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

André Akira Hayashi

Monografia apresentada ao
Instituto de Matemática e Estatística
da Universidade de São Paulo
para obtenção do título de
Bacharel em Ciência da Computação

Orientador: Prof. Dr. Paulo Meirelles
Coorientadora: Profa. Dra. Renata Wasserman e Profa. Dra. Kelly Rosa Braghetto

São Paulo
Janeiro de 2020

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

André Akira Hayashi

Esta é a versão original da
monografia entregue como parte
do trabalho final de MAC0499.

Agradecimentos

Sou grato a um grupo de pessoas que me apoiaram ao longo não somente deste ano, mas por toda a minha vida. Agradeço primeiramente aos meus pais *Cecília Harumi Yamaguchi* e *Hélio Tsunefumi Hayashi*, que me amaram incondicionalmente, apoiaram todas as minhas decisões que fiz até hoje e se sacrificaram para formar a pessoa que eu sou hoje. Aos meus orientadores *Paulo Meirelles*, *Renata Wasserman* e *Kelly Rosa Braghetto* por me auxiliarem no desenvolvimento deste trabalho e estarem sempre presentes quando possível, para sanarem as minhas dúvidas. Ao aluno *Marcos Vinicius*, que me auxiliou no início deste trabalho com a modelagem do Data Warehouse e me ajudou na escolha da ferramenta utilizada, além de esclarecerem minhas dúvidas em relação ao projeto. A todos os professores do IME que fizeram parte da minha graduação e compartilharam seus conhecimentos valiosos que me acompanharão para o resto da minha vida. E finalmente, agradeço a todos os meus amigos que estão presentes na minha vida, e fizeram todo este processo muito mais fácil.

Resumo

André Akira Hayashi. **Um Modelo de Banco de Dados Analítico para Dados de Saúde Pública:** . Monografia (Bacharelado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2019.

Big data são dados com grande variedade que chegam em volumes crescentes e são atualizados com uma alta frequência. Essa definição se encaixa perfeitamente nos dados do setor da saúde, que possui um histórico em guardar dados, seja para a manuntenção de registros ou para estudos futuros. Sua variedade de tipos de dados é grande também, já que existem diversas sub-áreas da saúde, e a alta velocidade em que esses dados devem ser processados e analisados para sempre manter atualizado tanto os diagnósticos quanto os seus tratamentos. Para poder analisar diversas bases ao mesmo tempo, neste trabalho é proposta uma prova de conceito de uma modelagem de um Data Warehouse, para as bases SIM (Sistema de Informações sobre Mortalidade), SINASC (Sistema de Informação sobre Nascidos Vivos) e SIH (Sistema de Informações sobre Internações Hospitalares) do DATASUS do município de São Paulo, para que seja possível analisar esses dados apoiado por visualizações de dados. Nesse contexto, somado ao desafio de unificar as bases de dados, a irregularidade em que os dados estão disponibilizados dificultou a população do Data Warehouse proposto. Assim, foi necessário um grande esforço na limpeza dod dados para que fosse possível ter um Data Warehouse com os dados e pronto para a análise. As análises e visualizações prototipadas neste trabalho conseguiram responder algumas questões, como por exemplo, se há alguma relação entre a escolaridade da mãe e o tipo de parto realizado. Mostramos que a integração das bases é possível, sendo necessário um trabalho de tratamento dos dados para que eles fiquem uniformes para então serem carregados no Data Warehouse. As análises via o Data Warehouse prototipado, com o auxílio dos cubos pré-calculados desenvolvidos neste trabalho, resultaram em um conjunto de visualizações apontam em como é possível agilizar as pesquisas e deixá-las mais compreensíveis para os técnicos da área da saúde.

Palavras-chave: Big data, Data Warehouse, Análise de dados, Visualização de dados, Saúde pública, Superset Apache

Lista de Abreviaturas

SMS-SP	Secretaria Municipal de Saúde de São Paulo
SIH	Sistema de Informações sobre Internações Hospitalares
SIM	Sistema de Informações sobre Mortalidade
SINASC	Sistema de Informação sobre Nascidos Vivos
SQL	Linguagem de Consulta Estruturada (*Structured Query Language*)
CSV	*Comma-separated values*
IME	Instituto de Matemática e Estatística
USP	Universidade de São Paulo
BI	Business Intelligence

Lista de Figuras

Sumário

Capítulo 1

Introdução

O setor de saúde, históricamente, gera grandes volumes de dados, incentivados pela manuntenção dos registros, requerimentos constantes de documentos e do cuidado ao paciente [12]. Até pouco tempo atrás, os dados eram armazenados na forma física como em papel, que traz diversos problemas como o espaço necessário para guardá-los e a deterioração rápida com o passar do tempo. Mas, atualmente, a tendência é a da digitalização desses documentos.

Isso está no contexto do *Big data*, que são os conjuntos de dados tão grandes e complexos que são quase impossíveis de se gerenciar com software ou hardware tradicionais [7]. *Big data* também é utilizada de forma a ajudar uma grande quantidade de funções e de assistências médicas, e futuramente auxiliar à decisão clínica, como por exemplo, auxiliar o médico ou enfermeiro no diagnóstico de um paciente, apenas com os sintomas observados pelo responsável. Além disso, também será possível auxiliar na observação de doenças com potencial de se tornarem epidêmicas e na gestão da saúde da população [6]. Somente os dados do setor de saúde dos EUA chegaram à 150 exabytes em 2011 se esse ritmo for mantido, o *big data* no setor de saúde dos EUA chegará rapidamente à escala dos zettabytes e até a yottabytes [22].

O *Big data* no setor de saúde não é assustador somente pelo seu tamanho, mas também pela sua diversidade de tipos de dados, já que suas áreas vão desde o financeiro de um posto de saúde até ao cuidado com os pacientes, além da velocidade em que esses dados devem ser processados e analisados. Esses dados no Brasil são coletados pelo DATASUS, e variam desde arquivos com a distribuição da população brasileira segundo censos demográficos até dados mais comuns da área da saúde como arquivos disseminâveis para tabulação do Sistema de informação de Nascidos Vivos.

Para a área de pesquisa de *Big data*, o setor da saúde oferece uma grande quantidade e variedade de dados, para analisar e descobrir padrões e tendências nos dados, podendo trazer benefícios, como, uma melhora no atendimento ao paciente, auxílio em diagnósticos médicos e na redução de custos, como por exemplo, treinar os membros da família dos pacientes para prestar cuidados pós-operatórios [21]. As ferramentas que trabalham com a análise de dados em *Big data* do setor de saúde, aproveitam do momento em que ocorre uma grande atualização dos dados estudados, para melhorar ou afirmar as análises feitas

anteriormente. Após a limpeza e análise dos dados do *Big data* ser feita, os especialistas da área utilizam-o para melhorar os diagnósticos e nos desenvolvimentos de novos tratamentos, que podem resultar em novas curas para doenças ou até na diminuição do custo de algum tratamento.

Este trabalho apresenta um protótipo de modelagem para um Data Warehouse, para unificar as bases SIM (Sistema de Informações sobre Mortalidade), SIH (Sistema de Informações sobre Internações Hospitalares) e SINASC (Sistema de Informação sobre Nascidos Vivos) pertencentes ao SUS, do município de São Paulo. Para então, realizar as análises dos dados utilizando essa mesma modelagem, mostrando que é possível melhorar o modo em que esses dados são organizados e realizar uma análise mais rápida e fácil de se entender. Essas análises foram feitas utilizando o **Superset Apache**, que é uma aplicação de Business Intelligence com uma interface simples, com gráficos fáceis de se interpretar e que permite a criação e o compartilhamento de dashboards, que podem unificar vários gráficos em um mesmo local, para que seja mais fáci lde se analisar os dados em diferentes dimensões. Atraves dele foi possível montar os gráficos e realizar as análises dos dados no modelo do Data Warehouse proposto para a integração das bases. Alguns exemplos disso são: relacionar a escolaridade da mãe com o tipo de parto realizado, relacionar doenças com as estações do ano, entre outros. Após realizar as análises de algumas questões propostas pela prefeitura, foi possível retirar as dúvidas sobre elas. Sendo que com o auxílio do Data Warehouse houve uma grande simplificação na organização dos dados, se compararmos em como eles foram entregados para nós, facilitando as análises dos dados.

Capítulo 2

Fundamentação teórica

Este capítulo tratará de uma visão geral da parte teórica sobre big data e data warehouse que serão utilizados neste trabalho. Além de explicar a escolha da ferramenta que foi utilizada para realizar as análises dos dados sobre o data warehouse modelado.

2.1 Big data

Big data são dados com grande variedade que chegam em volumes crescentes e com velocidade cada vez maior [3], isso passou a ser conhecido como os três Vs. Simplificando, big data é um conjunto de dados grande e diversificado, que o software tradicional de processamento de dados simplesmente não consegue gerenciá-los. Os Três V's do Big Data são:

Volume é a quantidade de dados que importam para a análise que está sendo realizada. Com o big data, você terá que processar grandes volumes de dados não estruturados que podem facilmente chegar a centenas de petabytes.

Velocidade é a taxa na qual os dados são recebidos e administrados. Normalmente, a velocidade se torna mais alta quando os dados são transmitidos diretamente para a memória, em vez de ser gravada no disco. Alguns produtos para a internet operam em tempo real ou quase e necessitam de uma avaliação e ação em tempo real.

Variedade refere-se à diversidade dos tipos de dados disponíveis. Com o surgimento do big data, os dados vêm em novos tipos de dados não estruturados, como texto, áudio e vídeo exigem um pré-processamento diferenciado para, descobrir seu significado e dar suporte a metadados.

Neste trabalho estamos utilizando as base de dados da Secretaria Municipal da Saúde da cidade de São Paulo (SMS-SP), que geram grandes quantidades de dados devido às diversas áreas existente na área de saúde como, por exemplo, internações, maternidade, óbitos, partos, etc. Esses dados são gerados com uma frequência muito alta, já que são atendidos milhares de pessoas diariamente pelo SUS. As características anteriores classificam essas bases da SMS-SP como um big data, já que elas englobam os três V's, volume, velocidade e variedade.

2.2 Data Warehouse

Um Data Warehouse é um conjunto de dados que são unidos de acordo com um assunto em comum, integrado, não volátil e varia de acordo com o tempo, que é definida de acordo com o público ou cliente [9], porém essa definição restringe o Data Warehouse como uma "conjunto de dados", então definições mais recentes os descrevem de forma mais abrangente, como: “um sistema projetado com o propósito de dar apoio à extração, processamento e apresentação eficiente (dos dados) para fins analíticos e de tomada de decisão”.

Os Data Warehouses possuem as seguintes características: integram grandes quantidades de dados provenientes de diversas fontes, otimizados para a recuperação de dados, são mais preocupados com o armazenamento, a manutenção e a recuperação eficiente de dados históricos, a informação muda com pouca frequência, ou seja, não-volátil, portanto as suas atualizações são normalmente incrementais e possuem diferentes tipos de aplicações de análise, como por exemplo, o OLAP (Online Analytical Processing).

OLAP é um conceito de interface com o usuário que torna mais fácil a formulação de idéias e/ou perguntas sobre os dados, fazendo com que seja possivel analisá-los sob diversos ângulos. E normalmente utilizam-se de uma classe de consultas estilizadas, dentre elas temos: operadores de agrupamento e agregação, suporte para condições booleanas complexas, funções estatísticas e recursos para a análise de séries temporais.

Data Warehouses são baseasdos em um modelo de dados multidimensional nele os indicadores importantes são chamados de medidas ou fatos, e seus parâmetros são chamados de dimensões. Os modelos multidimensionais utilizam-se dessas relações com os dados para gerar matrizes multidimensionais chamadas de cubos de dados o desempenho de consultas realizados neles pode ser bem melhore do que se forem feitos em modelos de dados relacionais. Nessa estrutura de cubos, os dados podem ser consultados diretamente combinando qualquer uma de suas dimensões, fazendo com que seja evitado consultas complexas que seriam realizadas ao banco de dados. Considerando essa ideia dos cubos, hoje em dia existem ferramentas que realizam a vizualização dos dados de acordo com as dimensões escolhidas.

No modelo de armazenamento multidimensional, existem dois tipos de tabelas:

- **Tabela de dimensão:** são usadas para descrever as dimensões, eles contêm chaves, valores e atributos da dimensão;
- **Tabela de fatos:** nela estão contidas algumas variáveis medidas ou observadas que são identificadas por ponteiros (equivalente a uma chave estrangeira) para as tabelas de dimensões.

Nos esquemas de armazenamento, dois deles são os mais utilizados:

- **Esquema estrela:** Cada dimensão em um esquema estrela é representado com apenas uma tabela de dimensão única e essa tabela de dimensão contém um conjunto de atributos;
- **Esquema floco de neve:** Algumas tabelas de dimensões são padronizadas e isso divide os dados em tabelas adicionais.

Com as ferramentas OLAP são oferecidos um conjunto de operações para a agregação, seleção e projeção dos dados que estão organizados em um modelo multidimensional. Além da operação mais comum que é agregar uma medida sobre uma ou mais dimensões, temos também as operações de:

- **roll-up:** é um resumo em diferentes níveis de uma hierarquia de dimensões;
- **drill-down:** fornece uma visão de granularidade mais fina, desagregando elementos;
- **rotação (pivoting):** o cubo pode ser "girado"para exibir uma orientação diferente dos eixos;
- **fatiar (slice):** fazer uma seleção por igualdade em uma ou mais dimensões, possivelmente com algumas dimensões removidas;
- **cortar (dice):** fazer uma seleção por intervalo.

Considerando o grande volume de dados que é gerado pela SMS-SP, toda a variedade proveniente às suas bases de várias áreas da saúde e a grande frequência de atualizações, de acordo com essas características, a modelagem de um data warehouse foi proposto. Para que seja possível integrar todas essas diversas áreas em um único modelo e relacionar cada dado inserido com a data em que ele foi carregado no sistema para que haja um histórico desses dados, o que irá resultar em uma otimização na velocidade em que as pesquisas serão realizadas sobre esses dados. Considerando que para este trabalho as bases utilizadas foram as de mortalidade, internações e nascimentos, sendo que cada uma delas gerará pelo menos uma tabela fato neste modelo.

2.3 Data analysis

Data analysis é o processo de limpeza, transformação e modelagem dos dados para descobrir informações úteis para a área de estudo. Data analysis tem diversas peculiaridades e abordagens, abrangendo diversas técnicas que são usadas em diferentes situações. Nos dias de hoje, a data analysis desempenha um papel na tomada de decisões na área científica e no aumento da eficiência na operação de empresas [23].

O processo de data analysis é a coleta de informações utilizando um aplicativo ou uma ferramenta que permite a exploração dos dados e encontrar padrões neles. Considerando isso, pode-se tomar decisões ou obter conclusões definitivas. Esse processo consiste nas seguintes fases: Primeiro de tudo, na **Coleta de requisitos de dados**, que consiste em decidir o que analisar e como medi-lo precisa entender por que está investigando e quais medidas deve usar para fazer essa análise. Em seguida há a **Coleção de dados**, em que os dados são reunidos levando em consideração os requisitos discutidos na fase anterior. Além disso é necessário que haja registros para a data em que a coleta foi feita e a origem dos dados. Após a coleta é necessário a **Limpeza de dados**, pois nem todos os dados coletados serão úteis para a análise ou eles podem ter sido preenchidos de forma errada, portanto eles devem ser limpos. Por exemplo, registros duplicados, espaços em branco, datas com formato errado e etc, devem ser limpos nessa etapa. Após estas três etapas serem realizadas, os dados estão prontos para serem analisados durante esta etapa, é possível utilizar ferramentas e software de análise de dados que auxiliarão a entender, interpretar

e tirar conclusões com base nos requisitos. Depois de analisar os dados, deve-se fazer a **interpretação dos dados** ou dos seus resultados ela pode ser feita oralmente, escrita ou graficamente, por meio de tabelas e gráficos. E por último a **Visualização de dados** que é muito comum nos dias de hoje; geralmente são apresentados na forma de tabelas e gráficos. Em outras palavras, os dados são mostrados graficamente para facilitar o entendimento e o processamento do público alvo ou cliente. A visualização de dados geralmente é usada para descobrir fatos ou tendências desconhecidas ao observar isto e comparando conjuntos de dados, pode-se encontrar informações significativas que passaram despercebidas, pelos meios comuns de análise.

