INTER SCIENTIA

Vol. 5 • Nº 1 • Ano 2017

PROJETO DE ANÁLISE DE DADOS PARA IMPLANTAÇÃO DE DATA MART COMO FERRAMENTA PARA TOMADA DE DECISÃO EM COMBATE AOS VÍRUS DA DENGUE, ZIKA E CHIKUNGUNYA

PROJECT OF DATA ANALYSIS FOR DEPLOYMENT OF A DATA MART AS TOOL FOR DECISION SUPPORT IN COMBAT AGAINST DENGUE, ZIKA AND CHIKUNGUNYA VIRUS

Recebido em: 01/10/2017.
Aceito em: 28/11/2017.

Nivaldo Mariano Carvalho[1]
David Galdêncio Ferreira[2]
Moisés Erickison Brito de Araújo[3]
Ricardo Roberto de Lima[4]

RESUMO

O objetivo deste trabalho é utilizar de ferramentas e técnicas de *Business Intelligence* (BI) para análise e mapeamento de bases de dados em geral, com ênfase na saúde pública, para que através desta análise possa-se obter um sistema de apoio à tomada de decisão a favor do combate a doenças epidêmicas de Dengue, Zika e Chikungunya na cidade de Recife. Através da análise das bases de dados, disponibilizadas publicamente pela Prefeitura de Recife, será implantado um *Data Mart* como repositório de dados e ferramenta de inteligência empresarial para apoiar soluções de prevenção e combate a focos do mosquito *Aedes Aegypti*. Com isso, este projeto encaminha-se com a consolidação de dados que se inicia com a migração de dados para sistemas informatizados, um *Data Mart*, tornando o histórico on-line e disponível para tomadas de decisão (SAD) por parte de órgãos e autoridades públicas e ser consultado por todos os envolvidos, cidadãos ou equipe da saúde e gestores do governo e de organizações de saúde privadas.

Palavras-chave: Mineração de dados. Saúde pública. *Aedes aegypti*. DeCS.

1 Pós-graduando em Business Intelligence com Big Data pelo Centro Universitário de João Pessoa (UNIPÊ). E-mail: nivaldo_j@outlook.com

2 Graduado em Gestão da Tecnologia da Informação pelo Centro Universitário de João Pessoa (UNIPÊ). E-mail: davidgf_jp@hotmail.com

3 Pós-graduando em Business Intelligence com Big Data pelo Centro Universitário de João Pessoa (UNIPÊ). E-mail: moises201197@gmail.com

4 Mestre em Engenharia de Software pelo CESAR/UFPE. Docente do Centro Universitário de João Pessoa (UNIPÊ). E-mail: ricardo.roberto@unipe.br

ABSTRACT

The goal of this document is to make use of tools and techniques of Business Intelligence (BI) for analysis and mapping of databases in general with emphasis on public health, a decision support system to combat epidemic diseases as Dengue, Zika and Chikungunya in Recife will be obtained through this analysis. Analyzing public databases available on Recife city hall, a Data Mart as data repository and tool of business intelligence will be implanted to support solutions of prevention and combat the focus of mosquito *Aedes Aegypti*. With this, this project deals with data consolidation that begins with the migration of data to computerized systems, a Data Mart, making an online historic that is available for a Decision Support System (DSS) by public authorities and consultable by all involved, citizens, health staff and government managers or private health organizations.

Keywords: Data mining. Public health. *Aedes aegypti*. DeCS.

INTRODUÇÃO

Os avanços tecnológicos permitiram a incorporação em tarefas de análise de dados uma nova concepção chamada de *Business Intelligence* (BI), que ganha mais força e visão no mercado sendo catalogada de Inteligência do Negócio, com ênfase no apoio a tomada de decisão de profissionais de nível tático e estratégicos, possibilitando caminhos para solução de problemas.

O uso de softwares com singularidades *Online Analytical Processing* (OLAP) permite a estratificação de registros e *dashboards* (painéis gráficos) com desempenho de alta qualidade e diferentes amostras matemáticas e da estatística. Segundo Ziulkoski (2003), conforme citado por Santo (2006), a Tecnologia da Informação (TI) é vista como uma ferramenta estratégica central na busca pela vantagem competitiva.

Na década de 90, surgiu uma poderosa ferramenta para gerenciar as informações dentro de uma organização, seja ela privada ou pública: o *Data Warehouse*/*Data Mart*, ou DW/DM, respectivamente. De acordo com Inmon (1996), um DW é uma coleção de dados integrados, orientados por assunto, não-voláteis e variáveis com relação ao tempo, de apoio às tomadas de decisão gerenciais (apud SANTO, 2006).

O uso da referida tecnologia pode facilitar as tomadas de decisão em relação a assuntos importantes, como por exemplo, a saúde pública. Os crescentes casos de Dengue, Zika e Chikungunya tomaram os jornais físicos e eletrônicos, redes sociais e meios de comunicação como um todo, principalmente a partir do segundo semestre de 2015, segundo a Secretaria Estadual de Saúde (SES) do estado de Pernambuco (PE), em 2016, no Brasil.

Foram, somente no ano de 2016, 16,46 mil (dezesseis mil, quatrocentos e

sessenta) confirmações de dengue, 2,644 (dois mil, seiscentos e quarenta e quatro) de Chikungunya e 75 (setenta e cinco) de Zika somente em Pernambuco. Segundo o boletim epidemiológico da SES-PE, publicado em 07 de junho de 2016, um dos municípios que apresentaram as maiores incidências de dengue foi em Ipojuca, na Região Metropolitana do Recife. Só no Recife, no ano de 2015, a capital registrou 34,46 (trinta e quatro mil, quatrocentos e sessenta) mil casos e em 2016, 19,18 (dezenove mil, cento e oitenta) mil prováveis das três doenças transmitidas pelo mosquito *Aedes Aegypti*.

Segundo Kimball (2002), o BI tem a responsabilidade para arcar com os resultados, pois foca nas necessidades do negócio, neste caso, na própria saúde pública. As informações podem possibilitar ao governo um rastreamento mais preciso dos focos da epidemia na cidade do Recife, e, assim, contribuir para que medidas preventivas nos locais mais atingidos, como limpeza de bairros, monitoramento mais agressivo de agentes de saúde, presença mais assídua quanto ao assunto da unidade básica de saúde na área etc., sejam traçadas de forma mais rápida.

