

Implementação de Data Warehouse para área de saúde

Vanessa Fernandes¹ e Edinelson Aparecido Batista²

¹Engenheira de Computação, vanessa.fernandesb@gmail.com

²Mestre em Computação pela UNICAMP e professor do UNISAL, edinelson.batista@sj.unisal.br

Resumo – O volume de dados gerados por aplicações de sistemas de informação vem crescendo de forma exponencial nos últimos anos. Para analisar esses dados é necessária a criação de soluções de tecnologia que facilitem o tratamento e visualização dos mesmos, ajudando, assim, os gestores das organizações no processo de tomada de decisões. Essas soluções consistem na transformação de dados em informações ou fatos para que, a partir delas, sejam aplicadas ações com base na inteligência de negócios adquirida. Direcionado à área da saúde, esse projeto apresenta o estudo de aplicação para o setor público hospitalar utilizando metodologia de business intelligence e versão open source da ferramenta multi plataforma Pentaho.

Palavras-chave: Business intelligence, data warehouse, Pentaho.

Abstract – O volume de dados gerados por aplicações de sistemas de informação vem crescendo de forma exponencial nos últimos anos. Para analisar esses dados é necessária a criação de soluções de tecnologia que facilitem o tratamento e visualização dos mesmos, ajudando, assim, os gestores das organizações no processo de tomada de decisões. Essas soluções consistem na transformação de dados em informações ou fatos para que, a partir delas, sejam aplicadas ações com base na inteligência de negócios adquirida. Direcionado à área da saúde, esse projeto apresenta o estudo de aplicação para o setor público hospitalar utilizando metodologia de business intelligence e versão open source da ferramenta multi plataforma Pentaho.

Keywords: Business intelligence, data warehouse, Pentaho.

I. INTRODUÇÃO

Atualmente, vive-se na chamada era da informação, na qual a enorme quantidade de dados gerados e armazenados nas organizações gera um grande desafio aos gestores que precisam tomar decisões de forma rápida e eficiente. Por outro lado, ações tomadas a partir de uma análise adequada de dados, devidamente organizados, podem aumentar os lucros e diminuir os erros nas corporações.

Por apresentar uma tecnologia que permite aos gestores uma melhor visão sobre seus negócios, os

conceitos de *business intelligence* (BI) vêm sendo implementados ao longo dos últimos anos em diversas áreas do mercado. A implantação dessa metodologia na área da saúde contribui para, por exemplo, a criação de relatórios confiáveis de fatores críticos que podem determinar a qualidade de vida de uma população.

Como publicado em [1], pesquisas do IBGE revelam que em 2013 apenas 30 % da população brasileira possuía plano de saúde particular. Devido à crise enfrentada pelo país, este número vem diminuindo com a migração do setor privado para o setor público, aumentando assim, a quantidade de atendimentos nas unidades públicas. Este ponto preocupante do sistema de saúde também é abordado no G1, de acordo com [2], através de dados coletados pelo Ibope que mostram que 61 % dos brasileiros consideram a saúde como a área mais problemática do país. Pode-se atribuir essa preocupação os problemas de organização, qualidade de atendimento e falta de recursos básicos como leitos e remédios, além do déficit de médicos nos hospitais. Isso se confirma ao verificar na mesma pesquisa que 55 % dos entrevistados considera a demora no atendimento como o principal problema do sistema público, além disso, 57 % apontam como principal medida para melhorar o atendimento a contratação de mais médicos.

Um fator que pode colaborar com o nível de qualidade e diminuir a quantidade de atendimentos nos hospitais são a controle de epidemias e criação de campanhas preventivas. Com o uso da tecnologia o conhecimento rápido e fácil de regiões mais afetadas por epidemias e grupos de pessoas propícias a uma determinada doença, por exemplo, podem ser significantes para a eficácia das tomadas de decisão da gestão de saúde de um município para criação dessas medidas.

A proposta deste projeto é desenvolver uma aplicação utilizando os conceitos de BI. O projeto contemplará todas as fases do BI, desde a coleta e carga dos dados até a análise final destes dados, que possibilitará a geração de vários relatórios gerenciais que poderão dar apoio às tomadas de decisões. No desenvolvimento deste trabalho foi utilizada a versão *open source* da tecnologia Pentaho, que integra uma gama de recursos e aplicações, fornecendo uma plataforma completa que atende aos requisitos de uma solução de BI. Espera-se no final deste projeto a disponibilização de uma solução direcionada às necessidades hospitalares que contribuirá na tomada de decisões dos gestores da área da saúde.

II. REFERENCIAL TEÓRICO

Uma solução de BI se apoia, essencialmente, no uso de banco de dados. Antes da introdução dos conceitos de BI, no entanto, é importante destacar conceitos básicos de banco de dados e dos princípios da modelagem multidimensional que são usados na construção de um *Data Warehouse* (DW) para visão final das informações. O detalhamento e características particulares dos modelos serão explicados nas próximas sessões juntos aos demais conceitos que fundamentam a construção deste projeto e serão ilustrados pelos dados utilizados durante seu desenvolvimento.

A. Modelo transacional e relacional

De forma simples, modelos transacionais e relacionais são compostos de tabelas e relacionamentos entre tabelas. De acordo com [3], tabelas podem ser entendidas como um conjunto de linhas e colunas que qualificam cada elemento ou, de outra forma, qualificam cada ocorrência de um conjunto de dados que se deseja armazenar.

Os elementos ou ocorrências da tabela são identificados e relacionados por uma ou mais colunas, que são chamadas de chaves. Por sua vez, podem ser classificadas como primárias e estrangeiras e segundo [4], chaves primárias constituem a unicidade de valores que garantem a restrição de integridade dos dados de uma tabela dentro de um banco de dados. Elas podem ser simples ou compostas. Quando uma única coluna é suficiente para identificar uma única ocorrência é chamada de simples e quando uma combinação de colunas que juntas distinguem uma linha das demais dentro da tabela, a chave é denominada de chave primária composta.