2.3.1 Ferramentas

Para a análise de dados comparamos as ferramentas Pentaho[1] e Superset[2]. A decisão inicial pelo Pentaho foi devido a sua integração e análise de dados de um Data Warehouse. Isso permitiria que houvesse uma organização no acesso, preparo e análise de todos os dados de qualquer fonte em qualquer ambiente. Entretanto, a versão livre do Pentaho tinha algumas limitações, principalmente na área de análise de dados e gráficos, sendo que esses problemas são resolvidos na versão paga do Pentaho.

Considerando isso decidimos aderir à ferramenta Apache Superset que é software livre e uma aplicação de Business Intelligence para a web. Atualmente, o Superset está sendo usado pelo Airbnb, Twitter, GfK Data Lab, Yahoo!, Udemy e outros. Segundo a página do GitHub[3] o Superset já foi testado em grandes ambientes com centenas de utilizadores. O ambiente de produção do Airbnb serve mais de 600 utilizadores ativos que visualizam mais de 100 mil gráficos por dia.

O Superset tem como principais características um conjunto rico de visualizações de dados (Figura 2.1), criação intuitiva de uma dashboards com os gráficos montados (Figura 2.2), um modelo de segurança/permissão extensível e de alta granularidade permitindo regras complexas sobre quem pode acessar certos recursos e datasets, a integração com a maior parte dos sistemas gerenciadores de banco de dados relacionais com linguagem SQL através do SQLAlchemy (Figura 2.3), entre eles temos: MySQL, Oracle, MySQL, PostgreSQL, Snowflake, SQLite, SQL Server, entre outros e sua integração com o SQL Lab (Figura 2.4), que permite selecionar a base de dados, o schema e a tabela, que já foram previamente carregadas no superset, para então realizar querying interativas, para visualizar os dados filtrados pela pesquisa em SQL, além disso as queries são guardadas em um histórico[4].

Considerando tudo isso, e principalmente por ser um software livre e sua utilização por grandes empresas que possuem grandes cargas de dados, conclui-se que essa seria uma escolha mais adequada para a análise de dados do projeto.

[1] https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho-platform.html
[2] https://superset.incubator.apache.org/
[3] https://github.com/apache/incubator-superset
[4] https://superset.incubator.apache.org/sqllab.html

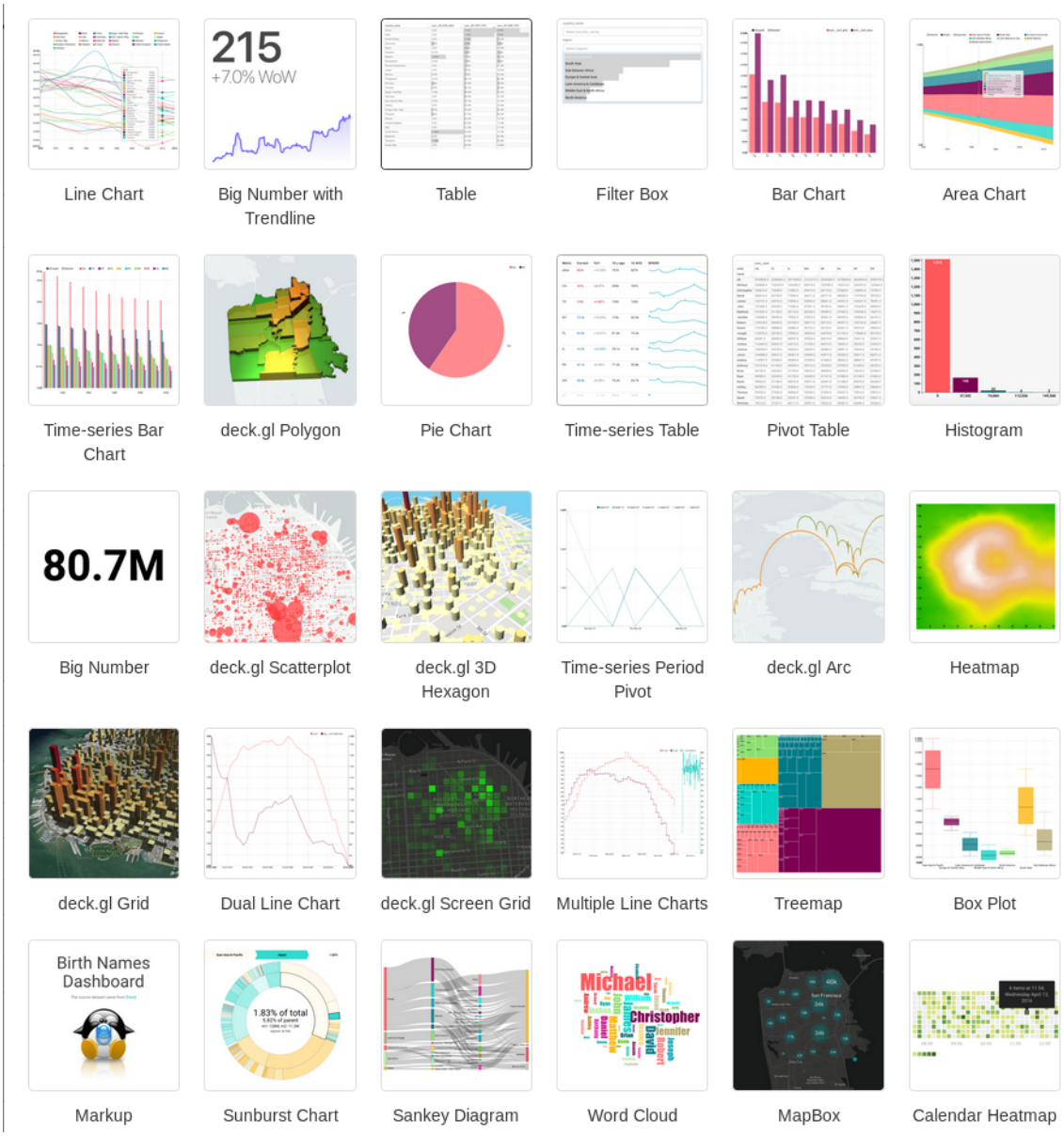


Figura 2.1: *Exemplos de visualizações de dados no Superset.*

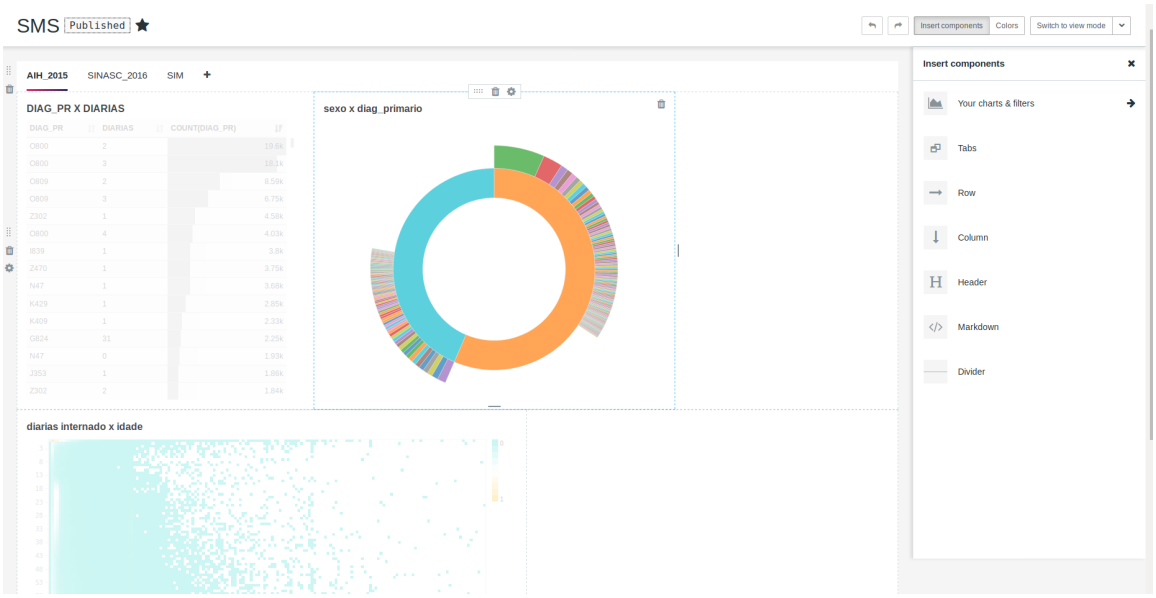


Figura 2.2: *Edição da dashboard no Superset.*

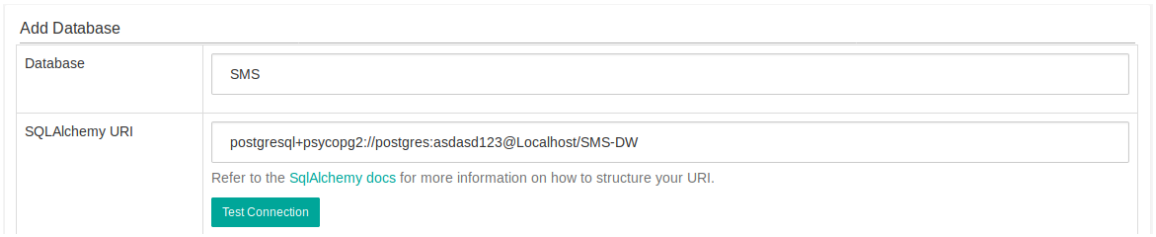


Figura 2.3: *Conexão com uma base de dados através do SQLAlchemy no Superset.*

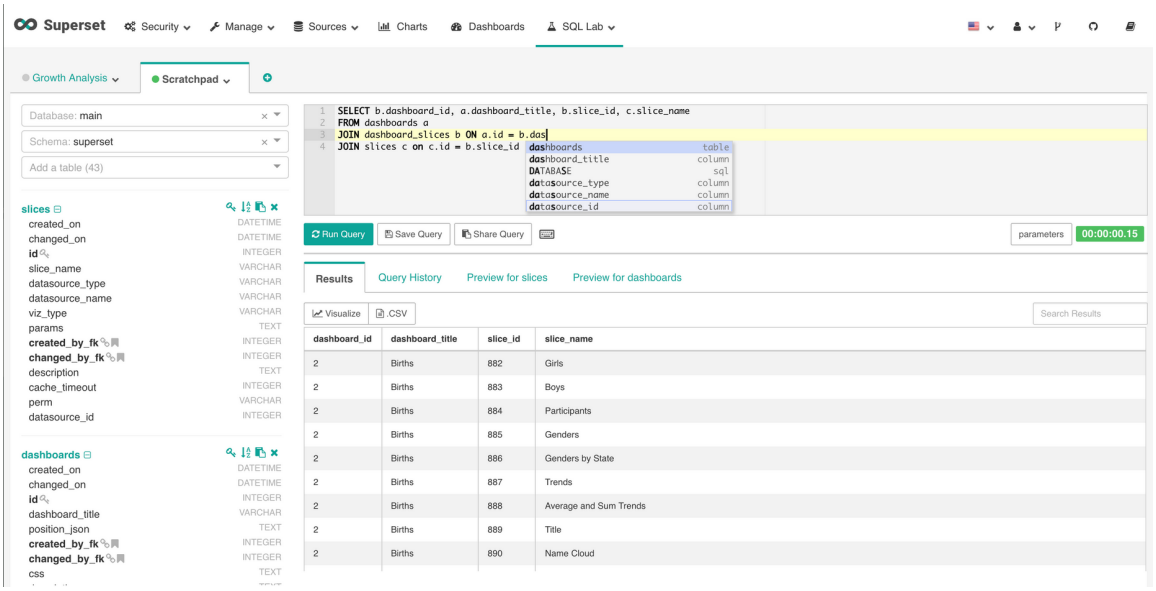


Figura 2.4: *Querie no SQL Lab do Superset.*

Como o Superset é utilizado por grandes empresas com grandes quantidade de dados como foi citado anteriormente, pode-se concluir que ela é uma boa ferramenta em tratar com grandes quantidades de dados como é um Big Data.

Para este trabalho está sendo utilizado principalmente a diversidade de gráficos disponível, a fácil conexão com o PostgreSQL através do SQLAlchemy, que será necessário já que o Superset não possui nenhum tipo de ferramenta específica para trabalhar com Data Warehouse, e com o PostgreSQL isso pode ser auxiliado através das cláusulas GROUP BY CUBE ou CREATE MATERIALIZED VIEW, ambos podem atuar como os cubos pré-calculados e ajudar a agilizar as consultas.

Capítulo 3

Exemplo de integração de bases de dados de saúde

O Departamento de Informática do Sistema Único de Saúde (DATASUS) foi criado em 1991 junto com a criação da Fundação Nacional de Saúde (Funasa). O DATASUS, tem a responsabilidade de prover aos órgãos do SUS de sistemas de informação o suporte de informática, necessário ao processo de planejamento, operação e controle.

Nos 25 anos de atuação do DATASUS já foi desenvolvido mais de 200 sistemas que auxiliam diretamente o Ministério da Saúde no processo de construção e fortalecimento do SUS. Atualmente, o departamento trabalha em sua maioria em soluções de software para a área da saúde, adaptando seus sistemas de acordo com às necessidades e integrando novas tecnologias.

A estrutura de armazenamento de dados do Departamento tem a capacidade de armazenar informações sobre saúde de toda população brasileira. Além disso, também são disponibilizados links espalhados em várias cidades brasileiras com conexões com todos os Núcleos Estaduais do Ministério da Saúde, Funasa, Agência Nacional de Vigilância Sanitária (ANVISA), Casa do Índio e com as 27 secretarias estaduais de saúde.

Em resumo, o DATASUS tem como missão promover a modernização por meio da tecnologia da informação para apoiar o Sistema Único de Saúde - SUS. Relacionado com o Ministério da Saúde, via o DATASUS foram criados, ao longo dos 27 anos de sua existência, mais de 200 sistemas ligados a gestão de saúde [11, p. 1]. A maior parte dos sistemas está relacionada à notificação de eventos do cuidado, tais como nascimento (SINASC) e mortalidade (SIM). Há também sistemas voltados para a gestão do gasto público, como o sistema para notificação de internações hospitalares (SIH). SINASC, SIM e SIH são os sistemas que obtivemos os dados para os exemplos usados neste trabalho:

SINASC: "O SINASC ou Sistema de Informações sobre Nascidos Vivos foi desenvolvido pelo DATASUS para reunir informações epidemiológicas referentes aos nascimentos informados em todo território nacional. Sua implantação ocorreu de forma lenta e gradual em todas as Unidades da Federação"[1].

[1] http://www2.datasus.gov.br/DATASUS/index.php?area=060702

SIM: "O SIM ou Sistema de Informações sobre Mortalidade foi criado pelo DATASUS para a obtenção regular de dados sobre mortalidade no país. A partir da criação do SIM foi possível a captação de dados sobre mortalidade, de forma abrangente, para subsidiar as diversas esferas de gestão na saúde pública. Com base nessas informações é possível realizar análises de situação, planejamento e avaliação das ações"[2].

SIH: "A finalidade do SIHSUS ou Sistema de Informações Hospitalares do SUS é de registrar todos os atendimentos provenientes de internações hospitalares que foram financiadas pelo SUS, e a partir deste processamento, gerar relatórios para que os gestores possam fazer os pagamentos dos estabelecimentos de saúde. Além disso, o nível Federal recebe mensalmente uma base de dados de todas as internações autorizadas (aprovadas ou não para pagamento) para que possam ser repassados às Secretarias de Saúde os valores de Produção de Média e Alta complexidade, além dos valores de CNRAC, FAEC e de Hospitais Universitários – em suas variadas formas de contrato de gestão"[3].

Muitos desses sistemas têm seus dados disponibilizados no site do DATASUS[4]. Os dados podem ser adquiridos em formato CSV a partir de uma interface online de filtragem, o TABNET[5]. Especificamente na cidade de São Paulo, que é o objeto de estudo deste trabalho, a Secretaria Municipal da Saúde de São Paulo é quem coordena o SUS da cidade e promove ações e projetos no intuito de proteger e gerar a saúde da população. As bases da SMS-SP foram criadas e preenchidas antes da criação do DATASUS, com isso foi gerado um número grande de dados inconsistente, em relação com as bases de todo o Brasil com isso está ocorrendo um esforço municipal para a unificação desses dados. Este trabalho está nesse cenário desafiador para a saúde pública brasileira, contribuindo com uma prova de conceito de uma abordagem para a visualização de dados utilizando-se de um modelo de Data Warehouse que está integrando três diferentes bases.