Após realização de análises dos dados e de notícias vinculadas à mídia, percebeu-se um aumento dos casos das doenças. Segundo o Diário de Pernambuco, publicado no início de 2016, o número de casos de microcefalia, registrados no estado de Pernambuco, subiu para 1829 (um mil, oitocentos e vinte e nove) na semanal final de março do ano de 2016. De acordo com a Secretaria Estadual de Saúde de Pernambuco (SES), 728 (setecentos e vinte e oito) atendem os parâmetros da OMS (Organização Mundial de Saúde) para a microcefalia que é causada em bebês pelo vírus da Zika que é epidemicamente distribuído pelo mosquito da dengue (*Aedes Aegypti*), o equivalente a 39,8% dos casos. É importante ressaltar que esses dados consideram as notificações entre 1º de agosto de 2015 e 26 de março de 2016.

Justamente nessa lógica é que o BI se insere na gestão pública, pois torna viável a produção do conhecimento através destas próprias informações que são armazenadas em sistemas distintos e diversos. Mesmo assim ainda existem muitas barreiras que devem ser superadas. Se mesmo em organizações tradicionais já existem inúmeras dificuldades em se obter excelência no aproveitamento das informações para produção de conhecimento, nos órgãos públicos, então, tais dificuldades se multiplicam em decorrência de se estar falando na sociedade como um todo (BOTH, 2012).

De posse destes dados e fatos, disponibilizados pela própria prefeitura e por meios de comunicação, o BI garante com responsabilidade, contanto que seja usado da forma correta, o seguro controle destes índices, fazendo com que a partição pública responsável tenha um suporte a mais em suas mãos.

Com o objetivo de melhorar a saúde pública no que se relaciona com os vírus do *Aedes Aegypti*, foi proposta a implementação de um *Data Mart* como repositório de dados conhecidos publicamente pelo site da prefeitura da cidade do Recife, com a situação de epidemia dos vírus da Dengue, Zika e Chikungunya.

CONCEITOS DO DESENVOLVIMENTO DO TRABALHO

Business intelligence

HISTÓRICO E CONCEITO

Segundo Elena (2011 apud BOTELHO, 2014), o termo BI, quando relacionado a outros termos dentro da tecnologia da informação, é mais recente e sua utilização foi feita pela primeira vez na década de 1950 por Hans Peter Luhn, o qual era pesquisador da IBM, em seu artigo intitulado de "*A Business Intelligence System*". Em seu artigo, Hans Peter Luhn utilizou do termo para referenciar o desenvolvimento de um sistema automático, baseado em máquinas de processamento de dados, que indexa e codifica automaticamente documentos e dissemina informações nas organizações conforme o ponto de ação (BOTELHO, 2014).

De acordo com Luhn (1958 apud BOTELHO, 2014), a "comunicação eficiente é uma chave para o progresso em todos os campos do esforço humano". O que Luhn objetiva explicar é que a comunicação é necessária, seja ela de qualquer sentido ou qualquer meio e passada de forma correta e eficiente, o ser humano consiga se orientar na informação dada pela comunicação e utilizá-la para entender o mundo ao seu redor e basear seus objetivos, sejam eles profissionais ou pessoais.

Segundo o histórico, na época de Luhn não existiam métodos para que a comunicação atingisse os objetivos das organizações, além do que a divisão e especialização das funções criavam novas barreiras para o fluxo de informação. De acordo com Luhn (1958), conforme citado por Botelho (2014), as empresas tinham dificuldades em gerir a informação de forma correta e, para isso, Luhn sugeriu o seu trabalho que consistia em um sistema de inteligência de negócios que aborda a coleta ou aquisição de novas informações; disseminação; armazenamento; recuperação; e, transmissão de informações.

A partir da década de 1980, impulsionados pela evolução dos *personals computers* e aumento da capacidade dos componentes que realizam processamento nestes computadores, os sistemas de BI tiveram maior desenvolvimento (ELENA, 2011; VERCELLIS, 2009 apud BOTELHO, 2014). Segundo Barbieri (2011), nessa época, os dados começaram a ganhar destaque, os estudos na área da tecnologia da informação avançaram e possibilitaram o surgimento das disciplinas de administração de dados, modelagem de dados, engenharia da informação e a análise de dados (apud BOTELHO, 2014).

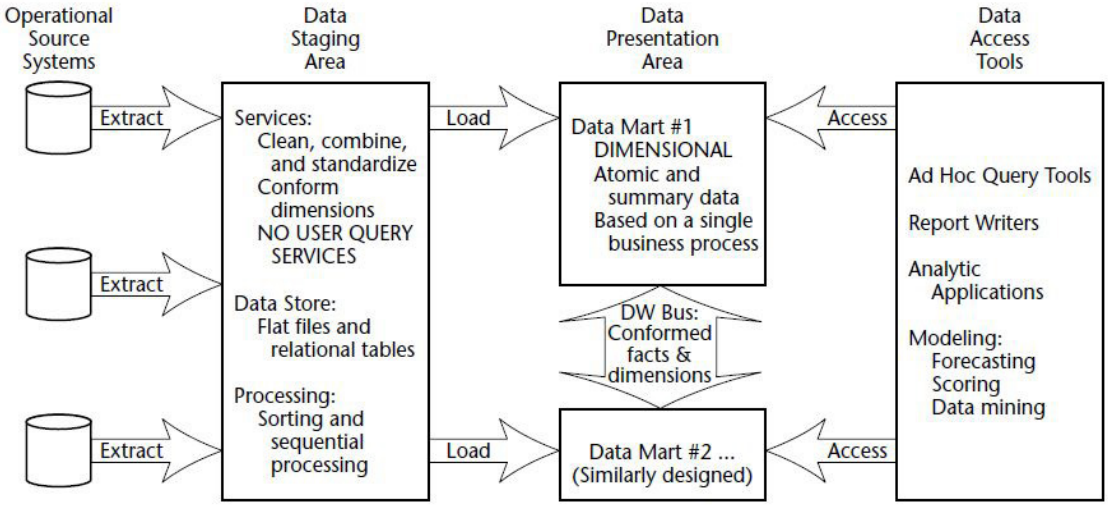
Mesmo com toda a literatura envolvida no *Business Intelligence*, ainda não se pode tornar um conceito concreto; é possível identificar termos e objetivos em comum nas definições dos autores, porém não há regra clara para se definir. Além disso, "é difícil compreender totalmente a BI, porque seus aplicativos não são sistemas

autônomos, nem dão suporte a objetivos específicos, como outros sistemas (SCM, CRM etc.)" (TURBAN; VOLONIMO, 2013 apud BOTELHO, 2014).

DATA WAREHOUSE E DATA MART

Para entender melhor o conceito de modelagem dimensional, se torna necessário compreender duas estruturas de DW, uma segundo Kimball (2002) e outra segundo Inmon (2001). Para Kimball, o processo de *Business Intelligence* começa extraindo os dados de sistemas operacionais onde estes são armazenados, posteriormente esses dados são jogados numa área temporária para tratamento chamada *Staging Area* e depois armazenados em suas respectivas divisões chamadas de *Data Mart* compondo o *Data Warehouse* em si; após isso, os dados podem ser acessados através de ferramentas BI. Este processo pode ser visto na Figura 1.

