

Proposta TCC
Um Modelo de Banco de Dados Analítico para
Dados de Saúde Pública

Aluno

André Akira Hayashi
NUSP: 9293011

Orientador

Paulo Meirelles

Co-orientadora

Renata Wasserman

20 de Abril 2019

1 Introdução

As tecnologias de Cidades Inteligentes emergem como uma solução para enfrentar problemas comuns em grandes centros urbanos, usando os recursos da cidade de forma eficiente e proporcionando serviços de qualidade para os cidadãos. Neste sentido, o IME-USP sedia o INCT da Internet do Futuro para Cidades Inteligentes - InterSCity [1], um consórcio de pesquisa multidisciplinar que tem como objetivo abordar os principais desafios de pesquisa em cidades inteligentes, com financiamento da FAPESP, CNPq e CAPES. Um dos objetivos de nosso INCT, é investigar abordagens e desenvolver aplicações para visualização espacial de dados de saúde de grandes cidades como o município de São Paulo, a partir de dados oficiais fornecidos pela Secretaria Municipal de Saúde de São Paulo (SMS-SP). A grande quantidade de dados heterogêneos sobre saúde nas grandes cidades torna necessário a criação de novas formas dinâmicas de visualização georreferenciada de dados da área da saúde. Nesse sentido, há diversas oportunidades de colaborações inerentes ao desenvolvimento dessas soluções, tanto em pesquisa quanto em desenvolvimento técnico. Portanto, este projeto visa oferecer oportunidades para que pesquisadores e gestores públicos possam contribuir com formas inovadoras para coleta, armazenamento, gestão, visualização e análise de grandes quantidades de dados de saúde de populações urbanas. Uma primeira versão desta plataforma já está disponível [2].

Neste trabalho temos como principal tópico a integração de dados da SMS-SP, principalmente das bases SINASC (Sistema de Informações sobre Nascidos Vivos), SIH (Sistema de Informações Hospitalares) e a SIM (Sistema de Informações de Mortalidade) em um modelo de banco de dados analítico para os dados das bases mencionadas acima, para então fazer a evolução da camada de visualização dos fatos, como:

- Os pontos de localização de onde a gestante reside e onde foi realizado o parto e obterá o deslocamento da mãe para a realização da internação para o parto;
- Viés de tempo de gestação relacionado à região que a gestante reside;
- Deslocamento da gestante para ser atendida nos serviços de saúde;

2 Objetivos

Este projeto ao seu fim terá como principais objetivos:

- Modelar um Data Warehouse para os dados dos sistemas da SMS.
- Instalar e configurar o SGBD para o Data Warehouse, e depois criar seu modelo físico do Data Warehouse no SGBD.
- Criar um banco de dados analítico a partir do Data Warehouse.

- A partir dos DataMarts criados a partir dos Data Warehouse, criar cubos OLAP, para agilizar a execução de consultas dinâmicas e a obtenção de informações sumariadas.
- Com os cubos criados, fazer o estudo de caso, conforme sugeridos pela secretaria.

3 Metodologia

Os bancos de dados transacionais geralmente se baseiam no modelo de dados relacional. Já os Data Warehouses se baseiam no modelo de dados (multi)dimensional, em que esse modelo de dados representa os indicadores importantes para uma área de negócio, que são chamados de medidas ou fatos, e os seu parâmetros, chamados de dimensões. Sendo que os fatos são as variáveis observadas ou medidas de interesse e as dimensões são diferentes perspectivas sob as quais os fatos podem ser vistos. Neste trabalho, trabalharemos com os fatos em que a SMS-SP possui maior interesse para cada uma das bases (SINASC, SIH e SIM). Como os modelos multidimensionais tiram proveito das relações inerentes aos dados para gerar dados em matrizes multidimensionais chamadas de cubos (se $n > 3$, ele é chamado de hipercubo) de dados, sendo que nessa estrutura os dados podem ser consultados diretamente em qualquer combinação das dimensões, formando assim diversas consultas com um único cubo, sendo que, existem ferramentas que permitem a visualização e a análise dos dados de acordo com a escolha das dimensões necessárias para a análise.

4 Cronograma

No início do projeto as tarefas serão feitas em conjunto com um outro aluno (de abril a junho de 2019):

1. Estudar as estruturas e o significado dos dados dos sistemas da SMS-SP, o modelo de dados multidimensionais, os Data Warehouses e as operações OLAP (Online Analytical Processing), as ferramentas para ETL (Extract, Transform and Load) e para OLAP e quais são os SGBD's geralmente usados para manter os Data Warehouses;
2. Modelar o Data Warehouse;
3. Instalação e configuração de um SGBD para o Data Warehouse;
4. Criação do modelo físico do Data Warehouse no SGBD.

Após a criação do Data Warehouse, cada aluno seguirá para seu respectivo tema, sendo o deste TCC a criação de um modelo de banco de dados analítico para dados de saúde pública (de julho a setembro de 2019)

1. Realizar um estudo mais aprofundado do OLAP ;

2. Estudar e escolher uma ferramenta de apoio para a análise dos dados como por exemplo: Pentaho CDE, Mondrian, Reporting, etc;
3. Criação do bancos de dados analíticos, os Data Marts a partir do Data Warehouse;
4. A partir dos Data Marts, criar cubos OLAP, para agilizar a execução de consultas dinâmicas e a obtenção de informações sumariadas;
5. Estudo de caso: implementar consultas/visualizações de dados referentes aos casos propostos pela SMS-SP.

5 Requisitos

Este projeto consta com alguns requisitos que devem ser respeitados, como:

- Utilizar um software livre;
- Ter suporte a dados georeferenciados;
- Possuir um bom desempenho em consultas de dados dinâmicos;
- Capacidade para armazenar grande volume de dados, no mínimo 1 milhão de registros;
- Priorizar soluções que dependam de uma infraestrutura computacional simples, acessível à SMS-SP.

Referências

- [1] <http://intercity.org>
- [2] <http://healthdashboard.intercity.org>
- [3] Sistemas de Bancos de Dados (6ª edição), *Elmasri e Navathe*, Capítulo 29
- [4] Sistemas de Gerenciamento de Banco de Dados (3ª edição), *Ramkrishnan e Gehrke*, Capítulo 25