Existem iniciativas via universidades, incluindo o IME-USP, e institutos ligados aos estudos da saúde pública brasileira. Oficialmente, o DATASUS possui o TABNET [6], que disponibiliza os dados de todas as suas bases no formato CSV. Especificamente, a SMS-SP é um caso especial, por trabalhar também com outras bases de dados e sistemas próprios, ela não possui nenhuma ferramenta oficial aprofundada para a visualização e análise de dados, e nenhum tipo de modelo que integra essas diferentes bases em um único local com dados normalizados.

3.1 Proposta de integração

Considerando a falta de integração que as bases do DATASUS possuem, que consequentemente dificulta tanto na busca quanto na análise de seus dados, está sendo proposto neste trabalho uma modelagem de um diagrama entidade relacionamento (DER) para unificar as bases do SINASC, SIM e SIH.

[2] http://www2.datasus.gov.br/DATASUS/index.php?area=060701
[3] http://tabnet.datasus.gov.br/cgi/sih/rxdescr.htm
[4] http://datasus.saude.gov.br/datasus
[5] http://www2.datasus.gov.br/DATASUS/index.php?area=060804
[6] http://www2.datasus.gov.br/DATASUS/index.php?area=02

Outra dificuldade está relacionada a uma grande irregularidade no dados que foram extraidos, como exemplo, nas três bases que trabalhamos o **SIM**, **SINASC** e **SIH**. As colunas do CSV eram diferentes das do dicionário de dados, disponibilizados no TABNET [20], dificultando a população do Data Warehouse. Há irregularidade em campos como datas, em que ela aparece em formatos diferentes dependendo da base utilizada, por exemplo a base SIH utilizava o formato "AAAAMMDD"e a SIM e SINASC utilizavam "DD-MM-AAAA", então foi necessário que houvesse uma formatação da data de "AAAAMMDD"para "DD-MM-AAAA"utilizada na SIH, já que o PostgreSQL não aceita esse formato de data, houve também espaços em branco à direita dos campos que seriam do tipo `int`, o que causa erro no PostgreSQL já que um valor inteiro não possui caracteres. Nesse mesmo caso, houve também colunas em que havia muitos espaços em branco, como pode ser observado na Figura 3.1. Essa quantidade de espaços fez com que o tamanho do arquivo aumentasse e o tamanho do campo `varchar` que tinhamos suposto desse atributo não suportou o tamanho que estava no CSV.

Figura 3.1: *Espaços em branco presente no csv da base SIM.*

Na Figura 3.2 podemos observar de uma forma geral a modelagem do Data Warehouse[7], onde as tabelas que estão expandidas são as tabelas de fatos e as que estão recolhidas são as tabelas de dimensões. Para este modelo foi utilizado o esquema de armazenamento floco de neve, onde há algumas tabelas que foram normalizadas e geraram tabelas adicionais, como por exemplo o atributo data que foi dividido nas tabelas "Ano", "Mes"e "Dia"que formam a "Data". Outras normalizações foram feitas nos atributos de endereço e do CID. As tabelas fato do modelo foram divididas em "Pessoa", "Parto", "NascidoVivo", "Natimorto", "Obito", "Atendimento"e "Internacao". Para ter um maior entendimento do significado de cada um dos atributos da tabelas a seguir, basta ler a documentação das bases utilizadas encontrado no TABNET [20].

[7]Esta modelagem foi feita em conjunto com o aluno Marcos Vinicius do Carmo Sousa, já que ambos os trabalhos iriam necessitar deste Data Warehouse integrando as mesmas bases, SIM, SIH e SINASC. O trabalho dele seria focado nos processos de ETL e o meu na visualização de dados.

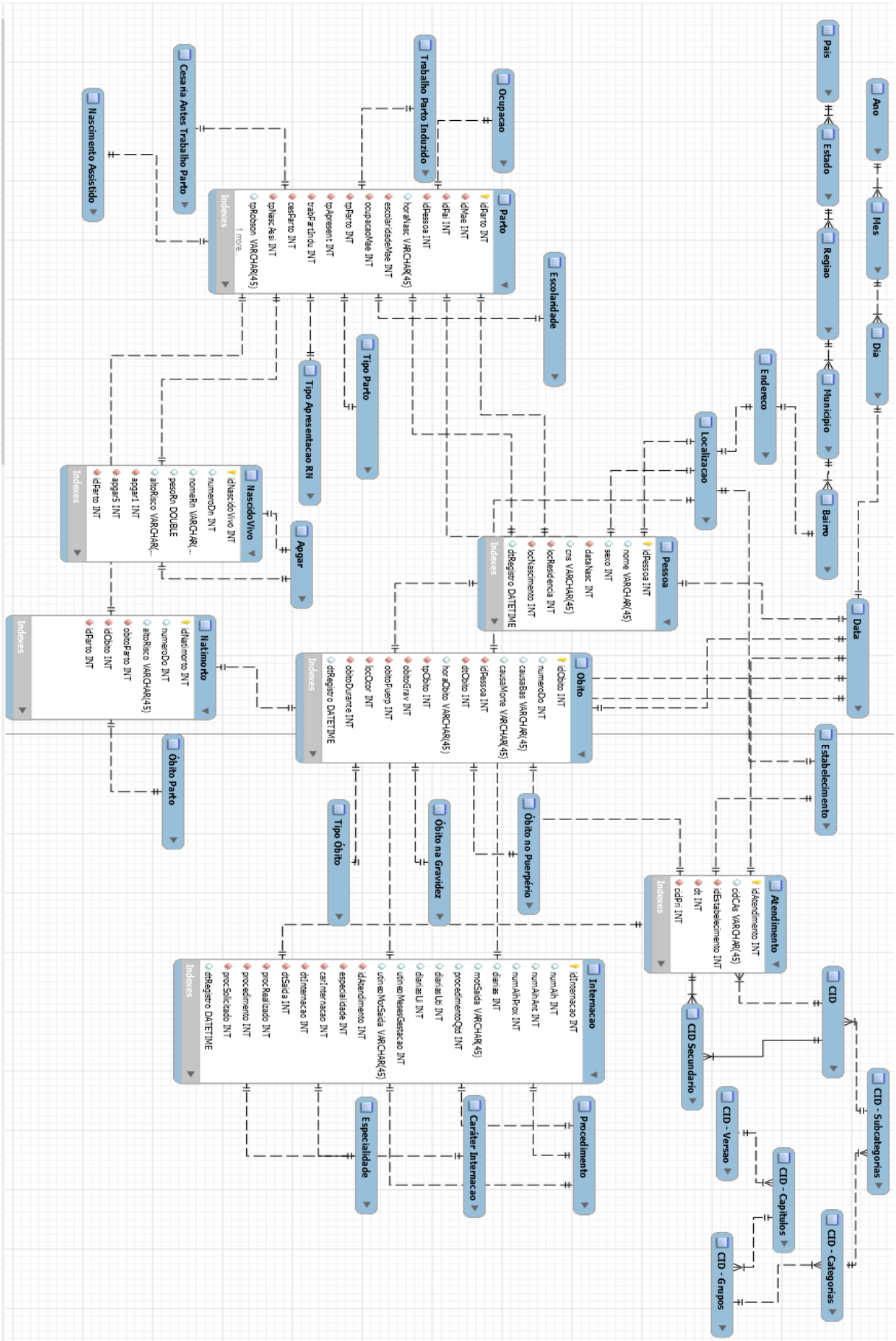


Figura 3.2: *DER do Data Warehouse*

3.1.1 Normalizações

Nessa seção iremos tratar das normalizações realizadas para o modelo floco de neve. Na normalização apresentada na Figura 3.3 há a padronização da data, em que dividimos ela em 4 tabelas em ordem de grandeza, em que elas fazem referencia entre si através de chaves estrangeiras sendo que na tabela dia o atributo `dia` é um inteiro que guarda o seu valor, na tabela `Mes` o atributo `mes` é um `VARCHAR` que guarda o nome do mês e na tabela `Ano` o atributo `ano` é um inteiro que guarda seu valor.

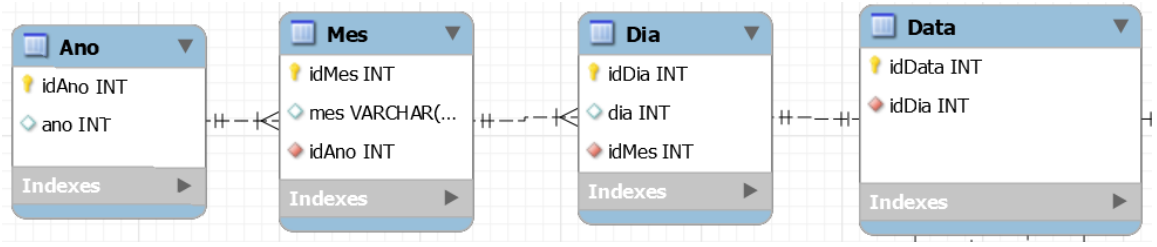


Figura 3.3: *Normalização da data*

Na normalização apresentada na Figura 3.4 há a padronização do CID como o CID é um conjunto de caracteres em que cada um deles significa algo, a normalização foi feita pensando nessa organização, em que um CID é composto por Capítulos, grupos, categorias e subcategorias, e a união de todos eles formam o CID, além disso temos que a cada versão do CID esses caracteres são alterados e o CID secundario é apenas um diagnóstico secundário para o atendimento.

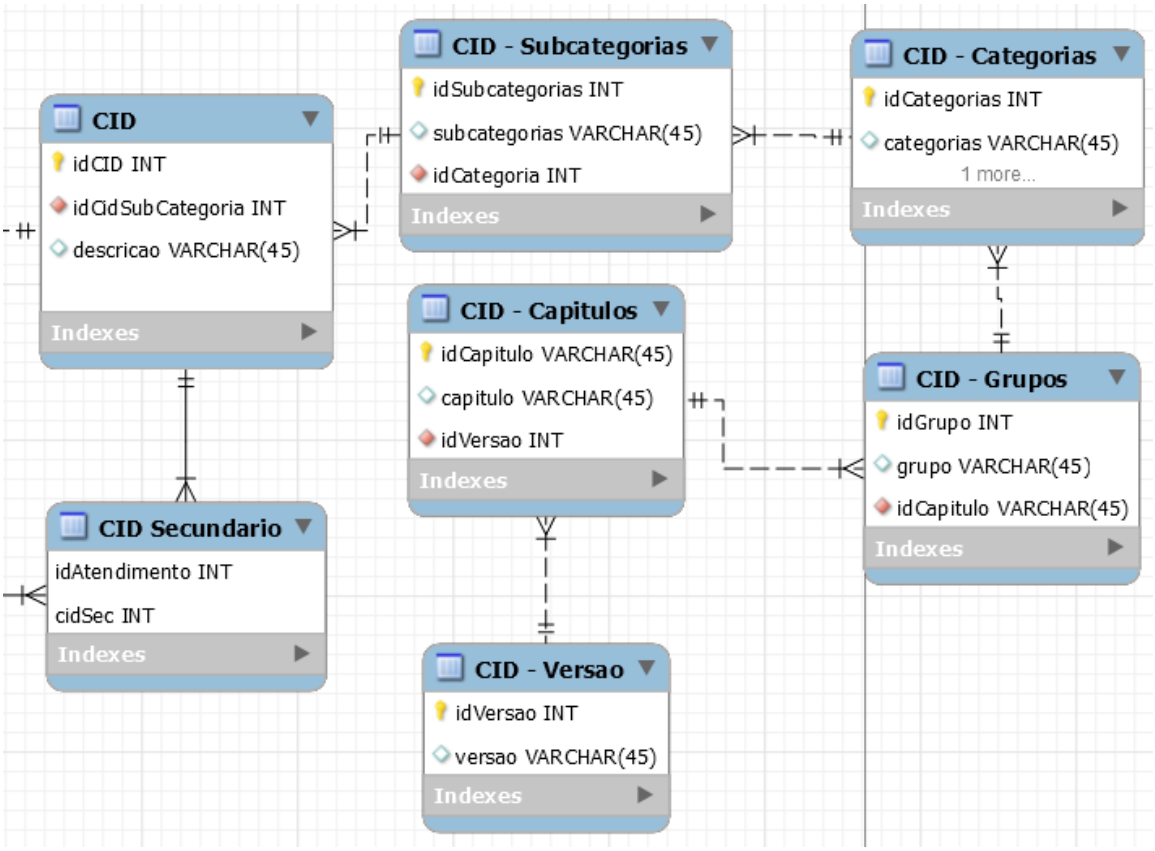


Figura 3.4: *Normalização do CID*

Na normalização apresentada na Figura 3.5 há a padronização da localização em que ela foi organizada em ordem de grandeza dividida em sete tabelas, em que uma localização possui um endereço, um complemento, um distrito e um número, um endereço é composto pelo seu código, bairro, logradouro e CEP, um bairro é composto pelo seu código, munícipio e CEP, um município é composto pelo seu código e região, uma região é composta pelo seu código e estado, um estado é composto pelo seu código e país.

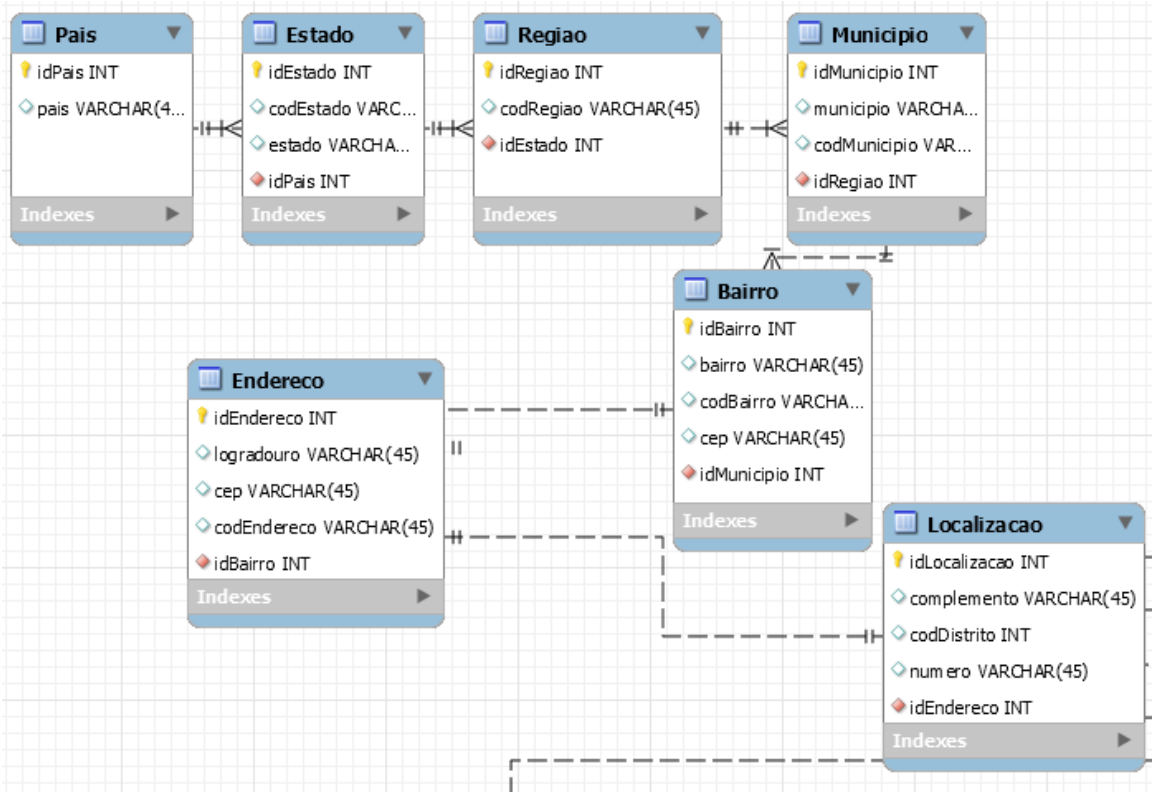


Figura 3.5: *Normalização da localização*

3.1.2 Tabela de fatos e de dimensões

Nessa seção iremos tratar das tabela de fatos e de dimensões para o modelo floco de neve.