Figura 1: *Business Intelligence* segundo Kimball



Fonte: HENRIQUE (2012)

De acordo com HENRIQUE (2012), os componentes relevantes deste processo são:

- *Staging Area*:
  - Parte do *Data Warehouse* responsável por receber a extração, transformação e carga (ETL) das informações dos sistemas transacionais legados, para posterior geração dos *Data Marts* de destino;
  - *Staging Area* é considerada área fora do acesso dos usuários, por isso não deve suportar consultas dos Usuários;
  - Ela pode ser composta por arquivos textos ou tabelas de banco de dados normalizadas.

- *Data Presentation Area*:
    - Área responsável pela apresentação dos dados, não deve ser utilizada para limpeza ou transformação de dados.
    - Organizada em *Data Marts*, orientados a processos de negócios, e não a unidades de negócio, departamentos ou funções específicas.
    - Um *Data Mart* é composto por dados atômicos e dados sumarizados para uma melhor performance.

A produção de um DW tem um custo altíssimo e leva certo tempo para a sua construção. As características de um DW dependem do tamanho da organização, do número de bases de dados que compõem o projeto, número de pessoas que compõe a equipe, as ferramentas a serem utilizadas e dentre outros fatores.

Nesse contexto, os arquitetos decidiram segmentar a construção de um DW iniciando por departamentos. Esses começaram a construção pelos DMs para os mesmos alimentarem o DW, ao invés da forma tradicional, que é construir um DW primeiro e a partir dele os DMs. Eles perceberam que essa abordagem tem grandes vantagens, sendo a principal delas o tempo de implementação reduzido. Um DM pode surgir de duas maneiras:

- **Top-Down:** É quando a empresa constrói um DW e depois passa a segmenta-lo, dividindo-o em partes menores e criando DM orientados a assuntos.
- **Bottom-up:** É a construção quando a organização decide primeiro criar um banco de dados apenas para uma área, e depois de acordo com os resultados a empresa vai criando para outras áreas até resultar em um DW.

A infraestrutura de hardware e software tem perfis semelhantes, porém a arquitetura entre as duas são diferentes. De acordo com Henrique (2012), o DW, por ser um modelo hibrido possa também ser um modelo relacional, enquanto o DM passa por um conceito dimensional.

A dificuldade do tratamento dos dados em um DW em uma empresa é maior do que um DM, por exemplo, pois com um DM a preocupação é com apenas uma parte da empresa, enquanto que em um DW a organização deve se preocupar com toda a organização na fase de implementação.

Kimball (2002) defende que um DW deve ser planejado de vários DMs, para depois serem integrados em um único DW. Em sua análise, Kimball argumenta que as empresas devem construir DMs por assuntos, para posteriormente ter várias conexões entre eles, nas quais seriam as tabelas Fato e Dimensão, na qual as informações criadas entre os diferentes DMs poderiam ser geradas de maneira segura.

Bill Inmon (2001) tem como teoria que deve ser construído primeiro o DW, analisando toda a organização para um modelo único, para após isso ser produzido os

DMs por assuntos ou departamentos.

EXTRACT, TRANSFORM AND LOAD (ETL)

O processo de *Extract, Transform and Load* (ETL) tem como objetivo a extração, transformação e carga dos dados de bases e fontes externas de dados para uma ou várias bases que ocupam o DW. Este método se torna obrigatório para o processo, sendo a transformação ou limpeza um subprocesso opcional.

O ETL é o processo mais complexo e crítico, além de demorado na construção, de um DW/DM, porque se baseia na extração dos dados de bases não homogêneas, na transformação e limpeza destes dados e na carga dos dados na base do DW ou DM. Como dito, as decisões gerenciais são tomadas com base nas informações geradas pelas ferramentas do tipo *front-end* (*dashboards*, por exemplo). Estas informações são geradas através dos dados que estão localizados e armazenados no DW/DM. Se estes dados não estiverem íntegros e não forem trabalhados de forma correta no processo de ETL, as informações que foram geradas através deles farão com que decisões sejam tomadas erroneamente, podendo afetar diretamente os negócios da organização (ABREU, 2008).

Dessa forma, os dados devem representar a verdade, a mais pura verdade, nada mais que a verdade (KIMBALL, 1998 apud ABREU, 2008). Segundo Inmon, a maior parte do esforço exigido no desenvolvimento de uma solução BI atrelada a um DW/DM é consumido neste momento e não é incomum que oitenta por cento de todo o esforço seja empregado na construção do processo e da ferramenta de ETL (INMON, 1997 apud ABREU, 2008).

Segundo Kimball (1998 apud ABREU, 2008), as características mais relevantes para garantir a qualidade dos dados são:

- Unicidade, evitando assim duplicações de informação;
- Precisão, pois os dados não podem perder suas características originais assim que são carregados para o DW;
- Completude, não gerando dados parciais de todo o conjunto relevantes às análises; e
- Consistências, ou seja, os fatos devem apresentar consistências com as dimensões que o compõem.

Segundo Abreu (2008), é necessário que os dados fiquem homogêneos para serem carregados no DW/DM e para isso a construção da ferramenta de ETL precisa ser feita de maneira correta e com isso garantir a qualidade dos dados.

Modelagem dimensional

O modelo dimensional é uma forma de modelagem que as informações podem ser representadas na forma de um cubo. Esse modelo permite visualizar dados abstratos de forma simples e confrontar dados de diferentes setores de uma organização, que muitas vezes pode ser muito complicado de ser analisado usualmente em um modelo de dados relacional (MOREIRA, 2006).

A grande vantagem do DW, e o que o torna tão eficiente é que as informações que estão em vários sistemas ou formatos de arquivo são convergidas em um banco de dados de forma dimensional, obtendo assim informações unificadas e padronizadas sobre a organização em um mesmo local.

Segundo Moreira (2006), no exemplo de uma empresa na qual possui várias filiais e deseja acompanhar o desempenho de suas vendas para uma tomada de decisão, por exemplo, um DW poderia ser descrito em três dimensões principais: Produto, na qual representa o produto vendido na filial, Loja, que é a filial da empresa na qual vendeu o produto e Tempo, que corresponde as informações de data de quando a venda foi concretizada.

Um modelo dimensional pode ter quantas dimensões forem necessárias, e, dependendo de quantas dimensões são, pode não ser possível de ser representadas graficamente. Mesmo não sendo possível de ser caracterizadas, essas dimensões podem ser detalhadas e pode ser possível acompanhar o desempenho dessas dimensões (MOREIRA, 2006).