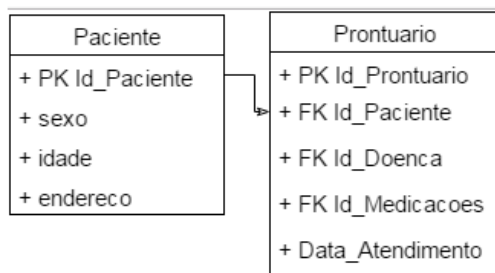
As chaves primárias podem estar presentes em outras tabelas que não são suas de origem, neste caso elas são chamadas de chaves estrangeiras. Assim, as chaves estrangeiras são utilizadas para implementar os relacionamentos entre as tabelas dentro do modelo relacional.

O modelo relacional é o mais comum nas organizações, pois são utilizados nos sistemas que dão apoio às operações diárias. Sua normalização deve retratar sempre uma realidade, o mundo real, tanto em sua nomenclatura de componentes como nos fatos que representa. Nesse modelo, as tabelas são chamadas de entidade e são classificadas por suas propriedades e características. Os relacionamentos existem entre duas entidades, ou seja, a relação existente entre um elemento de uma entidade com elementos de outras entidades. Utilizando o modelo apresentado neste projeto, pode-se visualizar na Figura 1 o relacionamento entre as entidades Paciente e Prontuário ligados pelas chaves primárias (PK, *primary key* em inglês) e estrangeiras

(FK, *foreign key* em inglês) correspondente ao elemento identificador de paciente.

Os relacionamentos podem possuir diversas cardinalidades que, por sua vez, pode ser entendida como quantidade de ocorrências de um relacionamento ou com quantos elementos cada elemento de uma entidade se relaciona com elementos de outra entidade

Figura 1 – Exemplo de modelo relacional entre paciente e prontuário.



Fonte: Acervo do autor.

Esses relacionamentos são definidos segundo [3] como:

1:1 - Ocorre sempre que uma entidade tiver uma única ocorrência para cada ocorrência na outra entidade.

1:n - Um dos mais comuns relacionamentos ocorre sempre que houver relacionamento de uma entidade com um ou mais elementos da outra entidade e essa outra se relacionar apenas com um elemento (também chamado de *tupla*) daquela entidade.

m:n - Ocorre sempre que uma entidade se relacionar com várias tuplas de outra entidade e essa, por sua vez, relacionar-se com várias *tuplas* daquela entidade.

B. Online Transaction Processing (OLTP)

Segundo [5], são sistemas utilizados nas transações organizacionais que suportam as operações denominadas CRUD (Consulta, Inserção, Alteração e Deleção de registros) no banco de dados transacionais. Essas operações são feitas através de uma interface de aplicação que executa comandos em linguagem SQL (*Structured Query Language*). De acordo com [3], SQL é o conjunto de comandos de manipulação de banco de dados utilizado para criar e manter a estrutura desse banco. Além de incluir, excluir, modificar e pesquisar informações nas suas tabelas.

O *software* responsável pelo gerenciamento do banco de dados é conhecido pela sigla SGDB (Sistemas Gerenciadores de Banco de Dados). Existem diversos SGDBs disponíveis no mercado, que podem ser usados em aplicações mais complexas e robustas ou em aplicações mais simples. Entre os mais populares destacam-se: *PostgreSQL*, *SQL Server*, *MySQL*, *DB2 da IBM*, *Oracle* e *Firebird*.

Neste projeto foi utilizado o MySQL, que é uma ferramenta atualmente mantida pela Oracle e a sua versão aberta, que é chamada de *Community Edition*,

possui uma interface gráfica cliente e servidor. O servidor é o responsável pelo armazenamento de dados e todo controle de transações e respostas de requisições, enquanto a interface gráfica cliente, chamada de *MySQL Workbench*, efetua a comunicação com o servidor através da linguagem SQL.

A ferramenta permite criar e configurar definições de banco de dados como tabelas, relacionamento, chaves primárias, chaves estrangeiras e executar comandos CRUD através de *scripts* ou pelo seu menu gráfico.

C. Modelagem multidimensional

De acordo com [6], a modelagem multidimensional é uma técnica de concepção e visualização de um modelo de dados de um conjunto de medidas que descrevem aspectos comuns de negócios. É utilizada especialmente para sumarizar e estruturar dados e apresentá-los em visões que suportem a análise dos valores dos mesmos. O modelo dimensional é formado pelos elementos básicos fato, dimensão e medidas.

Fatos são registros que acontecem no dia-a-dia de uma organização e que são considerados como as medições necessárias para avaliar o assunto desejado.

Ainda segundo [6], dimensões são as possíveis formas de visualizar os fatos, ou seja, são os “por” dos dados: “por mês”, “por país”, “por produto”, “por região”, etc. Para exemplificar pode-se utilizar o conceito de que as dimensões são formadas por membros:

“Um membro de uma dimensão é um nome diferente utilizado para determinar a posição de um item de dado. Por exemplo, todas as ocorrências de um ano, trimestre e mês fazem a dimensão tempo, e todas as cidades, estados e regiões fazem a dimensão geográfica”, [6].

Dentre esses membros, encontra-se a hierarquia como um meio de facilitar a visualização entre consultas, onde cada hierarquia possui diversos níveis. Já as medidas ou variáveis, são identificadas por serem quantificadas e possuem os valores das dimensões relacionados a um fato.

Tabela Fato

De acordo com [7], a tabela fato é a principal tabela de uma modelo dimensional onde os valores de interesse da empresa estão armazenados. A nomenclatura fato é utilizada para denominar medidas como quantidades, valores e indicadores. A tabela fato guarda esses valores e é composta por uma chave primária e pelas métricas referentes ao fato. Portanto, seus valores não devem ser zeros ou nulos para representar que nada aconteceu (por exemplo, que não houve atendimento de paciente em determinada data), pois isso faria com que a tabela de fatos crescesse demais.

Tabela Dimensão

A tabela dimensão é composta por atributos que agrupam e caracterizam o negócio e costuma ser bem mais simples e menor que a tabela fato. Para melhor visualização da estrutura dos dados pode-se utilizar dois tipos de modelos multidimensionais: modelo estrela e modelo foco de neve.