Na Tabela de fatos da Figura 3.6 temos como dimensões o `CID`, `CID Secundario`, `data` e `estabelecimento` e como atributo adicional da tabela atendimento o `cidCAs`.

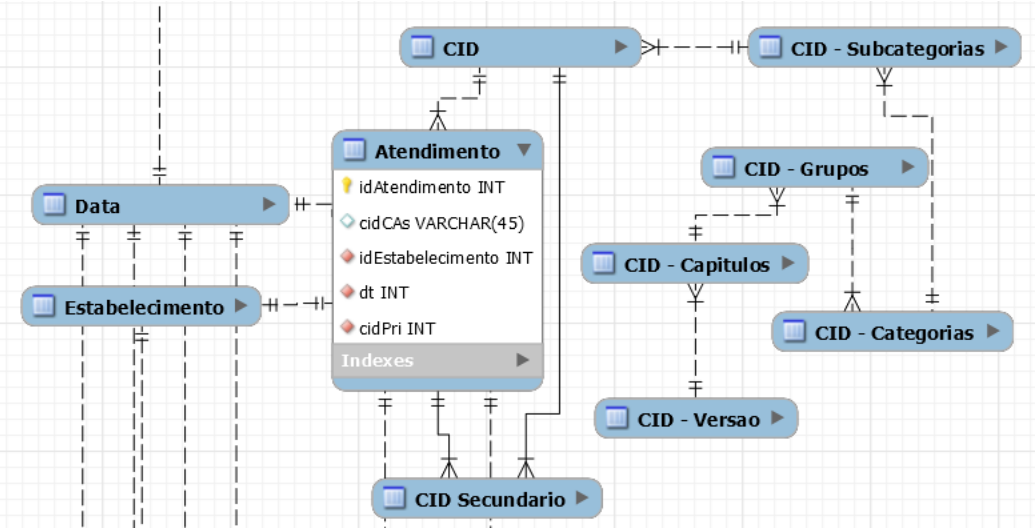


Figura 3.6: *Tabela de fatos do atendimento*

Na Tabela de fatos da Figura 3.7 temos como dimensões a `data` e a `localização` e como atributo adiciona da tabela pessoa temos o `nome`,`sexo`, `cns` e `dtRegistro`.

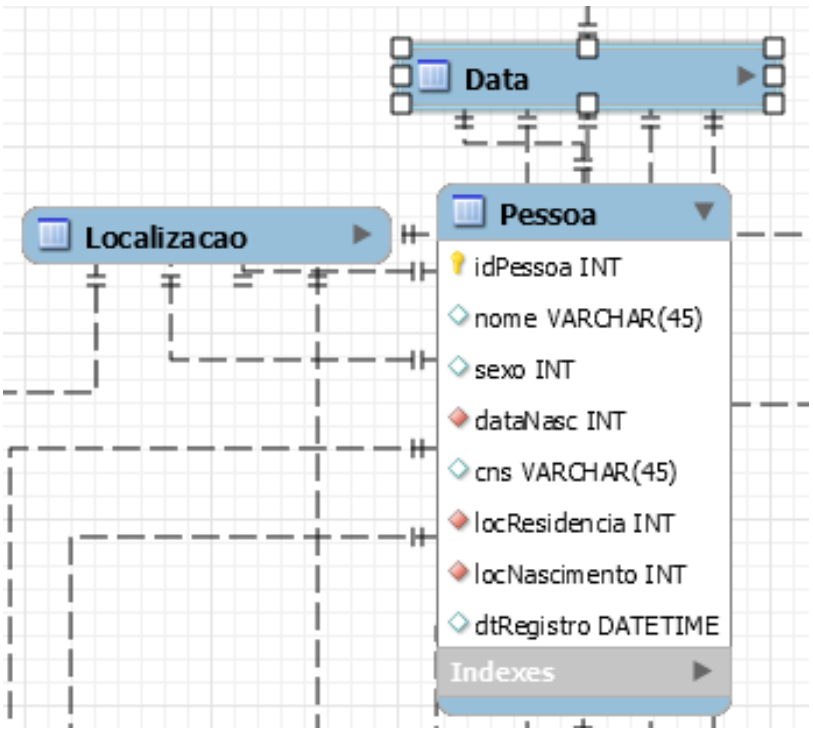


Figura 3.7: *Tabela de fatos das pessoas*

Na Tabela de fatos da Figura 3.8 temos como dimensões a `data`, `atendimento`, `procedimento`, `caráter internacao` e `especialidade`, e como atributos adicionais da tabela Internação temos o `numAih`, `numAihAnt`, `numAihProx`, `diarias`, `motSaida`, `procedimentoQtd`, `diariasUti`, `diariasUi`, `utineoMesesGestacao`, `utineoMotSaida` e `dtRegistro`.

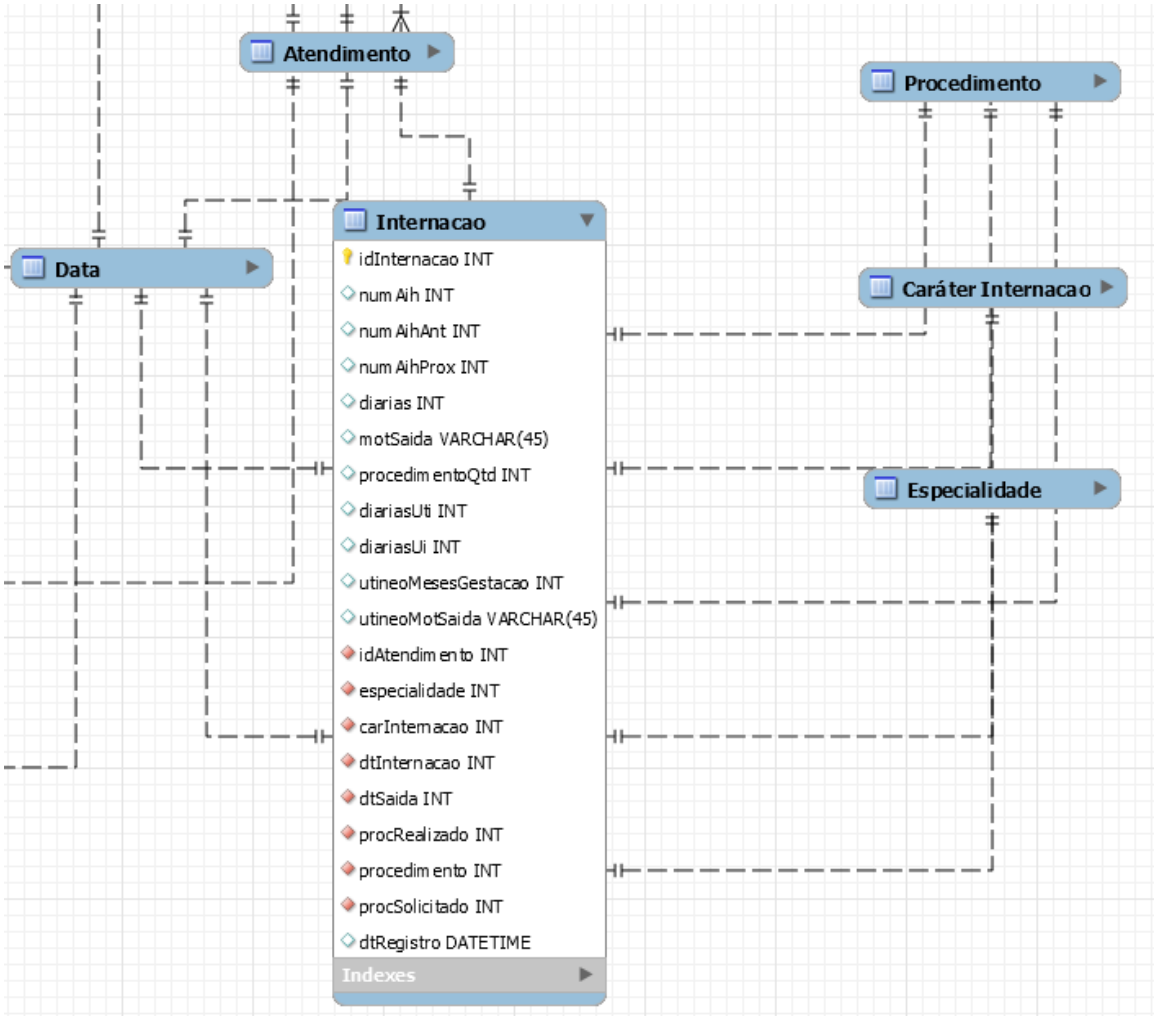


Figura 3.8: *Tabela de fatos das intenações*

Na Tabela de fatos da Figura 3.9 temos como dimensões a `Localização`, `Pessoa`, `Data`, `Óbito no Puerpério`, `Atendimento`, `Óbito na Gravidez` e `Tipo Óbito`, e como atributos adicionais da tabela Obito temos o `numeroDo`, `causaBas`, `causaMorte`, `horaObito` e `dtRegistro`.

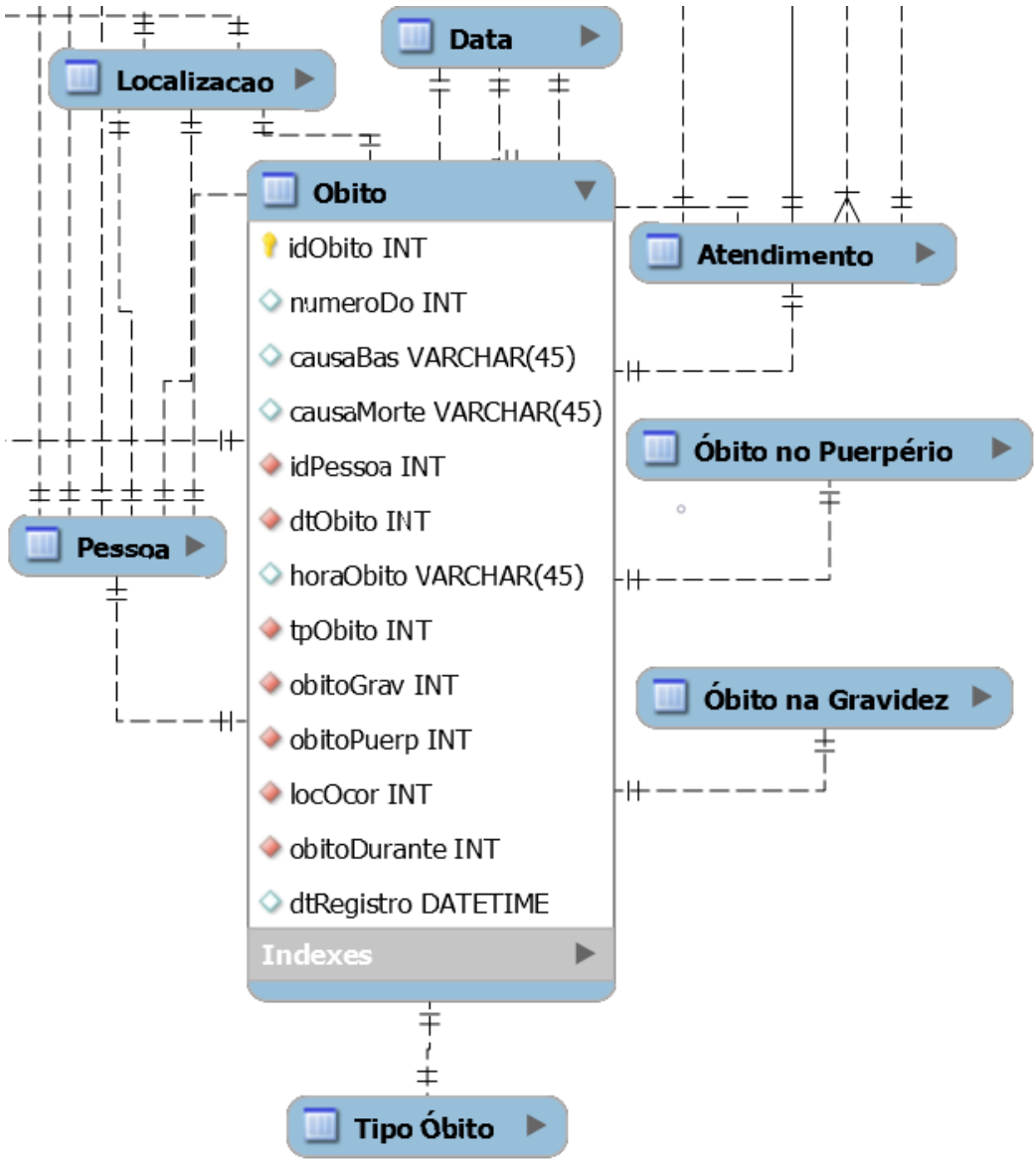


Figura 3.9: *Tabela de fatos dos óbitos*

Na Tabela de fatos da Figura 3.10 temos como dimensões `Nascimento Assistido`, `Cesaria Antes Trabalho Parto`, `Trabalho Parto Induzido`, `Ocupacao`, `Pessoa`, `Escolaridade`, `Tipo Parto` e `Tipo Apresentacao RN` e como atributos adicionais da tabela Parto temos a `horaNasc`, `tpRobson` e `dtRegistro`.

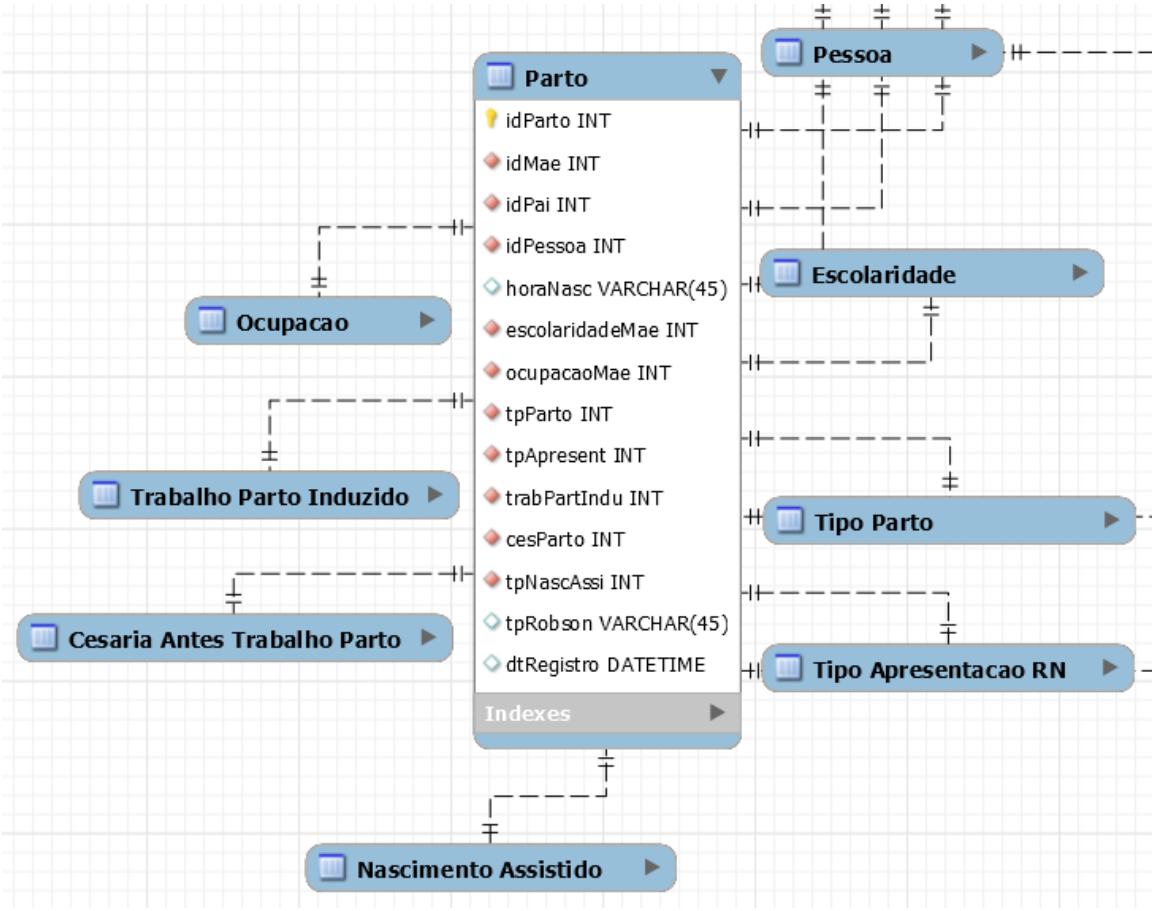


Figura 3.10: *Tabela de fatos dos partos*

Na Tabela de fatos da Figura 3.11 temos como dimensões `Parto`, `Obito` e `Óbito Parto` e como atributos adicionais da tabela natimorto temos o `numeroDo` e `altoRisco`.

