**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO**

**Câmpus Campinas**

**Curso de Tecnologia em**

**Análise e Desenvolvimento de Sistemas**

**Avaliação de diferentes estratégias na engenharia de dados para a construção de um data warehouse com dados públicos voltado para cientistas sociais.**

**Autor: Deive Audieres Leal**

**Orientador: Prof. Dr. Samuel Botter Martins**

**CAMPINAS**

**2021**

Autor: Deive Audieres Leal

Orientador: Prof. Dr. Samuel Botter Martins

Avaliação de diferentes estratégias na engenharia de dados para a construção de um data warehouse com dados públicos voltado para cientistas sociais.

Plano Inicial de Trabalho de Conclusão de Curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de São Paulo, Câmpus Campinas, como requisito parcial para a aprovação na disciplina de Projeto de Sistemas I.

**CAMPINAS**

**2021**

**Sumário**

[1.](#_heading=h.1fob9te) Introdução 4

[2.](#_heading=h.3znysh7) Justificativa 5

[3.](#_heading=h.2et92p0) Objetivos 6

[4.](#_heading=h.tyjcwt) Fundamentação Teórica 7

[4.1.](#_heading=h.3dy6vkm) Engenharia de Dados 7

[4.2.](#_heading=h.1t3h5sf) ETL 7

[4.3.](#_heading=h.4d34og8) Data Warehouse 9

[5.](#_heading=h.17dp8vu) Metodologia 10

[6.](#_heading=h.3rdcrjn) Cronograma 11

[7.](#_heading=h.26in1rg) Bibliografia inicial 12

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.

1. **Justificativa**

A falta de diálogo entre diferentes áreas do conhecimento, se por um lado, possibilitam a focalização em temas, por outro, tende a diminuir o campo de visão das diferentes áreas. Se em um primeiro momento, a divisão entre humanas, biológicas e exatas, vem no bojo da divisão do trabalho e posteriormente à técnica, em um segundo criou obstáculos para a ciência observar ao todo.

Essa dificuldade de fazer uma análise mais holística dos problemas e levantar questões podem ter deixado a exploração mais difícil de ambientes tecnológicos por partes das humanas e levado as exatas por questões mais técnicas, mas que surtem pouco efeito no meio social.

Assim, esse trabalho é uma tentativa de aproximação entre as áreas. Agregando dados de diferentes fontes em uma só base através de técnicas de engenharia de dados que estão em voga, e, assim, podendo prová-las em ambientes controlados permitirá pensar abordagens para suportar de informações nos mais diferentes campos. Da mesma forma espera-se que a base de dados possa causar *insights* e boas perguntas aos cientistas sociais, trazendo uma compreensão significativa de aspectos sociais e abrindo a possibilidade de usar a tecnologia em diferentes contextos.

Assim, atacando por diferentes aspectos, espera-se que o presente trabalho possa ajudar a ciência a avançar de maneira interdisciplinar, ao mesmo tempo que permite pensar técnicas e analisar ferramentas que são de fundamental importância no mundo orientados a dados das organizações.

1. **Objetivos**
   1. **Objetivo Geral**

Este trabalho visa investigar a utilização de ferramentas *open source* para a realização das etapas do ETL e armazenamento em data warehouse. Como prova de conceito, avaliaremos tais soluções em bases de dados utilizadas por estudantes e pesquisadores da área de ciências sociais.

* 1. **Objetivos Específicos**

1. Implementar o ETL com diferentes abordagens.
2. Explorar qual ferramenta pode ser a mais adequada em diferentes cenários
3. Modelar e implementar um data warehouse com dados de economia e política facilitando a exploração por cientistas sociais.
4. **Fundamentação Teórica**

* 1. **Engenharia de Dados**

Entendido 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-0).

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.

* 1. **ETL**

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.

* 1. **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, à 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.

1. **Metodologia**

Primeiramente, optamos por utilizar o Pentaho – suite completa de *Business Intelligence* e com recursos para geração de relatórios, integração e armazenamento de dados, ETL, análise de informações, painéis para controle gerencial e mineração de dados – , que já é uma ferramenta bem conhecida no mercado e que vem sendo desenvolvida desde de 2004, contendo já muitos trabalhos que fundamentam sua utilização. Caminhando em direção a linguagem em si, utilizaremos o Python, linguagem de propósito geral criada nos anos 90, de scripts e orientada a objetos, para criar e automatizar o processo. Vale ressaltar que Python é hoje a segunda linguagem de programação mais utilizada no mundo e amplamente empregada na ciência de dados, graças a uma comunidade extremamente ativa que trabalha no desenvolvimento de suas bibliotecas, tais como: pandas, que contém uma série de ferramentas para análise e manipulação de dados e *data frames*; numpy, usada para cálculos em arrays multidimensionais; matplotlib, dedicada a visualização de dados; scikit-learn, utilizada para aplicações de aprendizado de máquina.

Com relação aos cenários em que os processos de ETL serão empregados, observamos que as ciências sociais têm problemas para encontrar dados abertos que facilitem a análise de situações sociais. Assim, iremos trabalhar em três eixos básicos, mas que possibilitam englobar diversos temas dentro das ciências humanas. Assim, buscamos desconstruir uma barreira entre humanas e exatas que dificulta, em muitos aspectos, a interdisciplinaridade entre as áreas de conhecimento.

Trabalharemos a princípio com os eixos de política e economia. Vale observar que os eixos estão conectados e são interdependentes em algumas análises. Refinando os processos e provando sua viabilidade, poderemos então abrir o escopo e, assim, possibilitar a expansão para mais áreas como saúde, educação, população e outras possíveis.

1. **Cronograma**

A delimitação do tema será feita no primeiro mês de trabalho. Posteriormente o levantamento de bibliografia e as primeiras explorações, seguido da versão inicial do plano de projeto. Nos meses posteriores a exploração da bibliografia será mais acentuada. A escrita do plano final decorrerá até o mês de julho, momento em que se iniciarão os primeiros experimentos com ETL e posteriormente, então, a modelagem do *data warehouse*.

De agosto a outubro, concomitante com os experimentos de ETL e *data warehouse*, serão realizadas as análises dos resultados das implementações nos cenários. Por fim, será realizada a redação do texto, revisão e finalização seguido da defesa.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Atividade | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Delimitação do tema |  |  |  |  |  |  |  |  |  |
| Leitura de bibliografia inicial |  |  |  |  |  |  |  |  |  |
| Escrita do plano inicial |  |  |  |  |  |  |  |  |  |
| Leitura de bibliografia geral |  |  |  |  |  |  |  |  |  |
| Escrita do plano final |  |  |  |  |  |  |  |  |  |
| Experimentos de ETL |  |  |  |  |  |  |  |  |  |
| Implementação de data warehouse |  |  |  |  |  |  |  |  |  |
| Análise dos resultados |  |  |  |  |  |  |  |  |  |
| Redação inicial |  |  |  |  |  |  |  |  |  |
| Revisão |  |  |  |  |  |  |  |  |  |
| Redação final |  |  |  |  |  |  |  |  |  |
| Defesa |  |  |  |  |  |  |  |  |  |

1. **Bibliografia inicial**

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-0)