Modelos estrela e Floco de Neve

O modelo Estrela (*Star Schema*) é composto pela tabela fato e pelas tabelas dimensões. O nome “estrela” é dado pela disposição em que as tabelas são organizadas, sendo a tabela fato ao centro e as dimensões ao seu redor interligando-se com a tabela fato, [7].

O modelo Floco de Neve (*Snowflake*) tem como base o modelo estrela, com adicional de divisão hierárquica entre os membros da tabela dimensão.

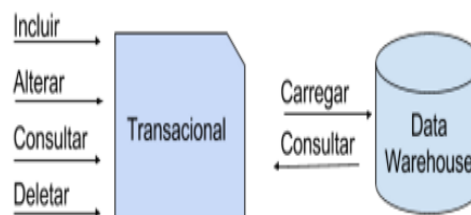
D. Data Warehouse

Segundo [8], *data warehouse* é um conjunto de dados baseado em assuntos, integrado, não-volátil, e variável em relação ao tempo, de apoio às decisões gerenciais. Ao projetar um DW, é importante se atentar ao nível de granularidade dos dados. A granularidade define o detalhamento desses dados, sendo que quanto menos detalhes o DW tiver, mais alto será seu nível de granularidade.

Outro ponto importante é a distinção dos bancos transacionais do ambiente de DW e como os dados se comportam na variação do tempo. Segundo [6], os dados do sistema transacional refletem o valor corrente de alguma coisa, sua exatidão é válida por determinado instante podendo ser alterado e atualizado. Já em DW os dados são precisos em relação ao tempo, pois a estrutura é projetada sobre janelas de períodos.

Como evidenciado por [8], a sumarização dos dados transacionais é relativamente pequena em relação ao DW. Isso é facilmente notado, visto que para ambientes transacionais períodos trimestrais são satisfatórios para análises e DW são significativamente maiores com períodos de 2 a 5 anos. Outra característica que diferencia esses ambientes são as funcionalidades realizadas nos sistemas. Por ser não volátil, o DW não permite utilização de funcionalidade de modificação e exclusão de dados em seu sistema, como ilustra a Figura 2.

Figura 2 – Diferença de funcionalidade entre modelo transacional versus Data Warehouse.



Fonte: Acervo do autor.

Para melhor acompanhamento durante o projeto, foi utilizado o conceito de [7], que segue uma abordagem iterativa centrada nas necessidades de entrega, aplicando as seguintes etapas:

- Levantamento das necessidades: devemos nos atentar aos requisitos do usuário. Nesse momento é realizado o levantamento das tabelas dimensões e tabela fato e como seus dados serão cruzados.
- Mapeamento dos dados: são mapeados os dados e as fontes externas que irão formar o DW. Aqui aproveita-se para validar a existência dos requisitos levantados na primeira etapa.
- Construção da *Staging Area*: trata-se do ambiente de transição dos dados do DW. Nessa fase, construímos sua estrutura e copiamos os dados do OLTP, que também passam pelo processo de ETL (do Inglês, *Extraction, Transformation and Loading*) para carga posterior das tabelas. ETL será explicado a seguir.
- Construção das tabelas: são construídas as estruturas e historicidade das tabelas dimensões e fato. Durante a construção da fato, definimos também a granularidade das informações que farão parte dela.
- Definição do processo geral de carga: são definidos como os dados serão processados e carregados no DW.
- Documentação: avaliamos e documentamos todo o processo da construção do DW, fornecendo apoio importante para a gestão do conhecimento.

Ao final destas etapas, o DW criado oferece os elementos básicos definidos também por [7]: *Fonte Operacional de Dados*, *Área de Retenção (Staging Area)*, *Servidores de Apresentação*, *Nível de acesso do Utilizador final*. Os dados armazenados em um DW podem ser visualizados pelos cubos multidimensionais, utilizando ferramentas OLAP (*OnLine Analytical Processing* em inglês).

Cubo

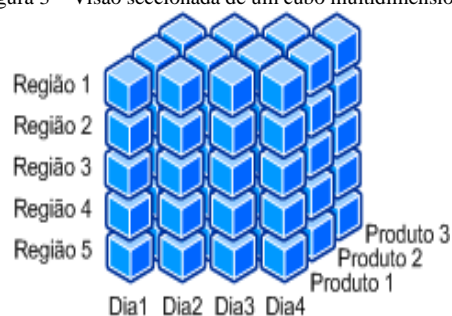
A metáfora denominada cubo é utilizada para transmitir a ideia de múltiplas dimensões. O modo como os dados são organizados permite visualizar os dados em fatias por diferentes perspectivas e cada “quadrado” fornece a informação de um elemento composto por diversas dimensões. Utilizando como exemplo a Figura 3 é possível observar as informações dos produtos vendidos em uma determinada região em um determinado dia.

Extract-transform-load (ETL)

Como já mencionado, após a coleta e armazenamento dos dados é necessário que eles sofram algumas transformações para atender às necessidades de negócio. De acordo com [7], o processo de ETL envolve extrair os dados de fontes distintas, transformá-los e carregá-los, preparando-os para ambiente de DW.

Como o nome diz, a etapa de extração é realizada para extrair dados de diferentes e distintos sistemas de origem e depois carregar para o ambiente de operações chamada de *Staging Area*, onde as operações de transformação serão aplicadas antes de seu carregamento em DW.

Figura 3 – Visão seccionada de um cubo multidimensional.



Fonte: [9].

Ao serem carregados para a *Staging Area*, os dados passam pela etapa de transformação que envolve limpeza, ajuste e consolidação dos dados. Nessa etapa os dados são tratados para que sejam consolidados de duas ou mais fontes, traduzidos e seus valores padronizados. Em [6], emprega-se um exemplo muito conhecido para exemplificar esse passo, utilizando como o tratamento da entidade ‘Sexo’, que pode ser carregada como “F/M”, “H/M” ou ainda “Feminino/Masculino” em diferentes sistemas de origem. Em um ambiente onde a entidade sexo foi extraída de fontes externas, transformamos essas nomenclaturas para que haja um padrão no armazenamento final F e M, por exemplo. Assim, após a carga de DW as informações relacionadas à entidade poderão ser facilmente encontradas e agrupadas.