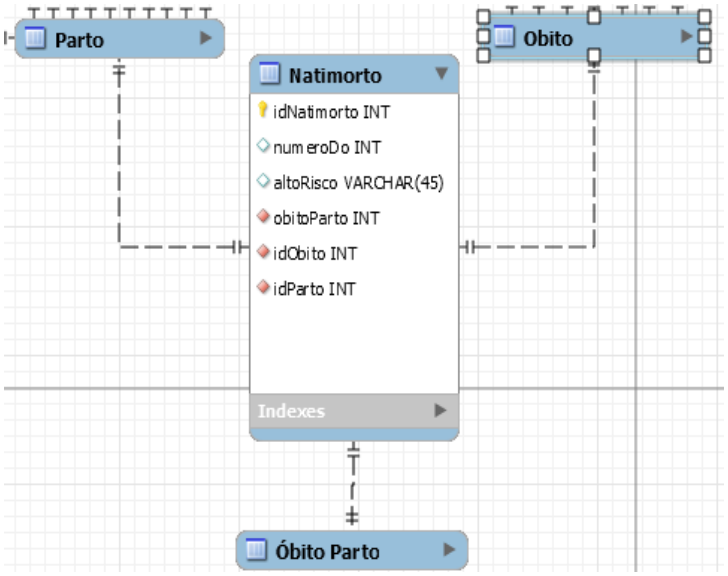


Figura 3.11: *Tabela de fatos dos natimortos*

Na Tabela de fatos da Figura 3.12 temos como dimensões `Parto` e `Apgar` e como atributos adicionais da tabela NascidoVivo temos o `numeroDo`, `nomeRn`, `pesoRn` e `altoRisco`.

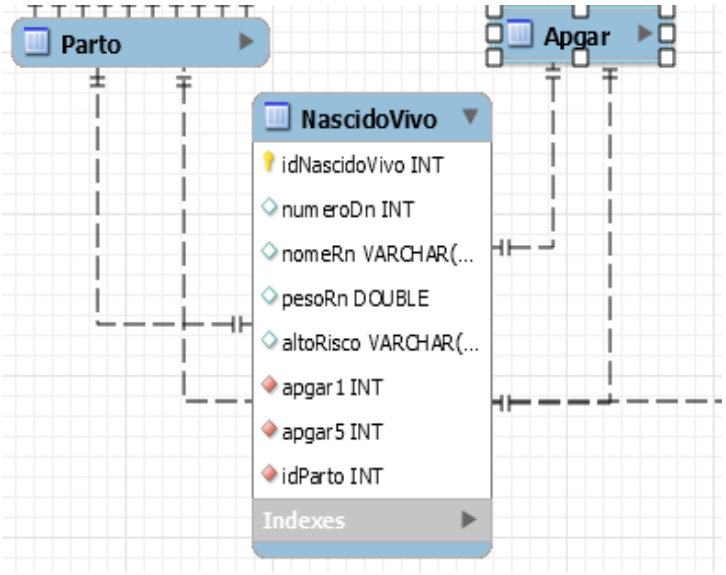


Figura 3.12: *Tabela de fatos dos nascidos vivos*

Esse será o Data Warehouse utilizado para as análises de dados, que então, montará os cubos ou hipercubos, a partir de algumas dessas tabelas fato pré-calculando seus atributos, para que agilize as pesquisas e montagens dos gráficos, que será feita pela ferramenta **Superset**.

Capítulo 4

Cenários de decisão

Neste capítulo, discutimos as análises que iremos realizar com os dados no modelo de Data Warehouse mostrado no capítulo anterior. Como o Superset não é uma ferramenta específica para Data Warehouse, foi necessário contornar essas dificuldades com comandos do PostgreSQL, sendo que para a visualização de dados o que é fundamental são os cubos e para isso foi utilizado o comando CREATE MATERIALIZED VIEW, onde é pré-montada a tabela de interesse para as pesquisas, sendo que ela é criada da seguinte forma:

```
CREATE MATERIALIZED VIEW [ IF NOT EXISTS ] table_name
    [ (column_name [, ...] ) ]
    [ WITH ( storage_parameter [= value] [, ... ] ) ]
    [ TABLESPACE tablespace_name ]
    AS query
    [ WITH [ NO ] DATA ]
```

Isto resulta na criação de uma tabela VIEW onde as análises serão feitas de forma mais ágil. Caso seja necessário atualizar os dados, o que é algo bem comum em data warehouse, existe um comando REFRESH, que executará novamente a query do view e preencherá a mesma tabela com os dados atualizados, esse comando é utilizado como:

```
REFRESH MATERIALIZED VIEW name
    [ WITH [ NO ] DATA ]
```

4.1 Análises na tabela fato parto

Nesta seção iremos tratar das análises que estão relacionadas a tabela fato `parto` da Figura 3.10.

4.1.1 Nascimentos por dias da semana

Esta pesquisa foi feita utilizando a operação ‘’dice” na tabela fato `parto`, no intervalo de tempo do ano de 2016. O gráfico mostra a quantidade de partos pelos dias do ano de

2016, sendo que na primeira linha da Figura 4.1 é a segunda-feira e a última o domingo e as cores mais claras representam uma menor quantidade de partos e as mais escuras uma maior.

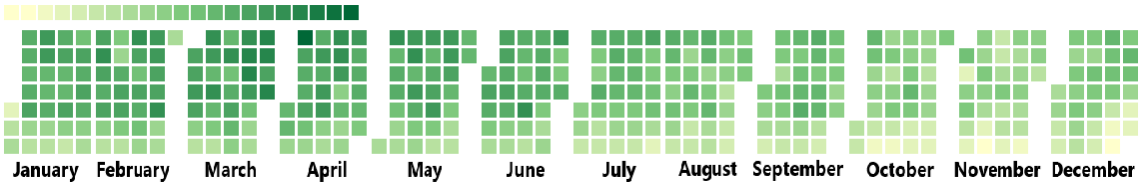


Figura 4.1: *Nascimento por mês*

Por meio da Figura 4.1, pode-se observar que há uma maior preferência de partos em dias de semana do que nos fins de semanas, pois pode-se perceber uma maior presença de cores escuras de segunda à sexta e de cores claras para sábado e domingo. O que responde à questão que a secretaria possui, em que os médicos preferem remarcar se possível os partos para os dias da semana.

Se realizarmos a operação ‘’dice” na tabela fato `parto`, no intervalo de tempo do mês de dezembro de 2016, obtemos o gráfico que mostra a quantidade de partos pelos dias do mês de dezembro de 2016, como podemos observar na Figura 4.2.

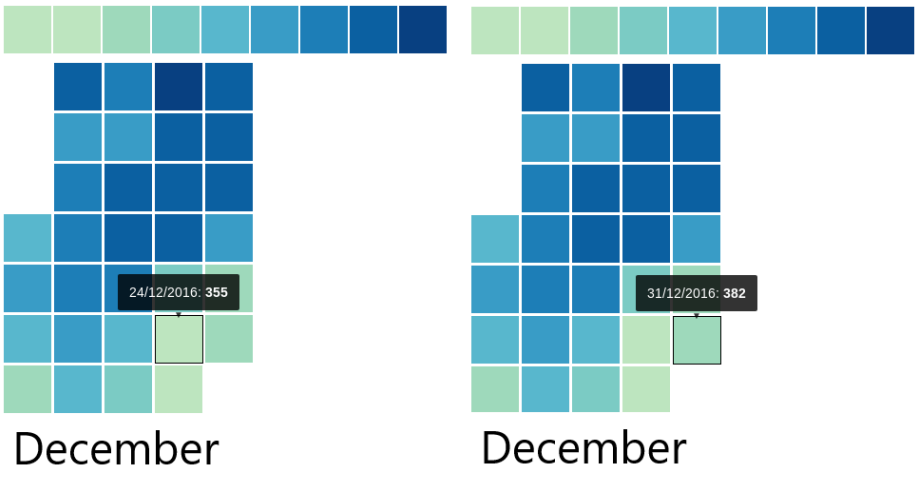
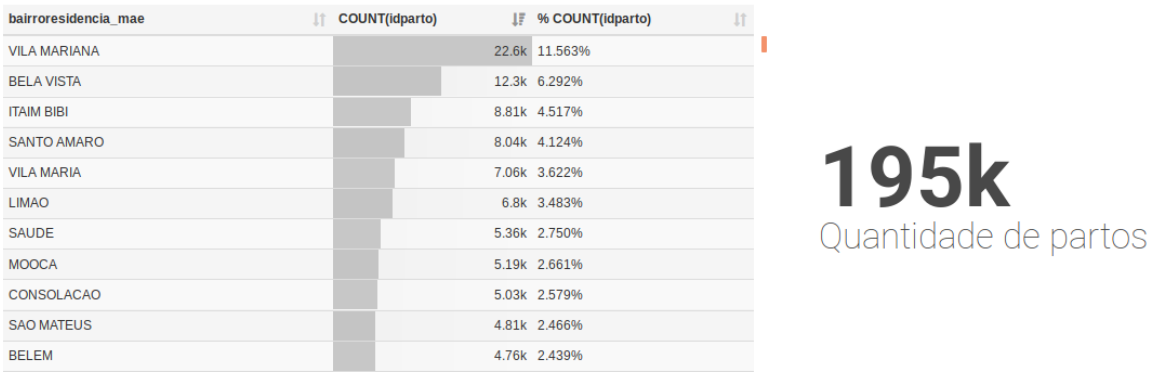


Figura 4.2: *Nascimento por dia no mês de dezembro de 2016.*

Na Figura 4.2 podemos observar que há uma maior presença de cores claras perto dos dias em que há feriado, como nos dias 23, 24, 25, dias próximos ao Natal e nos dias 30 e 31 próximos ao ano novo, isso é similar ao que ocorre nos partos durante fim de semana, nesse caso os feriados ocorreram no fim de semana também, então houve uma quantidade menor ainda de partos, a relação disso é o mesmo que no caso anterior, em que os médicos preferem remarcar se possível os partos para os dias que não são feriados.

4.1.2 Relação de escolaridade com o tipo do parto

Esta pesquisa foi feita utilizando a operação "dice" na tabela fato `parto`, no intervalo de tempo do ano de 2016, para analisar se há uma maior preferência de mães com baixa escolaridade em realizar partos normais e as com alta escolaridade em cesárias. Temos na Figura 4.3 o número total de partos por bairro.



bairroresidencia_mae	COUNT(idparto)	% COUNT(idparto)
VILA MARIANA	22.6k	11.563%
BELA VISTA	12.3k	6.292%
ITAIM BIBI	8.81k	4.517%
SANTO AMARO	8.04k	4.124%
VILA MARIA	7.06k	3.622%
LIMAO	6.8k	3.483%
SAUDE	5.36k	2.750%
MOOCA	5.19k	2.661%
CONSOLACAO	5.03k	2.579%
SAO MATEUS	4.81k	2.466%
BELEM	4.76k	2.439%

Figura 4.3: *Quantidade total de partos por bairros.*

Se fizermos no gráfico da Figura 4.3 a operação de "slice" no cubo nas dimensões da `escolaridade da mãe` e `tipo do parto` por meio do menu de seleções da Figura 4.4, geramos a Figura 4.5 e, se alterarmos o Tipo do parto para "Normal", obtém-se o gráfico da Figura 4.6.

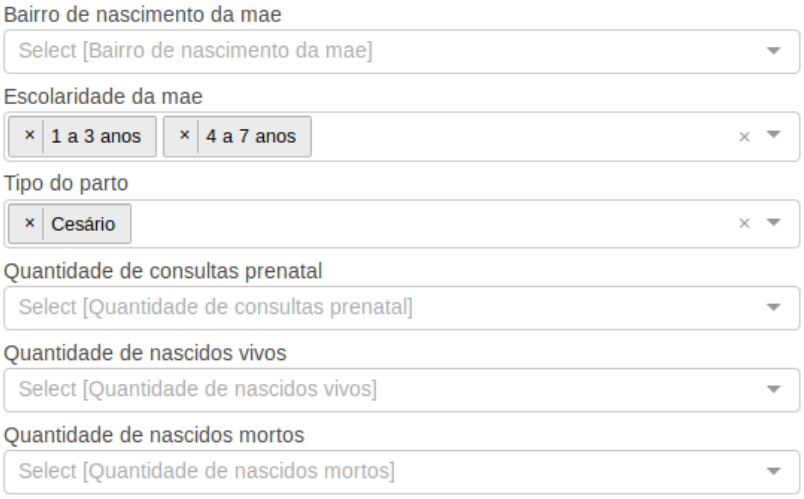


Figura 4.4: *Variações para as pesquisas.*

bairroresidencia_mae	COUNT(idparto)	% COUNT(idparto)
SANTO AMARO	332	6.206%
LIMAO	268	5.009%
JARDIM SAO LUIS	242	4.523%
JARDIM ANGELA	225	4.206%
SANTANA	223	4.168%
BUTANTA	216	4.037%
BELEM	200	3.738%
CIDADE DUTRA	184	3.439%
CONSOLACAO	174	3.252%
SAO MATEUS	168	3.140%
VILA MARIANA	156	2.916%

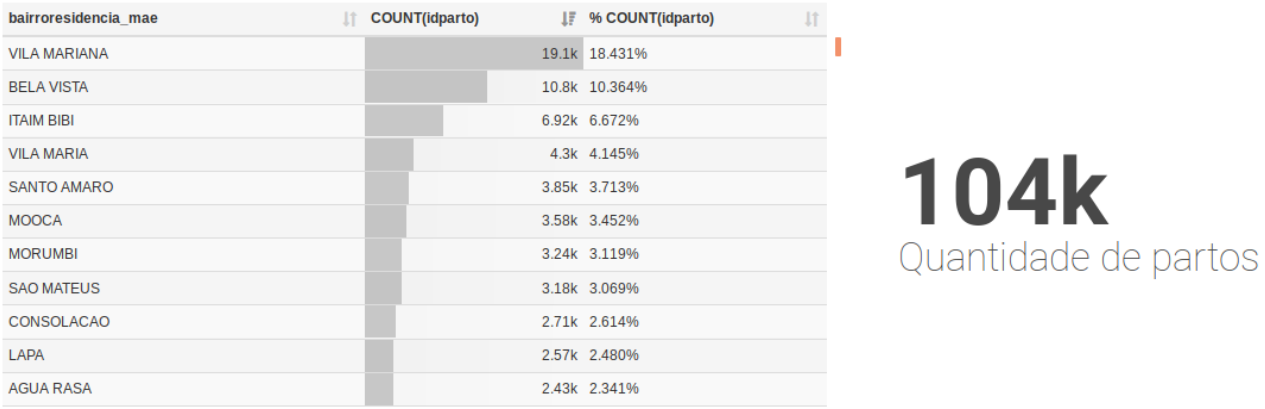
Figura 4.5: *Quantidade de nascimentos por Bairro, filtrados pelo seletor da Figura 4.4.*

bairroresidencia_mae	COUNT(idparto)	% COUNT(idparto)
LIMAO	569	5.559%
SANTO AMARO	525	5.129%
JARDIM ANGELA	481	4.699%
BUTANTA	420	4.103%
SAUDE	400	3.908%
JARDIM SAO LUIS	390	3.810%
CIDADE TIRADENTES	378	3.693%
JARAGUA	375	3.664%
ITAIM PAULISTA	364	3.556%
BELEM	345	3.370%
SAO MIGUEL	334	3.263%

Figura 4.6: *Quantidade de nascimentos por Bairro, filtrados pelo seletor da Figura 4.4, com o tipo de parto alterado para "Normal".*

Como podemos observar nas Figuras 4.5 e 4.6, os bairros de Santo Amaro e do Limão lideram a tabela com uma maior quantidade de partos, considerando a filtragem da Figura 4.4. Além disso, ao compararmos a quantidade total de partos, podemos observar que há cerca de 2 vezes mais partos normais, com mães com escolaridade entre 1 a 7 anos.

