclusão é substituído por Considerações Finais.

# REFERÊNCIAS

EL-SAPPAGH, Shaker H. Ali; HENDAWI, Abdeltawab M. Ahmed; BASTAWISSY, Ali Hamed El. **A proposed model for data warehouse ETL processes**. Journal Of King Saud University - Computer And Information Sciences, [S.L.], v. 23, n. 2, p. 91-104, jul. 2011. Elsevier BV. Disponível em: https://www.sciencedirect.com/science/article/pii/S131915781100019X. Acesso em: 27 de maio de 2021.

GARCIA, Marco. **Data Enginner ou Engenheiro de Dados - Conheça mais sobre**. Disponível em: https://www.cetax.com.br/blog/data-engineer-ou-engenheiro-de-dados/. Acesso em: 24 de maio de 2021.

GOUVEIA, Roberta Macêdo Marques; FREITAS, Charles Nicollas Cavalcante. **Implementação de um Data Warehouse para Análise de Dados Abertos Governamentais da Educação à Distância**. Revista de Educação Ciência e Tecnologia, Canoas, v. 7, n. 2, 2018. Disponível em: https://periodicos.ifrs.edu.br/index.php/tear/article/view/3037/2112. Acesso em: 25 maio 2021.

LIRA FILHO, Hermanny Alexandre dos Santos. **Análise comparativa das Ferramentas de ETL – Kettle e Talend**. 2013. 79 f. TCC (Graduação) - Curso de Bacharelado em Sistemas de Informação, Universidade Federal da Paraíba, Rio Tinto, 2013. Disponível em: https://repositorio.ufpb.br/jspui/handle/123456789/17132?locale=pt\_BR. Acesso em: 29 maio de 2021.

MICROSOFT. **Visão geral dos cubos OLAP do Service Manager para análise avançada**. 2018. Disponível em: https://docs.microsoft.com/pt-br/system-center/scsm/olap-cubes-overview?view=sc-sm-2019. Acesso em: 30 maio de 2021.

RASLAN, Daniela Andrade; CALAZANS, Angélica Toffano Seidel. **Data Warehouse: conceitos e aplicações**. Universitas: Gestão e TI, Brasília, ago. 2014. Centro de Ensino Unificado de Brasília. Disponível em: https://www.researchgate.net/publication/287517542. Acesso em: 25 de maio 2021.

SAS. **ETL: o que é e qual sua importância?**. [20--]. Disponível em: https://www.sas.com/pt\_br/insights/data-management/o-que-e-etl.html. Acesso em: 30 maio 2021.

SENE, Allan. **O que faz um Engenheiro de Dados?**. 2018. Disponível em: https://medium.com/data-hackers/o-que-faz-um-engenheiro-de-dados-fdcb0bca966b. Acesso em: 25 de maio 2021.

TAMIR, Mike; MILLER, Steven; GAGLIARDI, Alessandro. **The Data Engineer**. Ssrn Electronic Journal, [S.L.], 2015. Elsevier BV. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=2762013. Acesso em: 29 maio 2021.

WHITE, Sarah K.. **O que é e o que faz um engenheiro de dados?**.2018. Disponível em: https://cio.com.br/tendencias/o-que-e-e-o-que-faz-um-engenheiro-de-dados/. Acesso em: 24 de maio 2021.

1. Tradução do autor [↑](#footnote-ref-2)