Por último é executada a carga dos dados no banco do DW. Este processo de extração, transformação, limpeza e carga pode ser realizada de diferentes formas. Pode ser usado *scripts* SQL, que pode ser um processo bastante trabalhoso ou usar uma ferramenta ETL construída para este propósito. Um exemplo destas ferramentas é a Suíte Pentaho que foi utilizada neste trabalho.

OnLine Analytical Processing (OLAP)

Segundo [6], OLAP é uma tecnologia voltada para consultas de dados, permitindo a análise dos mesmos a partir de relatórios dinâmicos que satisfazem as necessidades gerenciais. Com o uso de operações como *drill down* e *roll up*, o usuário pode navegar entre os dados, através de uma visão clara e personalizada das diversas perspectivas das informações organizacionais movimentando-se entre os níveis hierárquicos de uma dimensão, aumentando ou diminuindo seu detalhamento. As Figuras 4 e 5 ilustram esses processos.

Figura 4 – Visão *Roll Up* de dados.

Volume de Produção (em milhares)		1999			
		Trim.1	Trim.2	Trim.3	Trim.4
Região Sul	RS	78	67	22	56
	SC	90	67	88	99

↑ Roll Up (Dimensão Tempo) ↑

Volume de Produção (em milhares)		Trimestre 1		
		Janeiro	Fevereiro	Março
Região Sul	RS	30	26	22
	SC	28	30	32

Fonte: [6]

Figura 5 – Visão *Drill Down*

Volume de Produção (em milhares)		Telefone Celular		Pagers	
		2001	2002	2001	2002
Região Sul	RS	33	12	8	12
	SC	45	34	20	23

↓ Drill Down (Dimensão localização geográfica) ↓

Volume de Produção (em milhares)		Telefone Celular		Pagers	
		2001	2002	2001	2002
RS	Canoas	13	4	2	5
	Porto Alegre	20	8	6	7

Fonte: [6].

Além do *drill down* e *drill up*, em [6] descreve-se outras formas de navegação OLAP:

- *Drill Across*: navegação entre níveis intermediários de uma dimensão, sem necessidade de sequência de suas posições.
- *Drill Through*: navegação entre dimensões diferentes.
- *Slice and Dice*: modifica a posição dos dados alterando linhas por colunas e gira o cubo.

Para facilitar a análise OLAP pode-se utilizar também alertas, *ranking*, filtros, *sorts* e ainda *breaks*. Eles são pontos que pode se aplicar em relatórios para destacar elementos, ordená-los, filtrá-los e agrupá-los de acordo com a necessidade do usuário final, podendo haver uma gama de variações como

- *DOLAP (Desktop On Line Analytical Processing)*: requisição de microcubo do cliente ao servidor para análise em Workstation (Cliente).
- *ROLAP (Relational On Line Analytical Processing)*: levantamento das necessidades: cubo é mantido no servidor, cliente apenas faz requisição de consulta ao servidor de banco de dados relacional.
- *MOLAP (Multidimensional On Line Analytical Processing)*: acesso e controle de manipulação do cubo diretamente no servidor multidimensional. É a arquitetura que mais gera benefícios de desempenho, pagando com o tempo de aquisição e problemas com escalabilidade.
- *HOLAP (Hybrid On Line Analytical Processing)*: É a junção das arquiteturas ROLAP e MOLAP trazendo benefício de alto desempenho do MOLAP com o processo de escalabilidade ROLAP.

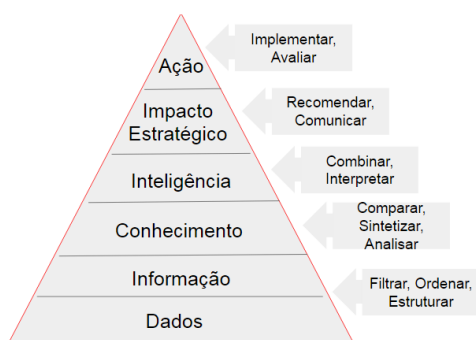
E. Business Intelligence

Segundo [6], *business intelligence* é um conjunto de tecnologias que permite o cruzamento de informações organizacionais que suportam análises dos indicadores de performance corporativa.

Entre suas vantagens, o BI fornece aos gestores uma melhor visão de seus negócios de uma maneira simples e dinâmica. As análises em tempo real e a praticidade de acesso e compartilhamento das informações auxiliam a identificar pontos de ação de negócio, colaborando na tomada de decisão.

Com a aplicação de BI pode se acompanhar a evolução dos registros armazenados em grandes bancos de dados em inteligência de negócio, como é visto na Figura 6.

Figura 6 – Processo de transformação de dados com a aplicação de BI.



Fonte: Acervo do Autor.

Os dados passam por um processo de filtragem e reordenação, transformando-se em informação. Após serem submetidas a análises e comparações, essas informações tornam-se conhecimento. Ao ser interpretado pelos gestores, esse conhecimento evolui a inteligência de negócio, possibilitando impactos estratégicos dentro da tomada de decisão e ações dos gestores. Pode-se dizer que o BI usa como base os dados armazenados no DW para produzir os relatórios e informações que serão a base para as análises descritas.

Pentaho

Pentaho é uma solução completa de inteligência de negócios que oferece uma gama de ferramentas gráficas e intuitivas que contemplam o cenário de BI. A ferramenta torna-se poderosa por sua flexibilidade e independência de plataforma, segurança e custo mínimo de implantação e manutenção.

Conforme disponibilizado em [10], a plataforma simplifica a preparação e combina os dados, oferecendo ferramentas para facilmente analisar, visualizar, explorar e gerar relatórios. Sua arquitetura oferece a possibilidade para que os membros de uma empresa possam traduzir dados em valor.

A plataforma oferece diversos pacotes e serviços e pode ser utilizada de acordo com as necessidades de cada cliente. Por ser multi plataforma e flexível, ela também pode ser utilizada em diversos sistemas operacionais e fornece conexão com os principais SGDBs, como Postgree, MySQL, Oracle, SQLServer.