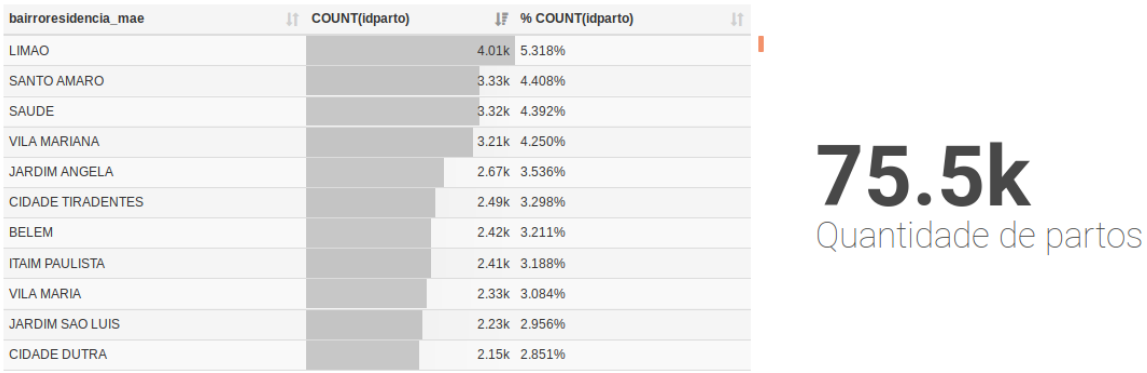
Se alterarmos o intervalo da escolaridade da mãe de 1 a 7 anos para 8 a 12, ou mais, na Figura 4.4, teremos a Figura 4.7, que representa a quantidade de cesárias realizadas por mães com maior grau de estudo.



bairroresidencia_mae	COUNT(idparto)	% COUNT(idparto)
VILA MARIANA	19.1k	18.431%
BELA VISTA	10.8k	10.364%
ITAIM BIBI	6.92k	6.672%
VILA MARIA	4.3k	4.145%
SANTO AMARO	3.85k	3.713%
MOOCA	3.58k	3.452%
MORUMBI	3.24k	3.119%
SAO MATEUS	3.18k	3.069%
CONSOLACAO	2.71k	2.614%
LAPA	2.57k	2.480%
AGUA RASA	2.43k	2.341%

Figura 4.7: *Quantidade de nascimentos por Bairro, filtrados pela 4.4, com a escolaridade da mãe alterada para "8 a 12 ou mais anos".*

Mudando novamente a opção para partos normais, teremos como resultado a Figura 4.8 que representa a quantidade de partos normais realizadas por mães com maior grau de estudo.



bairroresidencia_mae	COUNT(idparto)	% COUNT(idparto)
LIMAO	4.01k	5.318%
SANTO AMARO	3.33k	4.408%
SAUDE	3.32k	4.392%
VILA MARIANA	3.21k	4.250%
JARDIM ANGELA	2.67k	3.536%
CIDADE TIRADENTES	2.49k	3.298%
BELEM	2.42k	3.211%
ITAIM PAULISTA	2.41k	3.188%
VILA MARIA	2.33k	3.084%
JARDIM SAO LUIS	2.23k	2.956%
CIDADE DUTRA	2.15k	2.851%

Figura 4.8: *Quantidade de nascimentos por Bairro, filtrados pela 4.4, com a escolaridade da mãe alterada para "8 a 12 ou mais anos"e com o tipo de parto alterado para "Normal".*

Analisando as Figuras 4.7 e 4.8, percebemos que o bairro da Vila Mariana possui grandes quantidades de cesárias e para partos normais há uma menor desigualdade entre os nascimentos dos bairros. Também podemos observar que, a quantidade de cesáreas é cerca de 38% maior que a de partos normais, para mães com escolaridade entre 8 a 12 ou mais anos.

Considerando as duas análises realizadas nas Figuras 4.5, 4.6, 4.7 e 4.8, podemos observar que há uma maior preferência em mães com baixa escolaridade a ter partos normais, e em mães com uma maior escolaridade a preferência é a de cesárea, onde a mãe pode ter um maior controle da data em que o bebê irá nascer.

4.2 Análises na tabela fato óbito

Nesta seção iremos tratar das análises que estão relacionadas à tabela fato `obito` da Figura 3.9.

4.2.1 Doenças que mais causaram óbitos por estação do ano

Esta pesquisa foi feita utilizando a operação "dice" na tabela fato `obito`, no intervalo de tempo das estações do ano de 2016, para analisar se há ou não uma relação de algumas doenças com as estações do ano, sendo que a Figura 4.10 representa o verão, 4.11 o outono, 4.12 o inverno e 4.13 a primavera, e a operação "dice" no cubo foi feita com o filter box da Figura 4.9

Figura 4.9: *Filtro de intervalo de tempo, com o intervalo que representa o outono de 2016.*

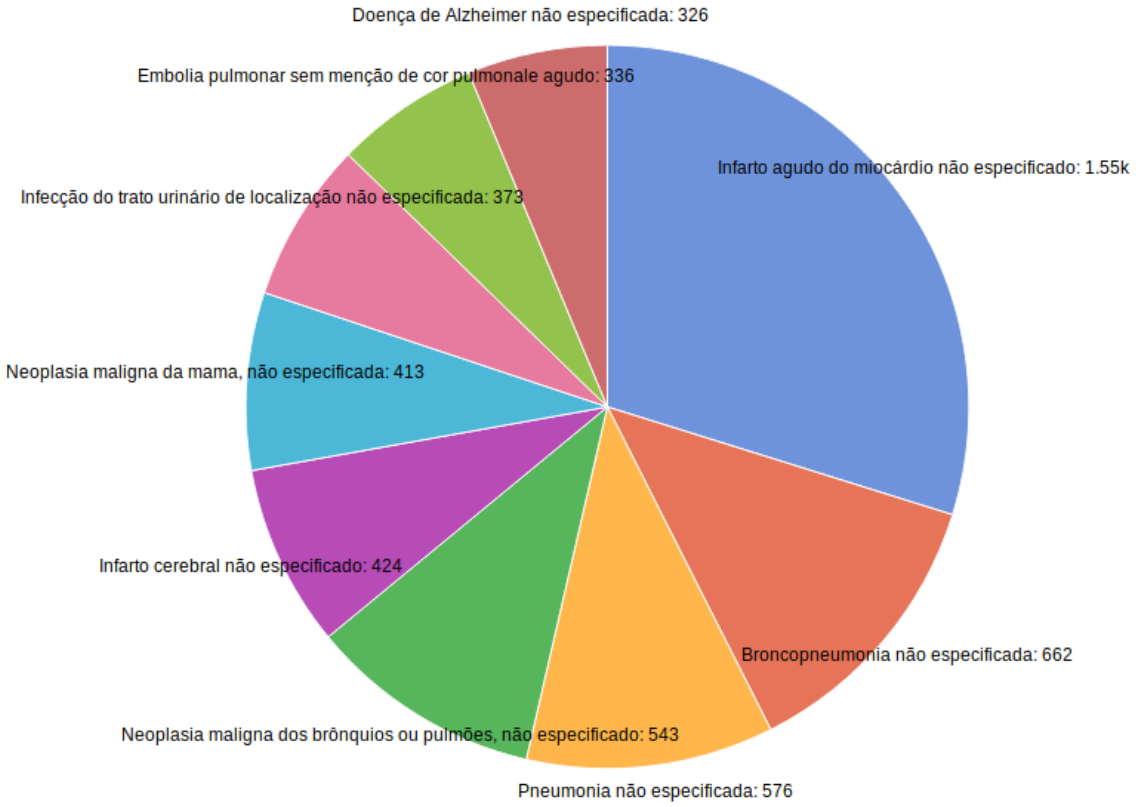


Figura 4.10: *9 doenças que mais causaram óbito no verão de 2016.*

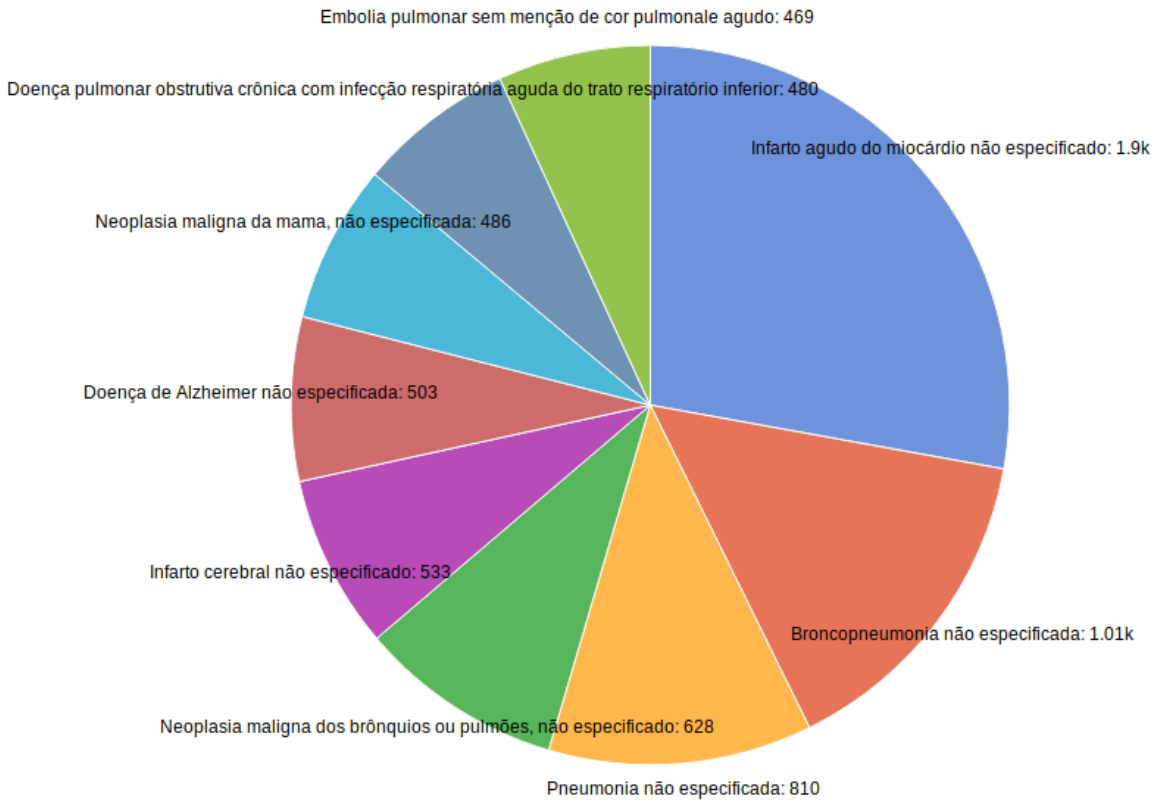


Figura 4.11: *9 doenças que mais causaram óbito no outono de 2016.*

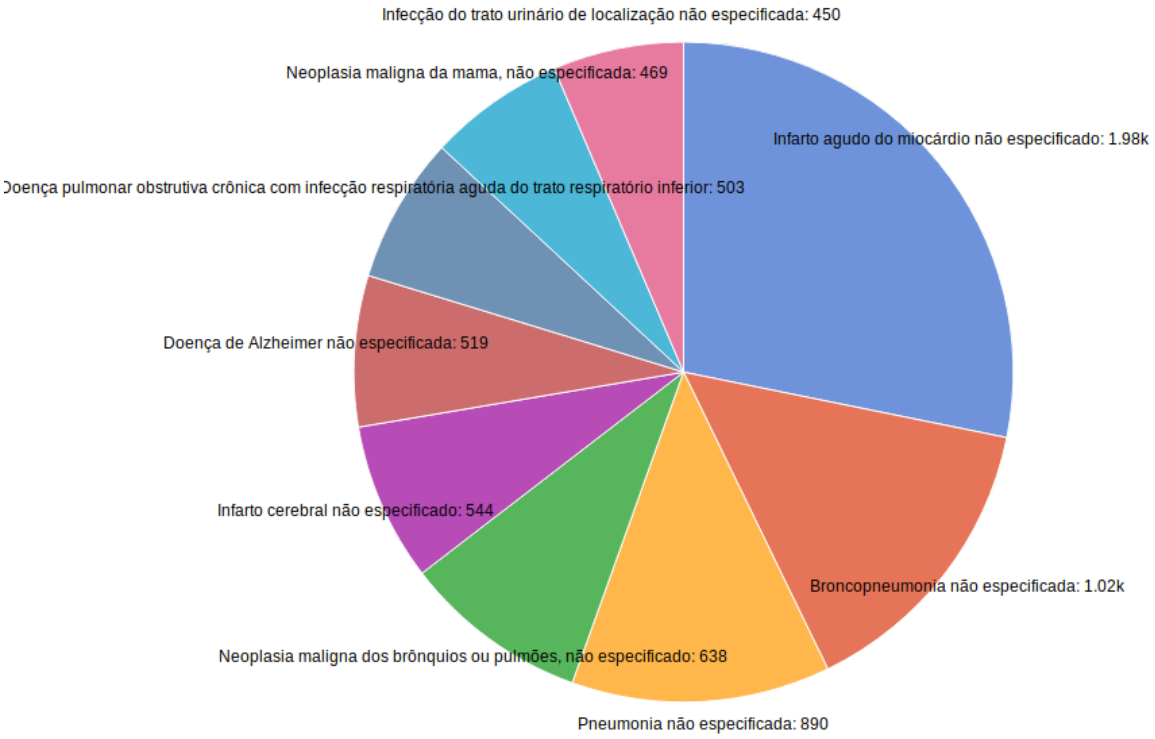


Figura 4.12: *9 doenças que mais causaram óbito no inverno de 2016.*

Figura 4.13: *9 doenças que mais causaram óbito no primavera de 2016.*

Analisando as Figuras 4.10, 4.11, 4.12 e 4.13, podemos observar que as estções do ano não influenciam nas 5 primeiras doenças. Porém, no outono e no inverno pode-se observar o surgimento de uma doença: a "Doença pulmonar obstrutiva crônica com infecção respiratória aguda do trato respiratório inferior", que é algo esperado já que essa doença é agravada quando há um acumulo de poluentes no ar, que normalmente ocorre em épocas com chuva reduzida, que é o caso do outono e do inverno.