Segundo Moreira (2006), o modelo dimensional conta basicamente com uma tabela de fatos central e tabelas dimensionais ligadas diretamente a essa tabela de fatos. A modelagem deve ser feita de modo a permitir velocidade de acesso a uma informação, proporcionando que os softwares naveguem por esses bancos de dados com eficiência.

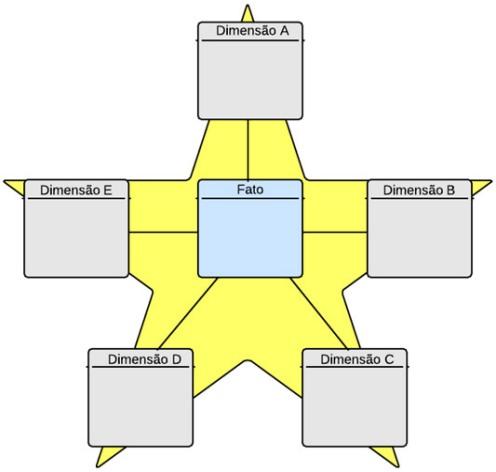
Uma tabela de Fatos contém medições sobre o negócio, tais como quantidade de produtos vendidos, valor de uma venda ou valor unitário de um produto, bem como chaves para as tabelas de dimensões (MOREIRA, 2006). Uma tabela de fatos pode-se tornar extremamente grande de acordo com a quantidade de registros armazenados.

As tabelas de dimensões contêm descrições sobre os dados que fazem parte do DM/DW. Essas tabelas possuem vários atributos que descrevem em detalhes as características que possam definir e serem úteis para pesquisas no DW (MOREIRA, 2006).

STAR SCHEMA (MODELO ESTRELA)

No modelo estrela, no qual está sendo utilizado nesse projeto, todas as tabelas têm relações diretas com a tabela de fatos. Nesse modelo, as tabelas dimensionais devem conter todas as informações que são necessárias para a definição de classes como, por exemplo, Produto, Tempo ou Loja na mesma tabela (MOREIRA, 2006).

Figura 2: Star Schema



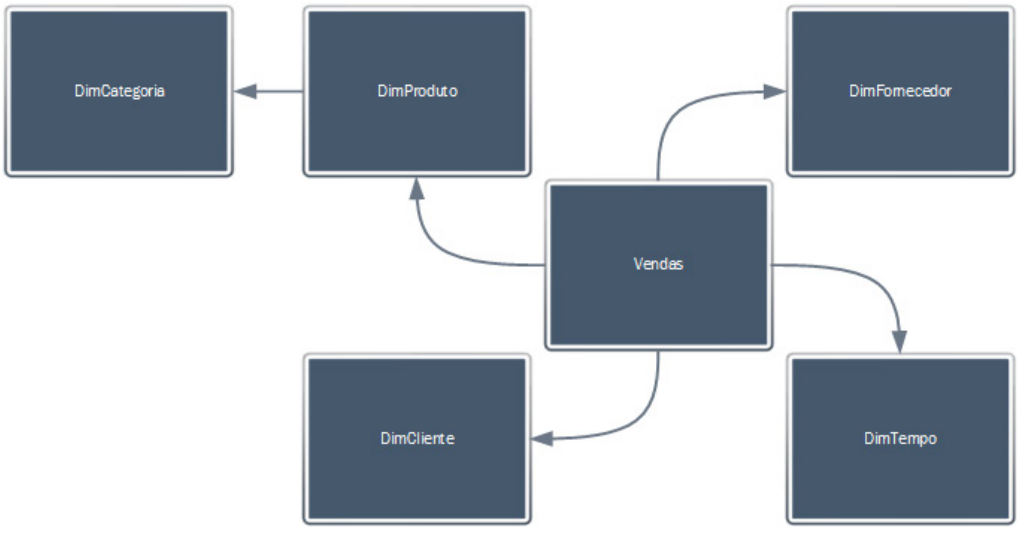
Fonte: ELIAS (2014)

Esse modelo é chamado de Estrela porque a tabela de fatos fica ao centro cercada das tabelas dimensionais assemelhando a uma estrela e as dimensões desse modelo não são normalizadas (CRUZ, 2014). A Figura 2 demonstra o *star schema* e como as informações integram a tabela fato.

SNOWFLAKE (MODELO FLOCO DE NEVE)

Segundo Moreira (2006), no modelo floco de neve (*snowflake*) as tabelas dimensionais se relacionam com a tabela de fatos, porém algumas dessas dimensões relacionam apenas com elas mesmas. Isso acontece para a normalização das tabelas dimensionais, visando a diminuição do espaço ocupado por essas tabelas, assim as informações tornam-se tabelas de dimensões auxiliares (BRITO, 2015).

Figura 3: Modelo Snowflake



Fonte: PEREIRA (2014)

Se construído o banco de dados dessa forma, é utilizado mais tabelas para representar as mesmas dimensões, mas ocupando um espaço em disco menor que o modelo estrela (MOREIRA, 2006). Esse modelo não foi utilizado em nosso projeto por adicionar complexidade desnecessária para a construção de um DW. O modelo floco de neve é chamado assim por as dimensões se dividirem em outras tabelas não normalizadas, lembrando assim um floco de neve (BRAGA, 2015). A Figura 3 exibe o modelo *snowflake* e mostra que este possibilita a relação de dimensões com dimensões.

METODOLOGIA DE DESENVOLVIMENTO

Para este trabalho ser executado foi necessário, além da pesquisa sobre BI, a coleta dos dados dos vírus disseminados pelo mosquito *Aedes Aegypti*, depois disso a modelagem do DM, a utilização da ferramenta de ETL *Pentaho Data Integration*® e criação de dashboards analíticos com o Power BI® da Microsoft® e, por fim, a organização dos dados já transformados em informações e monitoramento com estudo destes dados para ser implantado o DM. As bases de dados aqui dispostos foram disponibilizados publicamente pelo site da Prefeitura de Recife, em sua plataforma de dados abertos. Também foram utilizados, como base, os princípios de modelagem dimensional de autoria de Ralph Kimball.

Após o entendimento do problema, a abordagem realizada por este trabalho propôs as seguintes soluções, como pode ser visto no Quadro 1:

Quadro 1: Esquema de Justificativa Geral

Problema Identificado	Solução Proposta	Conceito e Tecnologia
Alto volume de dados que precisam de uma organização e análise	Construir um repositório de informações integradas e de nível tático e estratégico	Data Mart
Crescente aumento de casos relacionados ao arbovírus do mosquito Aedes Aegypti na cidade de Recife que precisam de monitoramento	Definição e implantação de um processo de Inteligência de Negócio para a entidade pública do Recife	Business Intelligence
Ausência de tecnologia de análise dimensional das bases de dados	Utilização da Tecnologia de processamento analítico de dados	OLAP (Pentaho® e PowerBI®)

Fonte: Elaborado pelo Autor.