Para suprir as necessidades levantadas neste projeto, foi utilizada a versão Suit *open source* (Community Edition) em conexão com MySQL. A Suit centraliza diversas aplicações que contemplam todo o ciclo de implementação de um sistema de apoio. São elas:

- PDI (*Pentaho Data Integration*): designer gráfico ETL que fornece de forma simples, diversas maneiras de conectar base de dados relacional, transformar os dados e fazer a carga no DW. Por ser ferramenta visual oferece uma ampla biblioteca de componentes chamados ‘steps’ para acessar, preparar e misturar os

dados relacionais, que são conectados interativamente com apenas alguns cliques para combinar transformações e carga no DW. Além dos componentes para construção e carga do DW é possível acompanhar em tempo real todos os processos e programar para que os fluxos sejam executados automaticamente.

- Mondrian (*Schema Workbench*): é uma ferramenta que permite criação de cubos e análise interativa a partir de consultas multidimensionais com mecanismo OLAP. Com integração com o BI Server, a Mondrian permite publicação de cubos diretamente no servidor. Outro ponto poderoso é que Mondrian utiliza a linguagem de consulta multidimensionais MDX que são esquemas de metadados XML criados em uma estrutura própria da Mondrian. Esses esquemas podem ser considerados estruturas de cubo que utilizam tabelas de fatos e dimensões.
- PUC (*Pentaho User Console*): Esta ferramenta é voltada para visão do usuário final. Permite o desenvolvimento e implantação de *dashboards* de forma rápida e eficiente.

De acordo com as informações em [10], para ajudar a construir painéis personalizados junto ao PUC, pode-se utilizar um conjunto de ferramentas e componentes chamado CTools. Sendo utilizado neste projeto o CDE (Painel Editor de comunidade). Nesta mesma fonte, pode-se encontrar a descrição de CDE como uma ferramenta gráfica que permite o desenvolvimento e implantação de *dashboards* avançados, sendo uma ferramenta muito poderosa e completa pois permite criar, editar e visualizar *dashboards*, combinando assim, *front-end* com fontes de dados e componentes personalizados de uma forma perfeita

III. METODOLOGIA

A metodologia de pesquisa empregada neste projeto assumiu o caráter exploratório, visto que a área de BI vem crescendo e sendo cada vez mais utilizada nas organizações. Foram analisados diversos artigos e livros que abordam essa tecnologia e realizado pesquisas relacionadas à melhor forma de aplicá-las.

Para obtenção de dados diretamente ao hospital e para delimitar pontos relevantes para geração de análise, foi realizado um estudo de caso e entrevistas direta com gestores, utilizando assim, a metodologia não experimental.

Com a finalidade de geração de um protótipo de BI, com relatórios e *dashboards* prontos para utilização do hospital, o presente projeto também assumiu caráter qualitativo e aplicado.

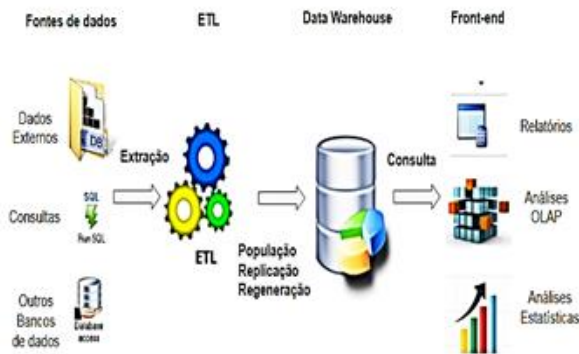
IV. DESENVOLVIMENTO DO PROJETO

As etapas que formam o desenvolvimento deste projeto seguem os passos gerais utilizados para implementação de um projeto de BI conforme descreve a Figura 7. A seguir será explicada cada uma dessas etapas detalhadamente.

Foi realizada uma vasta pesquisa, análise e seleção de artigos e livros que fundamentam a implementação de tecnologia BI com direcionamento para área da saúde. As pesquisas se estenderam às plataformas de BI disponíveis no mercado que disponibilizassem uma versão *open source* e de alto desempenho que atendessem todas as necessidades de uma solução de BI.

O passo seguinte foi apresentar a proposta de projeto para gestores do hospital municipal de Paulínia, coletar dados, dúvidas e necessidades frequentes de consulta que pudessem ser tratadas e respondidas com a utilização da tecnologia proposta. Com isso, foi possível a criação de um modelo de base de dados fictícia e em seguida essa base foi criada no banco de dados MySQL.

Figura 7 – Apresentação de etapas de desenvolvimento.



Fonte: Acervo do autor.

A base é composta por 5 tabelas principais que englobam os pontos relevantes para análise sendo elas:

Paciente: foram considerados apenas os dados de relevância para análise do paciente. Assim, foram levantados campos que ajudassem na identificação de pacientes por gênero, idade e região de moradia.

Doença: foram mapeados dados de descrição e categoria de doença.

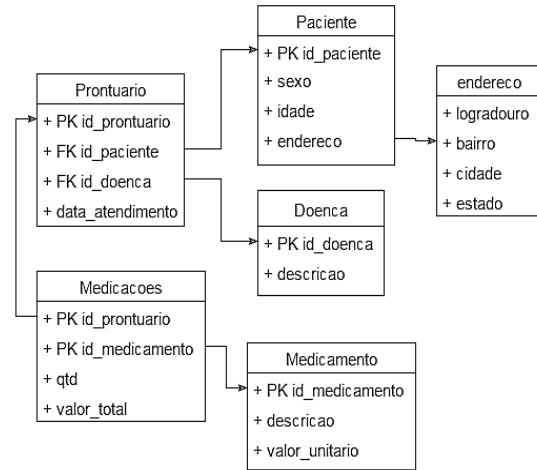
Itens de Medicação: nesta tabela foram armazenados os dados correspondentes aos medicamentos aplicados aos pacientes por atendimento. Os campos que formam parte dessa tabela são identificação de medicamento e prontuário, quantidade de medicamentos ministrados e o valor total calculado pela multiplicação da quantidade com o valor unitário do medicamento.