4.2.2 Óbitos relacionados ao puerpério

Esta pesquisa foi feita utilizando a operação “dice” na tabela fato `obito`, no intervalo de tempo do ano de 2016, e a operação de “slice” na dimensão do `óbito no puerpério`, para analisar se há alguma doença que afeta mais o óbito no puerpério, sendo que a Figura 4.10 representa o verão, 4.11 o outono, 4.12 o inverno e 4.13 a primavera, e a operação "dice"e "slice"no cubo foi feita com o “filter box” da Figura 4.14

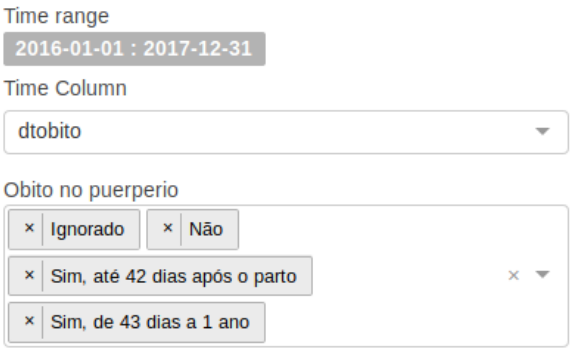


Figura 4.14: *Filtro de Óbito no puerpério com todas as opções selecionadas.*

descricao	COUNT(idobito)	% COUNT(idobito)
Outras doenças e afecções especificadas complicando a gravidez, o parto e o puerpério	18	11.321%
Doenças do aparelho circulatório complicando a gravidez, o parto e o puerpério	17	10.692%
Doenças do aparelho respiratório complicando a gravidez, o parto e o puerpério	14	8.805%
Infecção puerperal	10	6.289%
Outras hemorragias do pós-parto imediato	10	6.289%
Eclâmpsia na gravidez	8	5.031%
Embolia obstétrica por coágulo de sangue	6	3.774%
Pré-eclâmpsia grave	6	3.774%

159
Total de óbitos

Figura 4.15: *Doenças que mais causaram óbito até 42 dias após o parto.*

descricao	COUNT(idobito)	% COUNT(idobito)
Outras causas mal definidas e as não especificadas de mortalidade	8	6.957%
Cardiomiopatia no puerpério	7	6.087%
Neoplasia maligna da mama, não especificada	6	5.217%
Lúpus eritematoso disseminado [sistêmico] com comprometimento de outros órgãos e sistemas	6	5.217%
Infarto agudo do miocárdio não especificado	6	5.217%
Edema pulmonar, não especificado de outra forma	5	4.348%

115
Total de óbitos

Figura 4.16: *Doenças que mais causaram óbito de 43 dias a 1 ano após o parto.*

descricao	COUNT(idobito)	% COUNT(idobito)
Infarto agudo do miocárdio não especificado	14.1k	7.803%
Broncopneumonia não especificada	6.91k	3.815%
Pneumonia não especificada	5.72k	3.160%
Neoplasia maligna dos brônquios ou pulmões, não especificado	4.75k	2.620%
Infarto cerebral não especificado	4.16k	2.298%
Infecção do trato urinário de localização não especificada	3.48k	1.923%
Neoplasia maligna da mama, não especificada	3.47k	1.918%
Doença de Alzheimer não especificada	3.45k	1.904%

181k
Total de óbitos

Figura 4.17: *Doenças que mais causaram óbito sem relação com o puerpério.*

Analisando as Figuras 4.15, 4.16 e 4.17, percebemos que as seis principais doenças das duas primeiras são completamente diferentes, então não há nenhuma doença em comum entre as duas opções de puerpério. As únicas doenças que aparecem em comum são o Infarto agudo do miocárdio não especificado e Neoplastia maligna da mama, não especificada, entre as Figuras 4.16 e 4.17, sendo que o Infarto foi a doença que mais matou no ano de 2016, com mais de 14000 casos.

4.3 Análises na tabela fato internação

Nesta seção discutimos as análises que estão relacionadas a tabela fato `internacao` da Figura 3.8.

4.3.1 Média de diáreas na internação

Esta pesquisa foi feita utilizando a operação "dice" na tabela fato `internacao`, no intervalo de tempo de meados do ano de 2015 ao final de 2016, para analisar se há procedimentos para uma mesma doença que possui menor tempo de diárias na internação. Para isso, temos a Figura 4.18 que mostra os quinze procedimentos com maior número de diárias, não foi utilizado em ordem decrescente de diárias pois, as médias não variavam muito, sempre resultavam em uma dária. Realizamos as pesquisas nas três doenças que mais causaram internações apresentadas na Figura 4.19.

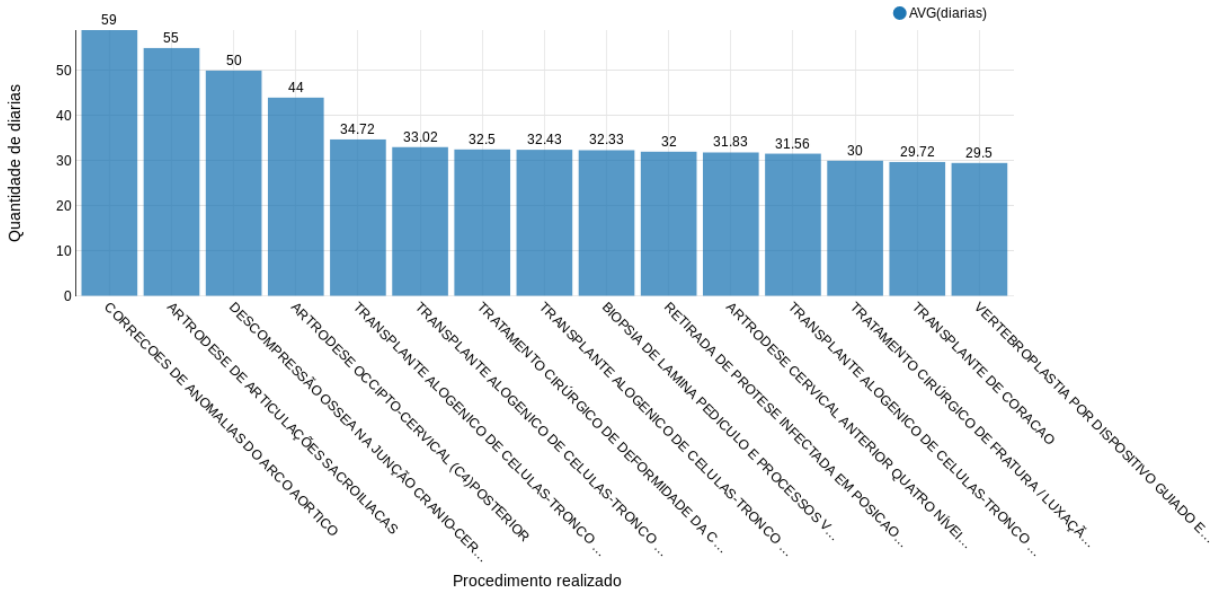


Figura 4.18: *Procedimentos com maior média de diárias.*

cid_descricao	COUNT(idinternacao)	% COUNT(idinternacao)
Parto espontâneo cefálico	43k	18.713%
Parto único espontâneo, não especificado	19.1k	8.305%
Broncopneumonia não especificada	15.2k	6.628%

Figura 4.19: *Três doenças que mais causaram internações.*

Considerando as doenças da Figura 4.19, alteramos a pesquisa com o auxílio do "filter box" da Figura 4.20. Utilizando a operação "slice" no cubo, na dimensão doença, foi escolhido o parto espontâneo cefálico como na Figura 4.21, o parto único espontâneo, não especificado que na Figura 4.22 e a broncopneumonia não especificada na Figura 4.23.

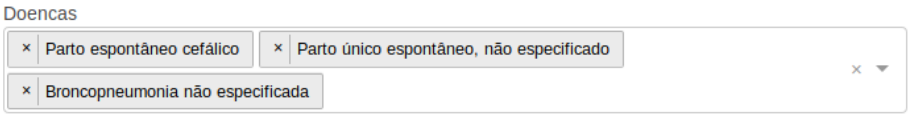


Figura 4.20: *Filtro utilizando as três doenças que mais causaram internações.*

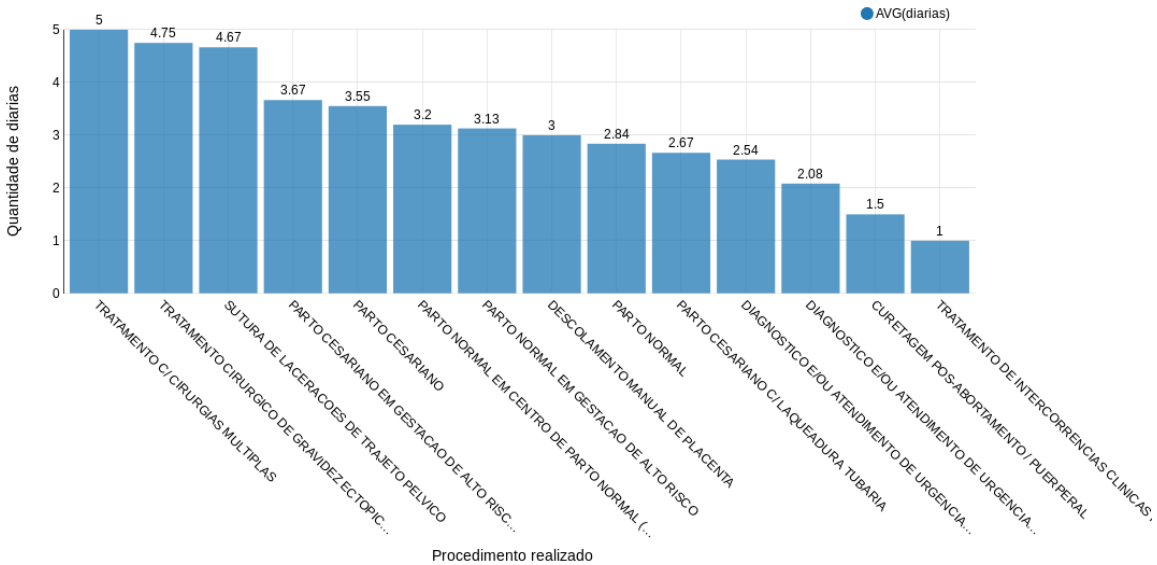


Figura 4.21: *Procedimentos com mais diárias para o parto espontâneo cefálico.*

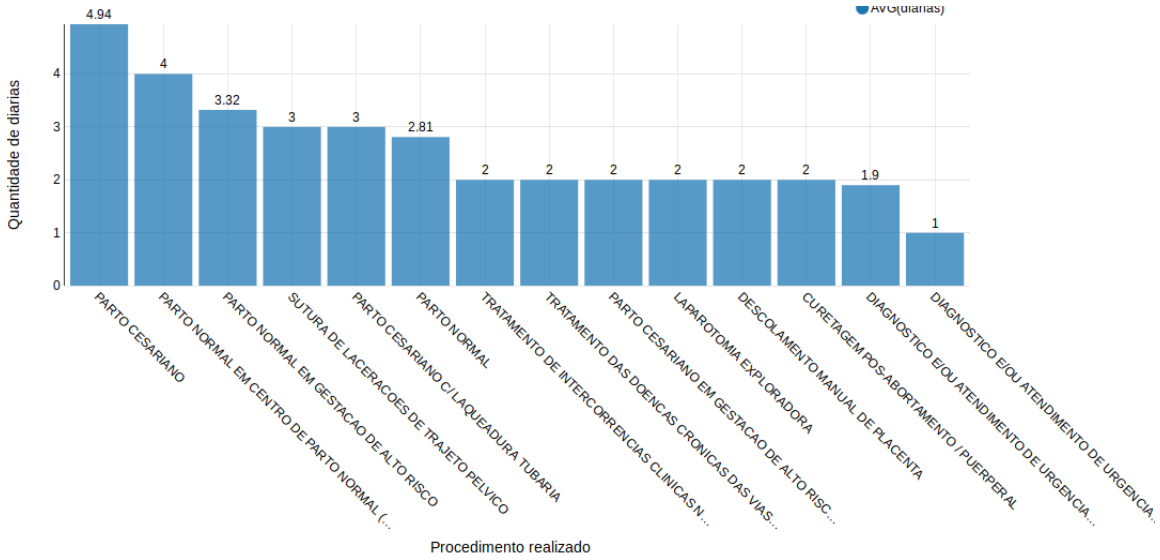


Figura 4.22: *Procedimentos com mais diárias para o parto único espontâneo, não especificado.*

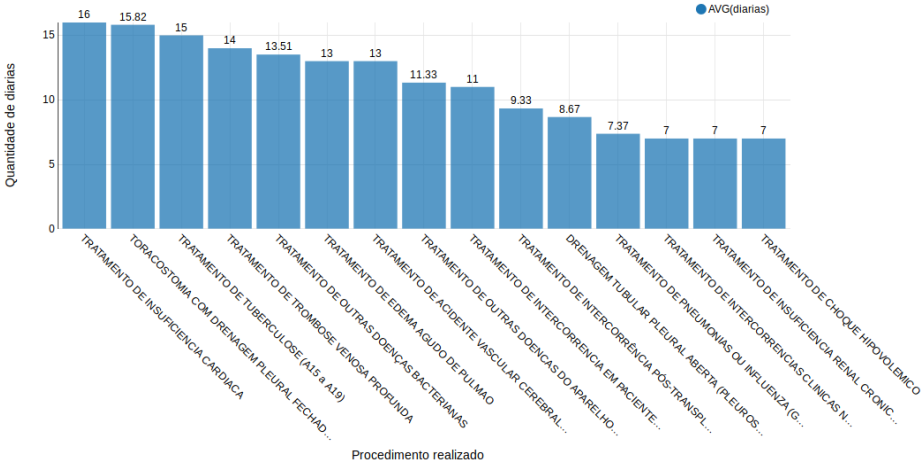


Figura 4.23: *Procedimentos com mais diárias para a broncopneumonia não especificada.*

Analisando as Figuras 4.21, 4.22 e 4.23, podemos observar que para as três doenças que mais causaram internações em 2015 e 2016, há pelo menos quinze procedimentos diferentes para curá-las, sendo que em todos os casos há uma diferença de pelo menos duas vezes da média de diárias mais alta para a mais baixa, como pode-se observar na Figura 4.23.

4.3.2 Especialidades mais frequentes nas internações

Esta pesquisa foi feita utilizando a operação "dice" na tabela fato `internacao`, no intervalo de tempo de meados do ano de 2015 ao final de 2016, para analisar quais especialidades são mais utilizadas para uma certa doença. Para isso temos a Figura 4.24 que mostra as especialidades que mais foram utilizadas em internações. Realizaremos as pesquisas nas três doenças que mais causaram internações apresentadas na Figura 4.19.

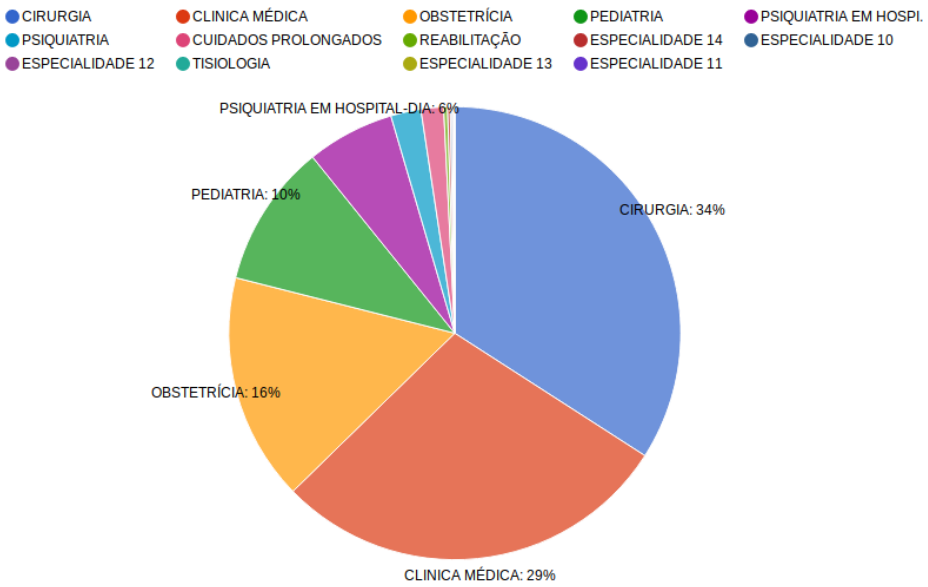


Figura 4.24: *Quantidade de internações por especialidades.*

Considerando as doenças da Figura 4.19, alteramos a pesquisa com o auxílio do "filter box" da Figura 4.20. Utilizando a operação "slice" no cubo, na dimensão doença, foi escolhido o parto espontâneo cefálico que gerou a Figura 4.25, o parto único espontâneo, não especificado que gerou a Figura 4.26 e a broncopneumonia não especificada que gerou a Figura 4.27.