Em primeiro lugar, para se elaborar o repositório foi realizada a coleta dos dados. Estes últimos estão expostos na plataforma de dados abertos da Prefeitura de Recife e foram disponibilizados publicamente em formatos *csv* e *json*.

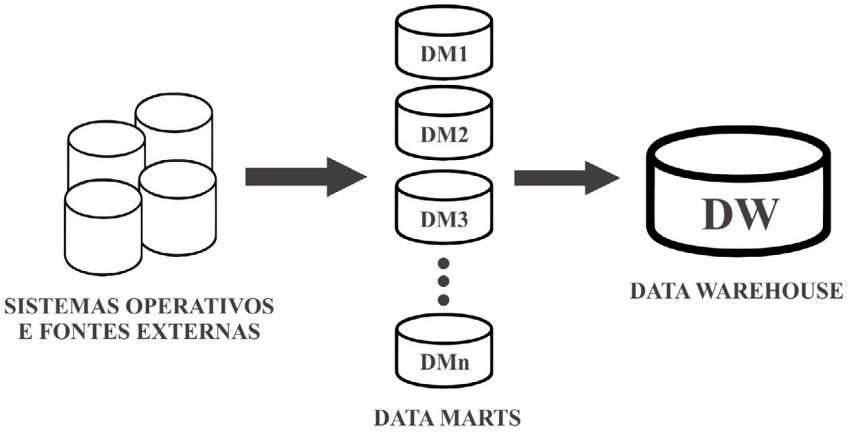
Para criação das ferramentas de coleta, ETL, organização (*Data Mart*), análise, compartilhamento e monitoramento das informações foi utilizado o software *open-source* chamado *Pentaho Data Integration*®. Foi utilizada, para a criação dos painéis de monitoramento, a ferramenta *PowerBI*®.

Em nível de estratégia de construção, foi utilizada a abordagem defendida por Ralph Kimball (2002) em seu livro "*The Data Warehouse Toolkit*" que é a abordagem bottom-up. Nesta abordagem, os DMs são criados primeiramente para prover

relatórios e analíticos capazes de especificar os processos do negócio, diferentemente da abordagem top-down na qual o DW é criado para depois os DM serem criados. Estes DMs, criados na abordagem bottom-up, são então integrados para a criação compreensiva e lógica do DW logo em seguida.

Foi escolhida a abordagem bottom-up, pois o nível de detalhes dos dados não é grande o suficiente para se ter uma visão mais consistente do DW e porque criar um DW é uma iniciativa cara, complexa e relativamente demorada (INMON, 2001). Uma ilustração desta abordagem *bottom-up* é apresentada na Figura 4:

Figura 4: Abordagem bottom-up para construção de Data Warehouse



Fonte: Elaborado pelo Autor

Para dar início ao desenvolvimento do sistema, foi necessário ter sob o domínio da equipe as bases de dados que seriam precisas para o caso. Estas bases são sobre os casos de dengue, zika e chikungunya e unidades de saúde.

As bases de dados foram obtidas através do site de dados abertos da Prefeitura de Recife. É importante frisar que estas bases estavam totalmente desnormalizadas e despadronizadas, com valores nulos, valores incoerentes e incorretos. Tais dados foram envolvidos na etapa de transformação para serem tratados e se tornarem úteis ao uso.

O sistema realizou a coleta dos dados direto da base da prefeitura de Recife por meio de *web service*. Depois de extraídos os dados, o *job* de extração que foi arquitetado na ferramenta *Pentaho* preencheu a *Staging Area* (SA) com estes dados, os quais ainda estavam em sua forma bruta.

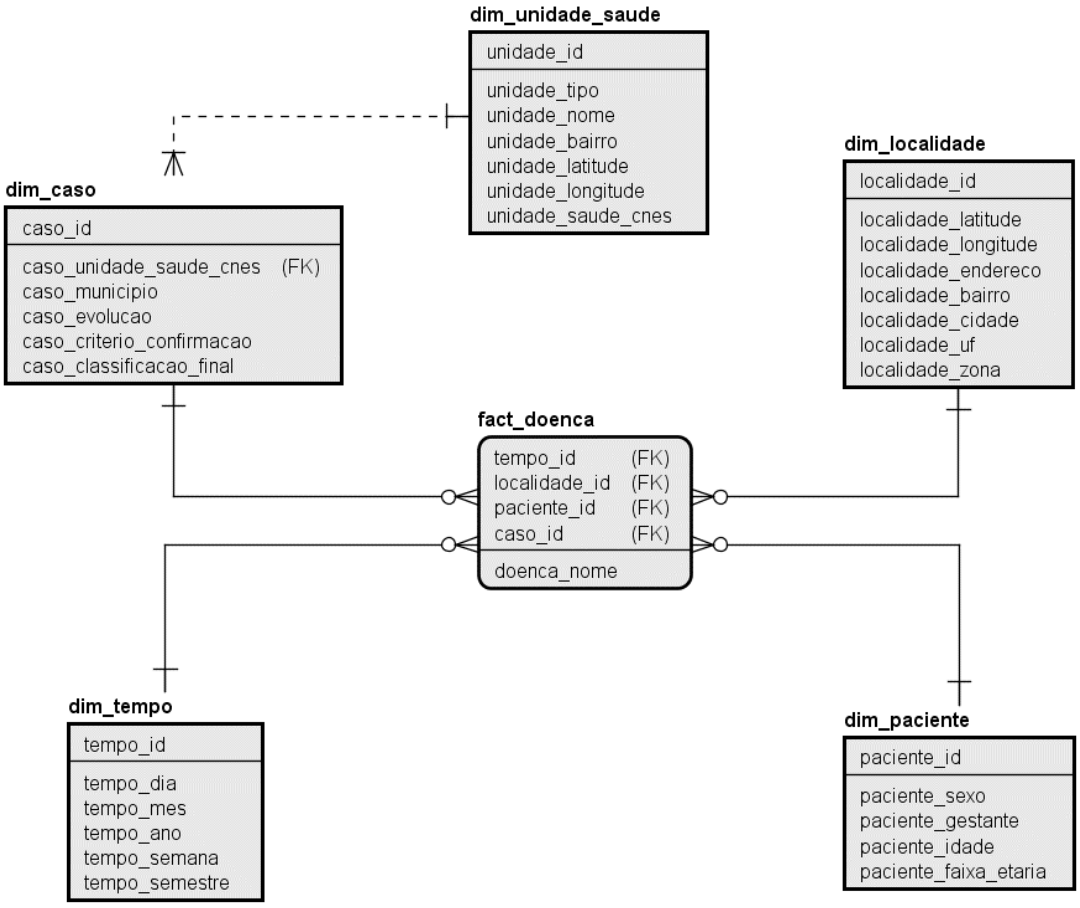
Após o preenchimento da SA, o *job* de transformação coletou os dados da própria SA e tratou-os conforme arquitetado, retirando do meio dos dados úteis as incoerências e erros, padronizando os mesmos e permitindo o uso nas ferramentas do OLAP. Com isso, o *job* de carga preencheu o DM em si com os dados já tratados e prontos para o uso analítico (*dashboards*).