Medicamento: nesta tabela são registrados apenas valores unitários de medicamento e sua descrição.

Prontuário: tabela responsável por armazenar os dados de prontuário com os identificadores de paciente, doença e data de atendimento.

O modelo de dados gerado pode ser visto na Figura 8. Este modelo também pode ser chamado de transacional ou OLTP.

Figura 8 – Modelo Relacional de dados utilizado em base fictícia de hospital.

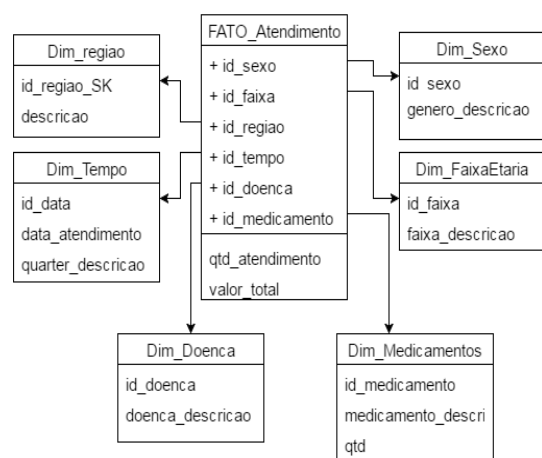


Fonte: Acervo do autor.

Como já mencionado, o modelo multidimensional é composto pela tabela fato e suas dimensões. Logo, foi definida a tabela fato que contém dados relevantes ao fato atendimento e tabelas dimensões que serão os “por” de nossas análises, ou seja, “por região”, “por data”, “por gênero”, “por doença”, “por faixa etária” e “por medicamentos”. Desse modo, perguntas como “quantidade de pacientes atendidos do sexo feminino em determinada época do ano”, “grupo de pessoas tendenciosas a um tipo de doença específica” poderão ser respondidas de acordo com o cruzamento e comparação de informações deste modelo.

O modelo multidimensional de dados pode ser visualizado na Figura 9 a seguir.

Figura 9 – Modelo multidimensional para construção de tabela fato e dimensões.



Fonte: Acervo do Autor

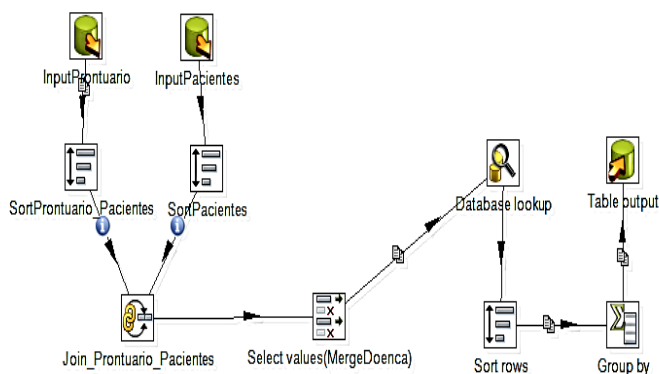
Para construção do DW, foi aplicado o processo de ETL utilizando a ferramenta *Data Integration Spoon do Pentaho* como pode ser visualizado na Figura 10.

Na ferramenta *Spoon* existem diversos componentes que são utilizados para fazer a movimentação dos dados. Através destes elementos os dados fluem pelo programa através de uma ligação entre eles e essas ligações são realizadas de forma interativa com apenas "clique e arraste". Essas ligações são chamadas de *hop*. Os principais componentes utilizados nesta etapa foram:

Table Input: componente que lê e carrega informações de um banco de dados, usando uma conexão e comandos SQL. O componente também oferece o carregamento dos dados do SQL com a utilização do botão 'Get' que retorna os campos da tabela selecionada.

Dimension Lookup/Update: componente usado para a carga das tabelas dimensões.

Figura 10– Visão de carregamento de tabela fato utilizando ferramenta Spoon.



Fonte: Acervo do autor.

Select Values: como o próprio nome diz, o componente é utilizado para selecionar e configurar os campos retornados de uma etapa de consulta de dados.

Sort Values: é responsável pela ordenação e organização dos dados para serem usados em outras etapas como *Merge Join*.

Merge Join: o *Merge Join* executa a clássica junção entre conjuntos de dados com dados de duas tabelas diferentes. Entre as opções de *merge* incluem *Inner*, *Left Outer*, *Right Outer*.

Table Output: esse componente trata da etapa final dos processos de carregamento, ou seja, o destino dos dados. Ele permite carregar dados em uma tabela do DW, oferecendo opções de configuração para a tabela de destino e inclusão de novos registros.

Como a modelagem estrela trata as conexões das dimensões e fatos com as chaves únicas, para a carga das tabelas dimensões foram carregadas as chaves (*IDs*) de cada tabela, respeitando as hierarquias que seriam

utilizadas nas análises OLAP. Assim, ao final das transformações pode-se visualizar as tabelas no DW carregadas com os dados correspondentes aos *IDs* de cada registro. Então, a ferramenta *Data Integration Spoon* foi utilizada para carregar os dados das tabelas transacionais para as tabelas do modelo dimensional.

Quando as dimensões e tabela fato estão carregadas é utilizada outra ferramenta da *Suit Pentaho*, o *Mondrian* para geração dos cubos. Da mesma forma como a *Spoon*, as definições e configurações das etapas de criação do cubo são feitas de forma simples e rápida. A partir da conexão definida e estabelecida com seu DW os campos de definições são preenchidos com as opções existentes em seu esquema.

Para a criação dos cubos é necessário definir tabelas, campos associativos e hierarquias das dimensões e suas relações com a tabela fato. A visão geral dessas definições pode ser visualizada na Figura 11.

Neste ponto foram criados dois cubos para que atendessem às navegações por valor total de medicações e quantidade de atendimentos.

Figura 11 – Visão geral de criação de cubo com Mondrian.



Fonte: Acervo do autor.

Após criação dos cubos eles são publicados para que possam ser utilizados nas demais ferramentas. Por já existir a conexão com o DW, a publicação dos cubos pode ser realizada no próprio *Mondrian* e ser acessado no PUC (ferramenta de controle de usuário).

