Data warehouse e ETL para dados da saúde pública: Uma aplicação prática

Aluno: Marcos Vinicius do Carmo Sousa **Orientadora:** Profa. Dra. Kelly Rosa Braghetto

Departamento de Ciência da Computação (DCC) – Instituto de Matemática e Estatística (IME) – Universidade de São Paulo (USP)
Contato: marcos.vinicius.wz@gmail.com

Introdução

- O Sistema Único de Saúde (SUS) tem centenas de Sistemas de Informação de Saúde (SIS), usados nas esferas federal, estadual e municipal
- Os dados dos usuários do SUS estão fragmentados e duplicados nas bases desses sistemas
- A integração dos dados auxilia o planejamento de ações e a visualização da situação da saúde pública
- A Secretária Municipal de Saúde de São Paulo (SMS-SP) quer integrar dados de três SIS: nascidos vivos (SINASC), mortalidade (SIM) e internações (SIH)

Objetivos

Este trabalho tem como objetivos:

- Propor um esquema de dados de *data warehouse* (baseado no modelo de dados multidimensional) para integrar dados provenientes de SIS-SUS da SMS-SP
- Desenvolver um processo de extração-transformação-carga (ou ETL, de *extract-transform-load*) para manter o *data warehouse* atualizado, ao mesmo tempo em que garante a conformidade e integridade dos dados unificados

Data warehouse

W. H. Inmon (1992) caracterizou um *data warehouse* como uma coleção de dados que varia no tempo, integrada, não volátil e orientada a assunto, que dá apoio às decisões de gerência.
O *data warehouse* neste trabalho é composto por quatro bancos de dados (BDs), produzidos por etapas diferentes do processo de ETL:

- o de Extração, que recebe os dados crus das bases de dados fontes
- o de Limpeza, em que os dados passam por procedimentos que removem campos vazios, erros de formatação, etc.
- o de Conformidade, em que as duplicações são removidas dos dados já limpos
- o Final, que possui o esquema principal do *data warehouse* – é o BD multidimensional que recebe os dados tratados e conformados das etapas anteriores e sobre o qual o analista ou aplicação poderá realizar suas consultas

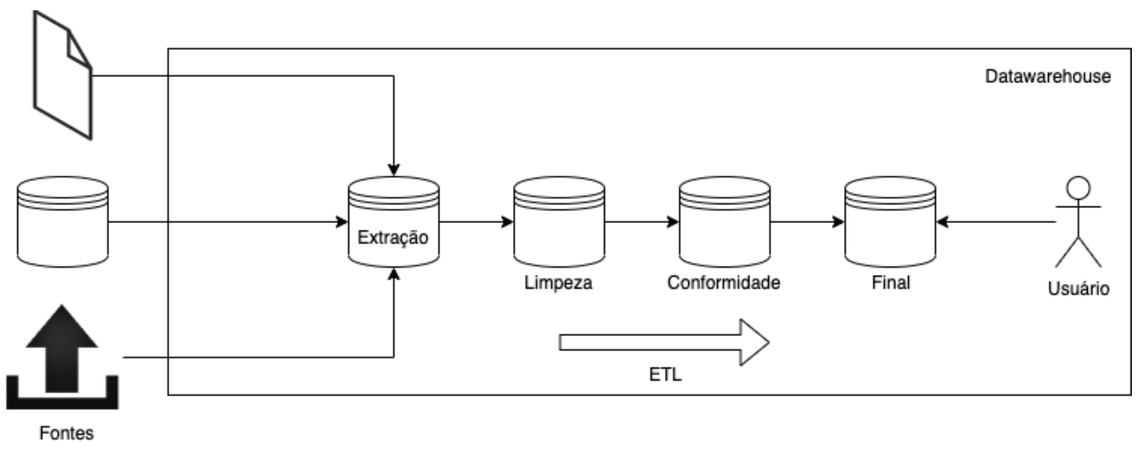

Figure 1:Etapas do processo de ETL para data warehouse desenvolvido

Modelo multidimensional

Neste trabalho, foi desenvolvido um esquema de *data warehouse* baseado no modelo de dados multidimensional, a partir dos esquemas dos BDs originais dos SIS-SUS da SMS-SP. O principal benefício desse esquema é ser altamente normalizado e ter os dados orientados a fatos e dimensões.