Após a transformação e tratamento dos dados na SA, os dados estavam prontos

para serem utilizados nas ferramentas de análise. Para a parte de análise OLAP, foi necessário antes que os dados fossem carregados em um DM devidamente modelado.

Primeiro, antes de iniciar o processo de carga dos dados, foi necessário a definição de uma fonte de dados onde os mesmos seriam armazenados. Este projeto utilizou o armazenamento em nuvem, no qua os dados foram armazenados em um banco de dados SQL na nuvem.

Figura 5: Modelo estrela



Fonte: Próprio autor (2017).

O relacionamento correto entre os dados foi implementado, utilizando o modelo estrela (*star schema*), para que os dados enviados ao banco de dados SQL estivessem com sua modelagem correta, facilitando a criação dos *dashboards*, conforme visto na Figura 5.

RESULTADOS E DISCUSSÃO

Antes de se iniciar, de fato, a construção dos *dashboards* foi necessária a conexão do Power BI® com o *Data Mart* que se encontra na nuvem. Após feita toda a conexão com o *Data Mart* na nuvem, o Power BI® esteve com os dados presentes na nuvem

sincronizados em sua interface. A partir disto, pode-se, então, começar a modelagem em si dos *dashboards*. O Quadro 2 garante melhor entendimento do conteúdo das páginas dos mesmos.

Quadro 2: Explicação das páginas do dashboard.

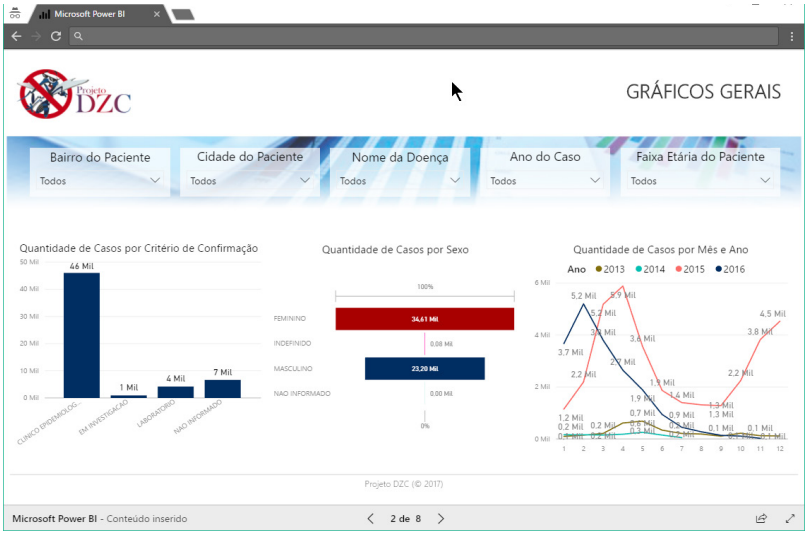
Página do Dashboard	Conteúdo da Página
1	Apresentação do Projeto
2	Gráficos Gerais (pizza, linhas, barras...)
3	TreeMap – Casos por Bairro
4	Mapa geográfico dos casos
5	Mapa geográfico das unidades de saúde
6	Diagrama de Sankey – Evolução/Faixa Etária
7	Gráfico SandDance – Class. Final do Caso
8	Sobre o Projeto

Fonte: Elaborado pelo Autor

Para este trabalho, foram feitas 8 (oito) páginas, sendo a primeira e a últimas páginas de apresentação e sobre o projeto, respectivamente. As restantes, são páginas que têm os gráficos de pizza, linhas, mapas, diagramas de Sankey, diagrama de *SandDance*, mapa de árvore (*TreeMap*), KPI's (*Key Perfomance Indicators* – Indicadores Chave de Performance), indicadores de número, contadores, entre outras visualizações oferecidas pelo Power BI®. Nas figuras 6 a 9 estão os gráficos gerais, *TreeMap* de casos por bairro, mapa geográfico dos casos, mapa das unidades de saúde, respectivamente.

Nos *dashboards*, a página 1 contém as informações sobre o projeto, com informações de como utilizar o painel, e na página 8 há as informações sobre os integrantes do projeto. Os *dashboards* que têm os dados vão da página 2 a página 7, contendo as seguintes informações:

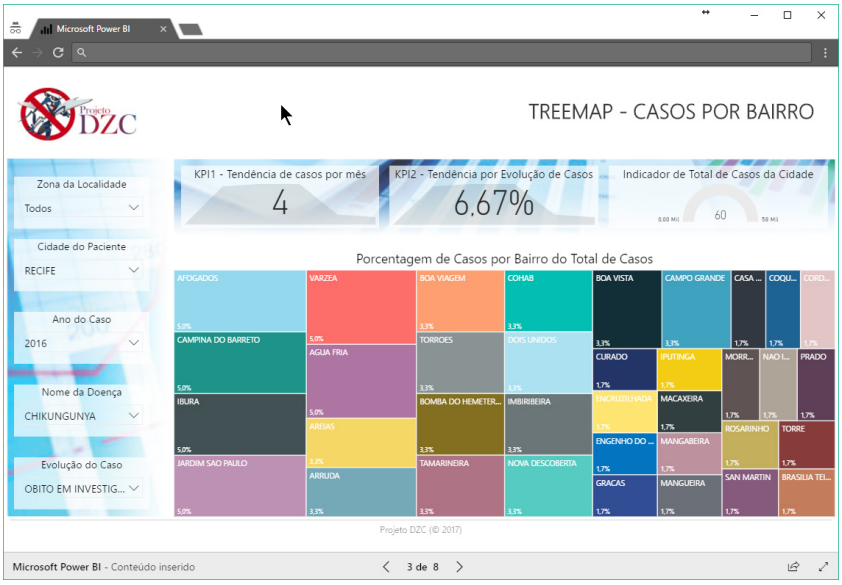
Figura 6: *Dashboard* de Gráficos Gerais