Com a utilização do PUC, iniciou-se a última etapa da implementação do processo de BI. Com os dados tratados e carregados no ambiente multidimensional é possível visualizar e analisar o cruzamento e agrupamento desses dados de diversas formas. Utilizando o PUC, essas análises ganham visões gráficas e dinâmicas para o usuário final e seu ambiente interativo tabular pode ser visualizado nas Figuras 12 e 13.

Figura 12 – Visão analítica gerada para seleção de campos.



Fonte: Acervo do autor.

Figura 13 – Visão tabela gerada com seleção de campos.

		Measures			
		Qtd_Atendimento			
		Faixa Etaria			
descricao_sexo	quarter	f1	f2	f3	f4
F	2014	506	480	7.960	3.973
	Q1	143	130	2.560	1.537
	Q2	121	70	1.920	870
	Q3	143	210	1.840	870
	Q4	99	70	1.640	696
	2015	253	320	5.360	2.726
	2016	605	820	12.280	7.163
M	2014	473	500	7.320	4.524
	2015	418	330	5.520	3.045
	2016	693	760	11.760	4.988

Fonte: Acervo do autor.

Nesta etapa, deu-se então, a criação de *dashboards* que “materializam” os dados em valores estatísticos.

Os valores dos gráficos são obtidos a partir do MDX gerado da seleção das dimensões, que podem ser classificados como linhas, colunas e filtros. O resultado do MDX pode ser pré-visualizado em tabela na visão analítica do PUC. Esse MDX é inserido no campo *query* de cada gráfico e pode ser alterado a qualquer momento de acordo com a necessidade do usuário. Além das *queries* MDX, *scripts* simples em SQL podem ser utilizados para carga dos componentes que serão visualizados nos *dashboards*.

Para a exibição dos gráficos é oferecido diversos tipos de *layouts* e gráficos como de setores, barras, linhas, entre outros e podem ser escolhidos com alguns cliques e personalizados posteriormente com comandos HTML e seleção de propriedades previamente carregadas para cada tipo de gráfico.

Para simplificar as análises dos gráficos é possível criar filtros utilizando componentes seletores como

¹ Indicadores de performance em que, a partir de definição de valores máximos e mínimos, podemos analisar se a informação medida está dentro do valor esperado.

checkboxes, *dropdownlists*, ou ainda filtros dinâmicos entre os gráficos utilizando o campo *clickable* e *scripts* básicos *JavaScript* que traduzem cliques de um gráfico em eventos.

A ferramenta também disponibiliza a customização de todo ambiente de visualização do usuário. Assim, com a utilização de comandos básicos de HTML foi customizado o *layout* do ambiente PUC para que atendesse aos padrões da apresentação final.

V. RESULTADOS E DISCUSSÕES

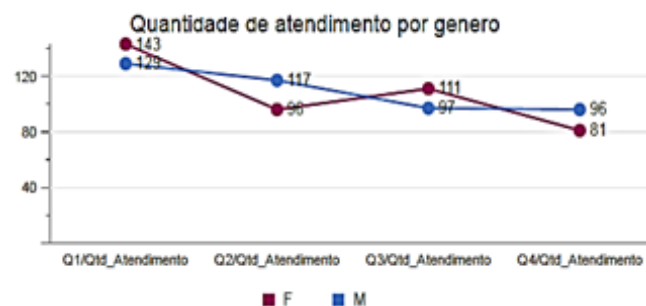
A utilização dos conceitos de BI junto às ferramentas de apoio MySQL e *Suit Pentaho* descritas neste artigo criaram o ambiente de desenvolvimento que fornece dados relevantes às análises de forma simples e organizadas.

Com a utilização de componentes tais como tabelas, gráficos de diversos estilos e ainda *gauges*¹ os dados foram distribuídos em *dashboards* agrupados de forma com que respondessem às questões levantadas na fase de planejamento.

Para uma questão de organização e direcionamento, o mapeamento dessas questões criou uma *task list*, onde cada *task* corresponde à “história do usuário”, ou seja, tarefas focadas em atender o objetivo do usuário. Assim, os gráficos que formam o resultado final deste projeto foram criados de acordo com a *task list*.

Pode-se exemplificar este ciclo de desenvolvimento seguindo a *task* “Entre os atendimentos no período de um ano, gostaria de saber quantas mulheres foram atendidas. Além disso, gostaria de identificar o período do ano onde houve maior ocorrência de atendimento”. A partir de sua descrição foi construído um gráfico de linhas que atende a tarefa citada e utilizando um filtro de ano os dados foram combinados para que estas informações fossem apresentadas como mostra a Figura 14.

Figura 14 – Gráfico de painel Perfil de Paciente correspondente a quantidade de atendimento por gênero nos períodos do ano.



Fonte: Acervo do autor.

Visando uma melhor apresentação final dos resultados, os gráficos foram classificados por áreas de análise, sendo elas perfil de paciente, doenças e atendimentos em geral. Estes grupos de gráficos foram dispostos em três painéis de controle ou *dashboards* para que as informações mais importantes de cada uma das áreas fossem visualizadas em uma só tela.

Entre a classificação da área de pacientes, outro ponto importante entre as análises para traçar um perfil é a faixa etária em que os pacientes são classificados. Seguindo o mesmo comportamento de cruzamento de informações por períodos do ano, pode se observar na Figura 15 o resultado do somatório dos atendimentos por faixa no ano e, entre eles, com destaque nos períodos de maior relevância.

Outros gráficos contendo os dados ordenados por região e ainda comparativos de atendimento de idade entre os três últimos anos, compõem o *dashboard* Perfil do Paciente. Assim, de forma rápida pode-se identificar, por exemplo, as regiões de moradia com maior frequência de atendimento no hospital nas estações do ano, auxiliando a direcionar possíveis reforços em postos de atendimento locais.

Figura 15 – Gráfico de painel Perfil de Paciente correspondente a quantidade de atendimento por faixa etária nos períodos do ano.