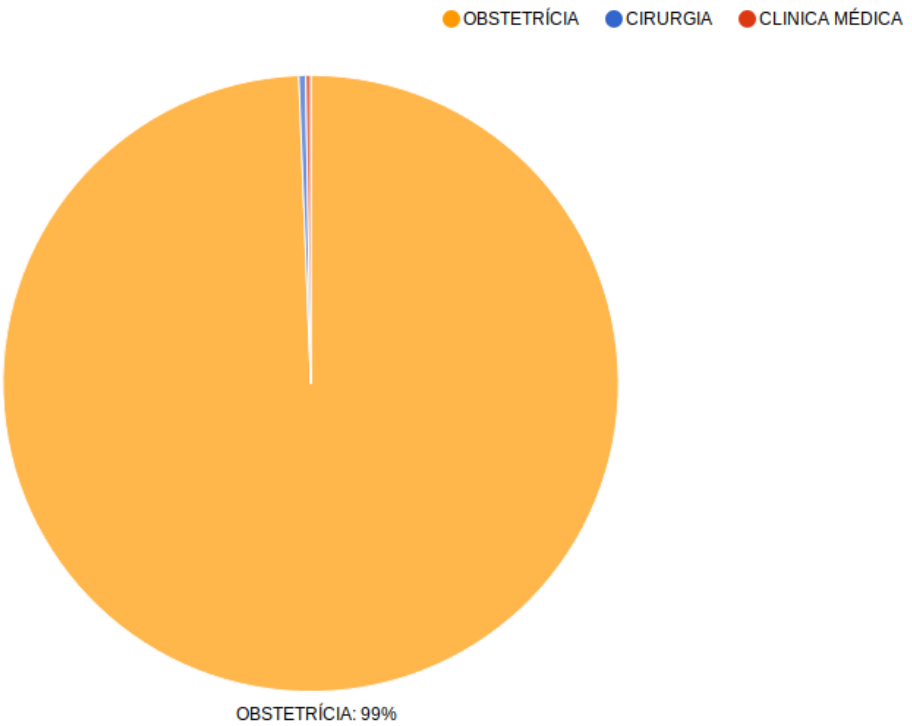


Figura 4.25: *Especialidades para o parto espontâneo cefálico.*

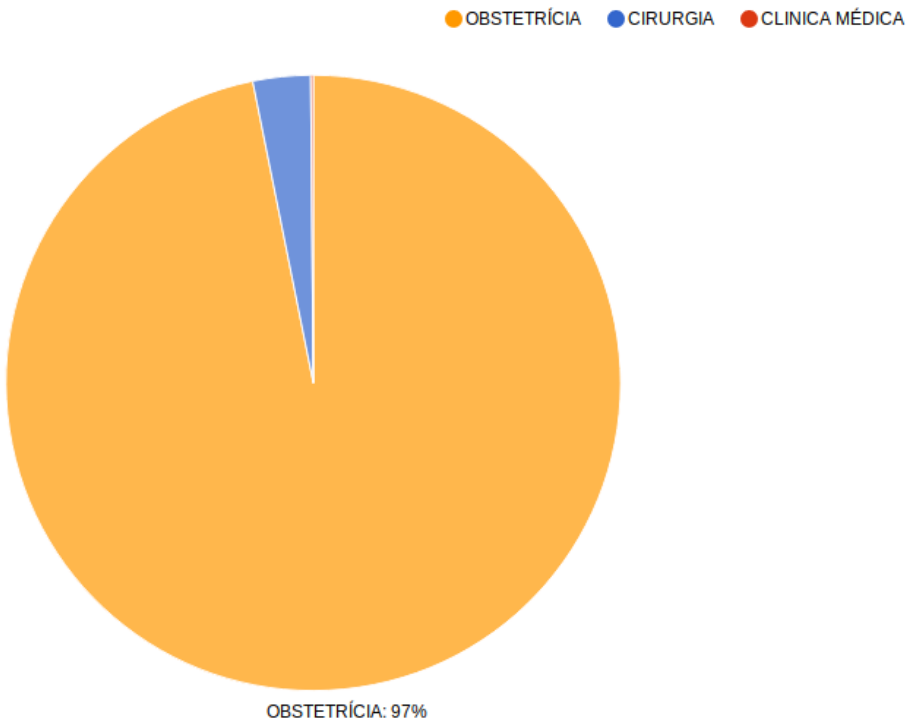


Figura 4.26: *Especialidades para o parto único espontâneo, não especificado.*

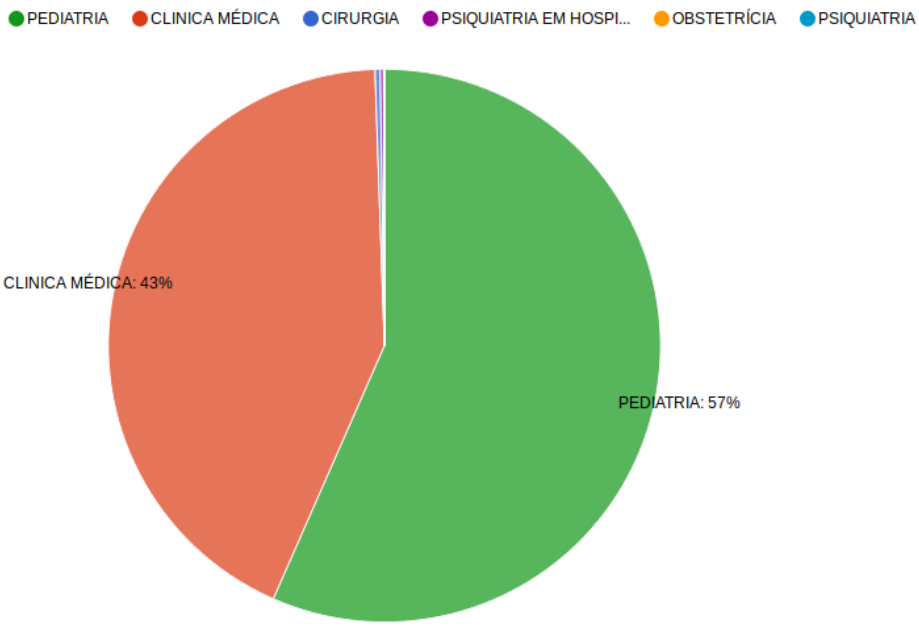


Figura 4.27: *Especialidades para a broncopneumonia não especificada.*

Analisando as Figuras 4.25, 4.26 e 4.27, como era de se esperar, como as duas primeiras tratam de doenças relacionadas ao parto, há uma grande predominância na especialidade obstetrícia. Já para a broncopneumonia não especificada, há uma maior quantidade de internações para a pediatria, o que indica uma maior quantidade de crianças com esse tipo de doenças. Porém, o que não era esperado era que a especialidade mais utilizada,

a cirurgia, como é visto na Figura 4.24, quase não aparece nas três doenças que mais causaram internações.

Capítulo 5

Considerações Finais

Para a modelagem do Data Warehouse foram utilizadas tabelas somente com as informações necessárias e com nomes de campos para fácil entendimento, de forma a evitarmos confusões com as análises e carregamentos de dados. Entre as principais características das bases que tivemos acesso para este trabalho está o tamanho e a diversidade. Em particular, a menor das bases tinha aproximadamente 180000 linhas e 54 colunas.

Nossa modelagem ajudou a agilizar as cargas dos dados, já que eles foram divididos em diversas tabelas com menos colunas em cada [1]. A população que demorou mais tempo foi a da tabela fato `internacao`, mostrada na Figura 3.8, que levou cerca de 40 segundos, o que é bem rápido considerando que o banco do SIH é o maior de todos com cerca de 700000 linhas e 20 colunas. Somado a isso, a montagem de cubos através do modelo de Data Warehouse fez com que as pesquisas ocorressem de forma mais rápida, já que eles já estava pré-calculados, diferente do que seria se toda pesquisa que fosse realizada precisasse realizar um "join" entre as tabelas fato e suas dimensões.

Durante o trabalho com as bases do SUS, foi possível observar que há uma grande irregularidade nos dados que foram extraidos, como exemplo, nas três bases que trabalhamos o **SIM**, **SINASC** e **SIH**. As colunas do CSV eram diferentes das do dicionário de dados, disponibilizados no TABNET [20], dificultando a população do Data Warehouse. Os dados que mais trouxeram problemas foram os campos de data e os campos que possuiam muitos espaços em branco à direita. Para contornar o problema da data foi feito um script em AWK, que é uma linguagem de programação interpretada, geralmente, usada para deixar os scripts de shell mais poderosos. Isso foi utilizado para alterar as datas do tipo "AAAAMMDD"para "DD-MM-AAAA". E para os espaços em branco foi necessário um trabalho bem mais manual para a sua remoção, através de um editor de texto substituir os espaços por um vazio. Esse foi o processo de limpeza de dados realizado para popular o Data Warehouse e realizar as visualizações e análises dos dados.

A utilização da ferramenta **Superset**, por mais que ela não fosse específica para Data Warehouse, a sua interface para criação de gráficos é muito simples de entender, e muito customizável. Por meio do SQL Lab, é possível editar a pesquisa que está sendo realizada

[1] O computador utilizado possui um processador Intel(R) Core(TM) i7-6560U CPU @ 2.20GHz, 16 GB de memória RAM e o sistema operacional é um Ubuntu 18.04.3 LTS.

alterando a query do SQL do gráfico que está sendo montado e como é observado na Figura 5.1, na coluna da esquerda pode-se visualizar os atributos da tabela selecionada "intern_espec_mor_proc_car_cid"e na direita é onde a query é escrita em que ela pode ser executada e o resultado aparece na aba “Results” logo abaixo, e se o resultado for o esperado basta clicar no botão “Explore” que irá direto para montagem dos gráficos com a query nova. Também é possível editar um arquivo JSON no dashboard, para alterar certas propriedades, como por exemplo na Figura 5.2, se alterar o campo “filter_immune_slices” para “filter_immune_slices”: [324, 65, 92], deixa os slices 324, 65 e 92 imunes a qualquer tipo filtro como por exemplo o filtro da Figura 4.9. Através dessa “filter box” foi possível realizar as operações “slice” e “dice” do cubo.

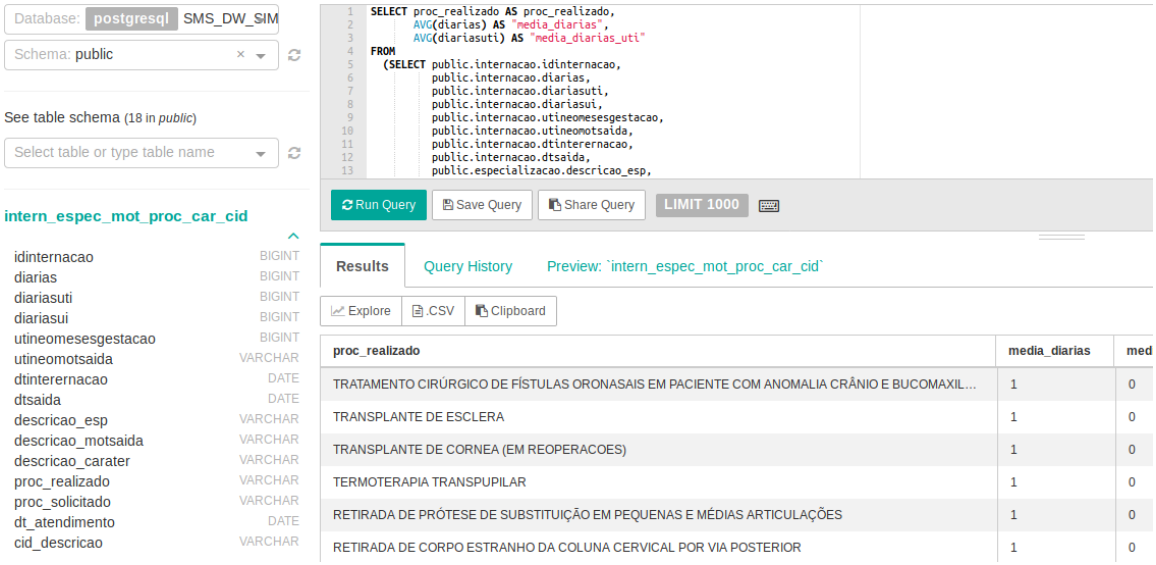


Figura 5.1: *Exemplo do SQL Lab.*

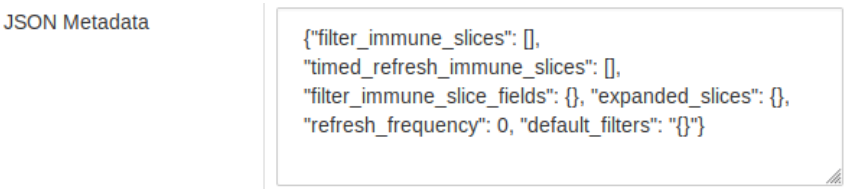


Figura 5.2: *Edição do JSON do dashboard.*

Por fim, neste trabalho foi observado que é possível criar um modelo de Data Warehouse nas bases do SUS, unificando-as de forma lógica e organizada, de modo que qualquer pessoa possa entender o que cada campo da tabela significa. Também, foi apresentada uma ferramenta o **Superset**, que pode manusear esse modelo, que é simples para montar e editar os gráficos, além de fornecer uma dashboard de fácil utilização e de compartilhamento, como foi mostrado na Figura 2.2, porém antes de realizar a modelagem seria necessário fazer uma limpeza dos dados que são extraídos da matriz.

Como trabalho futuro, o próximo passo é utilizar um processo de extração e carga de

dados completo, como seria em um Data Warehouse[2]. Cada uma dessas cargas teria um id único e uma data periódica em que ela seria carregada no sistema, simulando como seria o uso do modelo no dia-a-dia. Além disso, poderia ser feito gradualmente um aumento do modelo integrando cada vez mais bancos. Sendo que o primeiro deles, por exemplo, o SIGA (Sistema Integrado de Gestão de Assistência à Saúde de São Paulo), que foi proposto pela secretária como uma base importante, junto com o SIM, SIH e SINASC.

[2]Um processo foi desenvolvido no TCC do aluno Marcos Vinicius do Carmo Sousa, também do IME-USP, que pode ser encontrado em seu GitHub https://github.com/mvsousa/mac499 ou na página de seu TCC https://linux.ime.usp.br/ marcksm/mac0499/.

Bibliografia

[1] Mohammad Alkhatib, Amir Talaei-Khoei e Amir Ghapanchi. "Analysis of Research in Healthcare Data Analytics". Em: (jun. de 2016).

[2] Wiki Data analysis. URL: https://en.wikipedia.org/wiki/Data_analysis.

[3] *Big Data*. https://www.gartner.com/en/information-technology/glossary/big-data.

[4] DATASUS. URL: http://datasus.saude.gov.br/datasus.

[5] Ramez Elmasri e Shamkant B. Navathe. *Fundamentals of Database Systems*. Pearson, 2010.

[6] Skotnes T Feldman B Martin EM. *Big Data in Healthcare Hype and Hope*. Dr. Bonnie 360, 2012.

[7] Sullivan Frost. *Drowning in Big Data? Reducing Information Technology Complexities and Costs For Healthcare Organizations*.

[8] guru99. URL: https://www.guru99.com/what-is-data-analysis.html.

[9] W. H. Inmon. *Building the Data Warehouse*. Wiley Computer, 2002.

[10] Isabel Cristina Italiano, João Eduardo Ferreira e Osvaldo Kotaro Takai. *Aspectos conceituais em data warehouse*. Rel. técn. Departamento de ciência da computação IME-USP, 2001.

[11] Luiz Diana de Oliveira. "Estratégia de e-saúde do brasil: Plano de ações do ministério da saúde". Em: *Apresentação realizada para a I Oficina On-line: Estratégia e-saúde para o Brasil, da Sociedade Brasileira de Informática em Saúde* (fev. de 2017).

[12] Wullianallur Raghupathi. "Data Mining in Healthcare". Em: abr. de 2010, pp. 211–224. ISBN: 978-1-4398-0978-5. DOI: 10.1201/9781439809792-c11.

[13] Wullianallur Raghupathi e Viju Raghupathi. "Big data analytics in healthcare: Promise and potential". Em: *Health Information Science and Systems* 2 (fev. de 2014), p. 3. DOI: 10.1186/2047-2501-2-3.

[14] Raghu Ramakrishnan e Johannes Gehrke. *Database Management Systems*. McGraw-Hill, 2002.

[15] SAS. URL: https://www.sas.com/pt_br/insights/big-data/what-is-big-data.html.

[16] SIH. URL: http://datasus.saude.gov.br/sistemas-e-aplicativos/hospitalares/sihsus.

[17] SIM. URL: http://www2.datasus.gov.br/DATASUS/index.php?area=060701.

[18] SINASC. URL: http://www2.datasus.gov.br/DATASUS/index.php?area=060702.

[19] Apache Superset. URL: https://superset.incubator.apache.org/#.

[20] TABNET. URL: http://www2.datasus.gov.br/DATASUS/index.php?area=0901.

[21] *To Lower Health Care Costs, Look to International Innovations*. https://fortune.com/2019/09/10/health-care-costs-international-innovation/.

[22] *Transforming Health Care Through Big Data*. Institute for Health Technology Transformation.

[23] Belle Xia e Peng Gong. “Review of business intelligence through data analysis”. Em: *Benchmarking: An International Journal* 21 (abr. de 2014), pp. 300–311. DOI: 10.1108/BIJ-08-2012-0050.