Fonte: Elaborado pelo Autor

Gráficos gerais – Nesse painel, é possível visualizar um gráfico de barras verticais que indicam a quantidade de casos por critério de confirmação clínico, um gráfico de funil que exibe quantidade de casos por sexo e um gráfico de linhas que exibem quantidade de casos por mês e ano. Este painel tem ainda filtros que permitem a escolha do bairro, cidade, doença, ano e faixa etária do paciente;

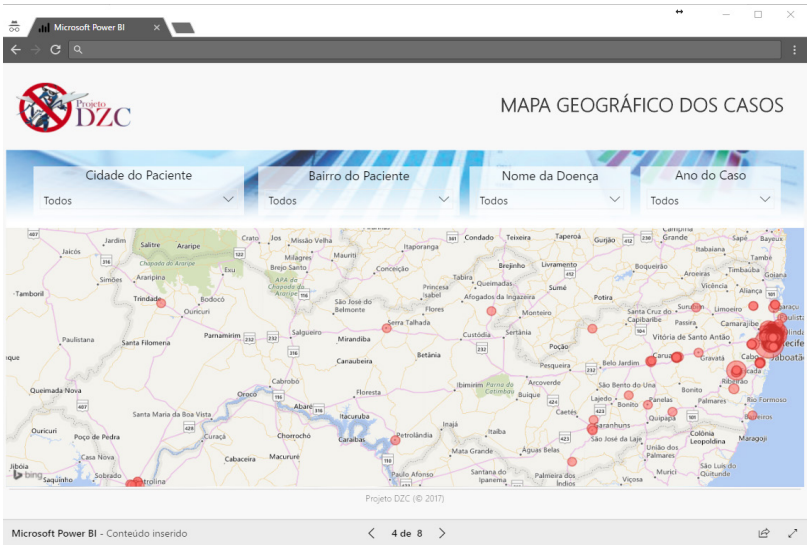
Figura 7: Dashboard de Casos por Bairro



Fonte: Elaborado pelo Autor

TreeMap – Painel que mostra a porcentagem de casos por bairro do total de casos acontecidos em um retângulo dividido com vários outros retângulos coloridos que representam municípios da região metropolitana de Recife, à medida que a porcentagem cresce para um município, este ganha uma área maior dentro do retângulo;

Figura 8: Dashboard de Mapa Geográfico dos Casos

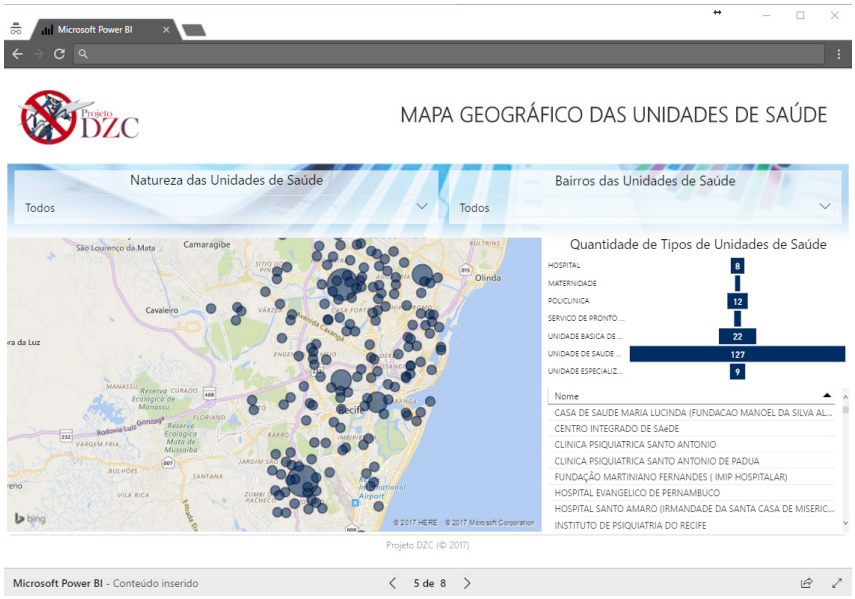


Fonte: Próprio Autor (2017).

Mapa Geográfico dos Casos – Exibe os casos de acordo com a seleção dos filtros dispostos no mapa com visão de satélite de acordo com latitude e longitude;

Mapa Geográfico das Unidades de Saúde – Semelhante ao anterior, exibe as unidades de saúde dispostas num mapa para, analisando o mapa anterior, indicar os casos que estão próximos às unidades de saúde;

Figura 9: Dashboard de Mapa Geográfico das Unidades de Saúde



Fonte: Próprio Autor (2017).

No *dashboard* também existem as páginas de Diagrama de Sankey e Gráfico *SandDance*. A página do diagrama de Sankey faz a ligação de casos baseados nas seleções de filtro relacionando os mesmos entre evolução do caso com a faixa etária e cidades e bairros, e o *SandDance* é um gráfico interativo que pode ser manipulado conforme seleção de filtros e à medida que um conjunto de dados é selecionado no gráfico dinamicamente é alimentado um relatório contendo detalhes sobre aquele dado.

CONSIDERAÇÕES FINAIS

Quando se aferem os *dashboards*, mais precisamente na página 3 (três) em que se localiza o *TreeMap*, quando selecionado apenas a cidade de Recife para a chikungunya nos anos de 2015 e 2016, em análise feita pela equipe em maio de 2017, tem-se que os bairros onde ocorreram maiores concentrações de casos foram os bairros de Ibura com 10,1 % dos casos e Afogados com 8,2 % dos casos para 2015 e Iputinga com 7,6 % dos casos e Várzea com 7,5 % dos casos para 2016, respectivamente.

Já quando é selecionado a zika para Recife nos anos de 2015 e 2016, tem-se que os bairros com maiores incidências foram Várzea com 15,6 % dos casos e Iputinga com 12,5 % dos casos para 2015; para 2016, foram os bairros Várzea com 16,9 % e Ibura com 15,5 % os que tiveram maiores incidências.

Para a dengue, ainda em Recife nos anos de 2013, 2014, 2015 e 2016, tem-se os seguintes bairros que obtiveram maiores incidências: para 2013, foram os bairros de Cohab e Ibura com 7,6 % e 7,0 % dos casos; para 2014, foram os bairros de Ibura e Boa Viagem com 6,6 % e 4,8 % dos casos; para 2015, foram os bairros de Cohab e Ibura com 6,8 % e 5,5 % dos casos; para 2016, foram os bairros de Várzea e Iputinga com 6,8 % e 5,9 % dos casos, respectivamente. É importante salientar que os outros bairros da cidade de Recife possuem porcentagens quebradas e menores divididas entre si.

A página do *TreeMap* também mostra um indicador chave de performance que indica, segundo estes filtros selecionados, a tendência de casos que, baseada nos dados reais da base, podem acontecer por mês de cada ano a que a base for atualizada.