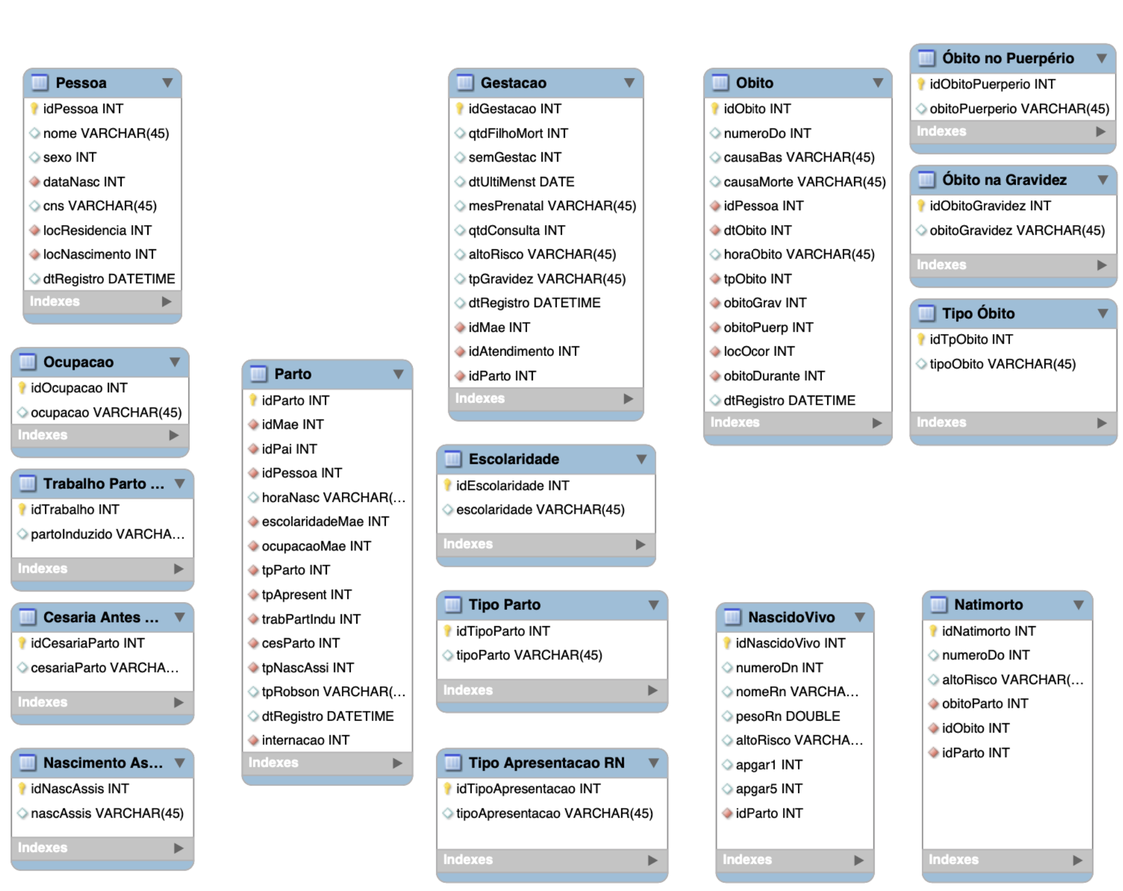

Figure 2:Esquema multidimensional (parcial) proposto para a integração de dados de SIS-SUS

Ferramentas de ETL

Neste trabalho, foram avaliadas diferentes ferramentas que auxiliam a criação e manutenção de processos de ETL. Elas foram selecionadas pela sua ampla aceitação na indústria e de acordo com a licença do seu código fonte.

Ferramenta	Tipo licença	Empresa/Organização	Escrito em	Vantagens	Desvantagens
Kettle	código proprietário	Pentaho	Java	Criar sua própria dashboard para ETL	Maioria dos recursos são da versão empresarial
				Análise de dados de logs do processo ETL	Suporte é somente na versão empresarial
				Suporte dedicado	Não é muito utilizado pela comunidade open source
				Suporte pré-definido de conectores de diversas fontes	Documentação de difícil acesso
Luigi	código aberto	Spotify	Python	Definição de fluxos de dados utilizando Python scripts	Não tem suporte nativo para processos distribuídos
				Código modular, tornando mais fácil de manter e atualizar fluxos	Interface gráfica não é intuitiva, tornando mais difícil de navegar.
				Integração com linha de comando	Escalabilidade é muito limitada
Airflow	código aberto	Airbnb	Python	Definição de fluxos de dados utilizando Python scripts	Airflow não tem suporte para data streaming
				Utilização de grafos acíclicos (DAG) como padrão de definição de trabalho	
				Possui agendamento de tarefas mais completo	
				Interface web muito completa	
				Altamente escalável	
				Ferramenta é muito utilizado pela comunidade open source	

Figure 3:Tabela comparativa ferramentas ETL

Principais Resultados

- BD multidimensional da figura 2 criado em um servidor PostgreSQL
- Processo de ETL implementado usando Python e Airflow
- Automatização da detecção da chegada de novos dados das fontes, do processamento dos dados e da atualização do *data warehouse*
- Testes com quase 2.5 milhões de registros do SINASC, SIM e SIH fornecidos pela SMS-SP em arquivos CSV

Referências

KIMBALL, R., e CASERTA, J. The data warehouse ETL toolkit: Practical techniques for extracting, cleaning, conforming, and delivering data. pg 3–25.
NAVATHE, S. B., e ELMASRI, R. Fundamentals of database systems. Overview of Data Warehousing and OLAP, 6 ed (2016), pg 1067–1069.