Para a zika, o indicador informa que em 2015, a cada mês, havia uma tendência de 25 casos acontecerem; para 2016 esta tendência diminuiu para 3 casos. Para a chikungunya, o indicador informa que em 2015, a cada mês, havia uma tendência de 654 casos acontecerem em toda a Recife; em 2016 esta tendência diminui para 55 casos. Para a dengue, o indicador informa que no ano de 2013, a cada mês, 113 casos aconteceriam; em 2014, 52 casos; em 2015, 3642 casos; em 2016, 30 casos.

Se realizar um somatório, nos anos de 2015 e 2016, foram ao todo 103 casos notificados de zika e 3692 casos notificados de chikungunya. Para a dengue, de 2013 a 2016, foram 51,70 mil casos notificados.

É, também, importante frisar que a dengue têm uma abrangência anual maior (2013 a 2016) que as outras doenças, devido a poucos dados notificados, relembrando aqueles que estavam inconsistentes e desestruturados.

Há também um segundo indicador que mostra uma tendência, a depender do filtro de evolução do caso (cura, óbito, não informado, etc.) selecionado, de quantos por cento a evolução selecionada no filtro pode ocorrer. Quando aplicada a zika, nos anos de 2015 e 2016, a chance de cura era de 25 %, segundo os dados; para a chikungunya, nos anos de 2015 e 2016, a chance de cura era de 17,20 %.

Baseado nos trabalhos de Ralph Kimball (2013) e Bill Inmon (2001), para os quais implementação de um *Data Mart* funcional não é uma tarefa trivial, enfatizando que 80% (oitenta por cento) do esforço gerado por esta implementação se concentrou na fase da criação das transformações de ETL, pois, os dados da fonte estão brutos, inconsistentes, divergentes e não servem para a aplicação do OLAP; sendo, assim, necessária uma infraestrutura que garanta todo o tratamento correto para a implementação do OLAP.

Como pontos positivos teve-se o envolvimento do BI e tecnologias afins com a saúde pública no quesito das doenças citadas, a sumarização dos dados em formato compreensível para a maioria da população em dados, gráficos e estimativas legíveis para que se obtenha a informação adequada sobre o problema e o desafio da equipe em estudar uma tecnologia recente no mercado implementando uma solução através dela.

No decorrer deste trabalho, teve-se problemas com a forma de obter os dados da API de dados abertos da prefeitura de Recife-PE, devido aos administradores desta

modificarem sua estrutura de consulta de dados. Isso atrelado à falta de atualização constante dos dados são pontos negativos do projeto.

Demandou-se tempo e dedicação para finalizar todo o projeto, porém ainda existem melhorias que podem ser feitas como trabalhos futuros, tais como: incorporar a tecnologia de mineração de dados com os *dashboards* para procura de padrões consistentes e a forma de comportamento dos dados, a criação de aplicativo mobile e sistema desktop, o uso do BI para outros problemas de saúde, educação, política, etc. e a criação de *chatbots* interativos com o usuário.

REFERÊNCIAS

ABREU, Fábio. Desmistificando o Conceito de ETL. **Revista de Sistemas de Informação.** Nº 2. 2008. Disponível em: <http://www.fsma.edu.br/si/Artigos/V2_Artigo1.pdf>. Acesso em 30 Set. 2017.

BOTELHO, Fernando; FILHO, Edelvino. Conceituando o Termo Business Intelligence: Origem e Principais Objetivos. **Revista Sistemas, Cibernética e Informática.** Volume 11. Nº 1. 2014. Disponível em: <http://www.iiisci.org/journal/CV$/risci/pdfs/CB793JN14.pdf>. Acesso em 30 Set. 2017.

BOTH, Eder; DILL, Sérgio. **Business Intelligence Aplicado em Saúde Pública.** Universidade Regional do Noroeste do RS. 2012. Disponível em: <periodicos.unesc.net/sulcomp/article/download/793/744>. Acesso em 30 Set. 2017.

BRAGA, Miguel. **Esquema Estrela vs Esquema de Flocos de Neve.** 2015. Disponível em: <https://community.qlik.com/docs/DOC-12579>. Acesso em 30 Set. 2017.

BRITO, Elias. **Modelagem Dimensional.** 2015. Disponível em: <https://consultoremti.wordpress.com/2015/06/08/modelagem-dimensional/>. Acesso em 30 Set. 2017.

CRUZ, Bruno; MIRANDA, Bruno; TURCHETTE, Fellipe. **Conceitos de Business Intelligence por meio de Estudo de Caso: Ferramentas Pentaho e QlikView.** Universidade São Francisco. Itatiba. 2014.

DIÁRIO DE PERNAMBUCO. **Pernambuco registra mais de 45 mil casos de dengue somente esse ano.** Portal Diário de Pernambuco Pernambuco, 29 de março de 2016. Disponível em: http://www.diariodepernambuco.com.br/app/noticia/vida-urbana/2016/03/29/interna_vidaurbana,635489/pernambuco-registra-mais-de-1-8-mil-casos-de-microcefalia.shtml. Acesso em: 27 ago. 2016.

ELIAS, Diego. **Dimensões e Fatos no contexto do Business Intelligence.** 2014. Disponível em: <https://canaltech.com.br/business-intelligence/dimensoes-e-fatos-

no-contexto-do-business-intelligence-bi-18710/>. Acesso em 30 Set. 2017.

HENRIQUE, Ozimar. **Estruturas de DW – Kimball x Inmon.** 2012. Disponível em:<https://social.technet.microsoft.com/wiki/pt-br/contents/articles/10275.estruturas-de-dw-kimball-x-inmon.aspx>. Acesso em 30 Set. 2017.

INMON, W. H.; TERDEMAN, R. H.; IMHOFF, C. **Data Warehousing:** Como transformar informações em oportunidades de negócios. 3. ed. São Paulo: Berkeley, 2001.

KIMBALL, Ralph; ROSS, Margy. **The Data Warehouse Toolkit Second Edition:** the complete guide to dimension modeling. New York: John Wiley & Sons, Inc., 2002.

MOREIRA, Eduardo. **Modelo Dimensional para Data Warehouse.** 2006. Disponível em: <https://imasters.com.br/artigo/3836/gerencia-de-ti/modelo-dimensional-para-data-warehouse/>. Acesso em 30 Set. 2017.

PEREIRA, Altieri. **BI Introdução a SnowFlake e Star Schema.** 2014. Disponível em: <http://altieripereira.blogspot.com.br/2014/04/bi-introducao-snowflake-e-star-schema.html>. Acesso em 30 Set. 2017.

SANTO, Frederico. **Construção de um Data Mart para Apoio às Tomadas de Decisão das Empresas Proembarque e Casacon.** 102 f. Universidade do Vale do Itajaí, 2006.