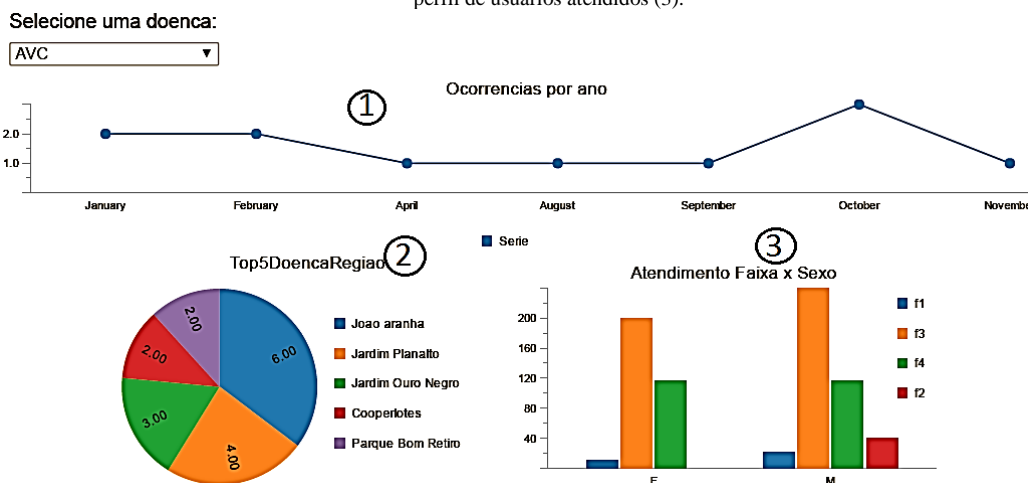
Quantidade de Atendimento por Faixa Etária Anual

f4	1363	928	9860
f3	2160	1920	19960
f2	170	190	1220
f1	132	110	1056
	Qtd_Atendimento/Q1	Qtd_Atendimento/Q2	Qtd_Atendimento/Q3

Fonte: Acervo do autor.

Além do perfil de Paciente, foram geradas análises referentes às doenças com maiores ocorrências entre os atendimentos. Por existir diversas doenças mapeadas, foi utilizado um filtro de seleção por doença, que detalhasse informações para identificar possíveis epidemias. Deste modo, a partir do evento de seleção é exibido o período do ano, com granularidade mensal, a idade e o sexo, além das cinco regiões de moradias dos pacientes em que a doença foi mais diagnosticada como pode ser observado na Figura 16.

Figura 16 – Grupo de gráficos pertencentes a painel Perfil de Epidemias mapeamentos por período do ano (1), região de maior ocorrência (2) e perfil de usuários atendidos (3).

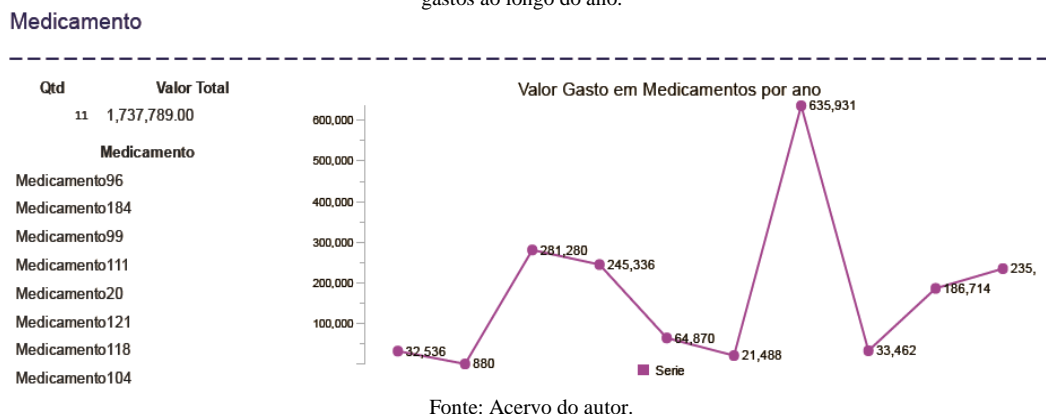


Fonte: Acervo do Autor.

As doenças podem possuir ainda registros dos medicamentos ministrados durante os atendimentos e, conseqüentemente, pode-se extrair os gastos gerados nestas aplicações. A Figura 17 mostra um exemplo de como gestores podem visualizar tanto a listagem desses

medicamentos quanto o gasto ao longo do ano. Os dados contidos nestas informações tornam-se eficazes no auxílio de planejamento de compras dos medicamentos evitando, por exemplo, falta de remédios nos hospitais.

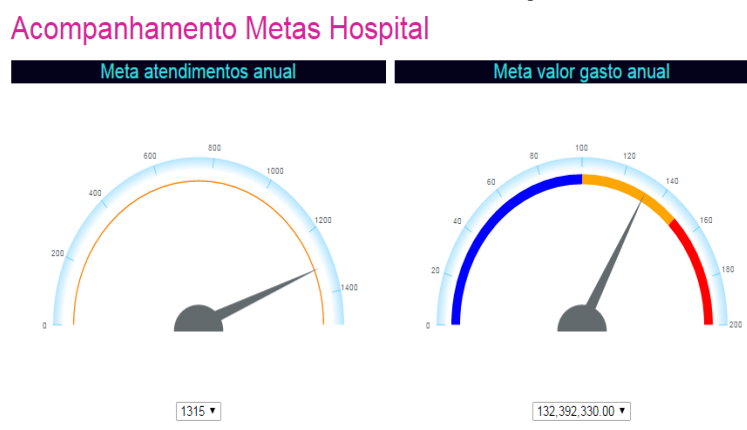
Figura 17 – Gráficos de painel *Perfil Epidemia* equivalente a informações de medicamentos ministrados por doença e seus respectivos valores gastos ao longo do ano.



As análises de BI trazem diversos componentes que podem auxiliar a mensurar valores alcançados a partir de metas definidas pelos gestores. A Figura 18 a seguir, exemplifica o uso de *gauges* aplicados para exibição dos

indicadores de performance de quantidade de atendimentos realizados no período de um ano e indicativo de valor em reais gasto pelo hospital em um intervalo de valores definidos.

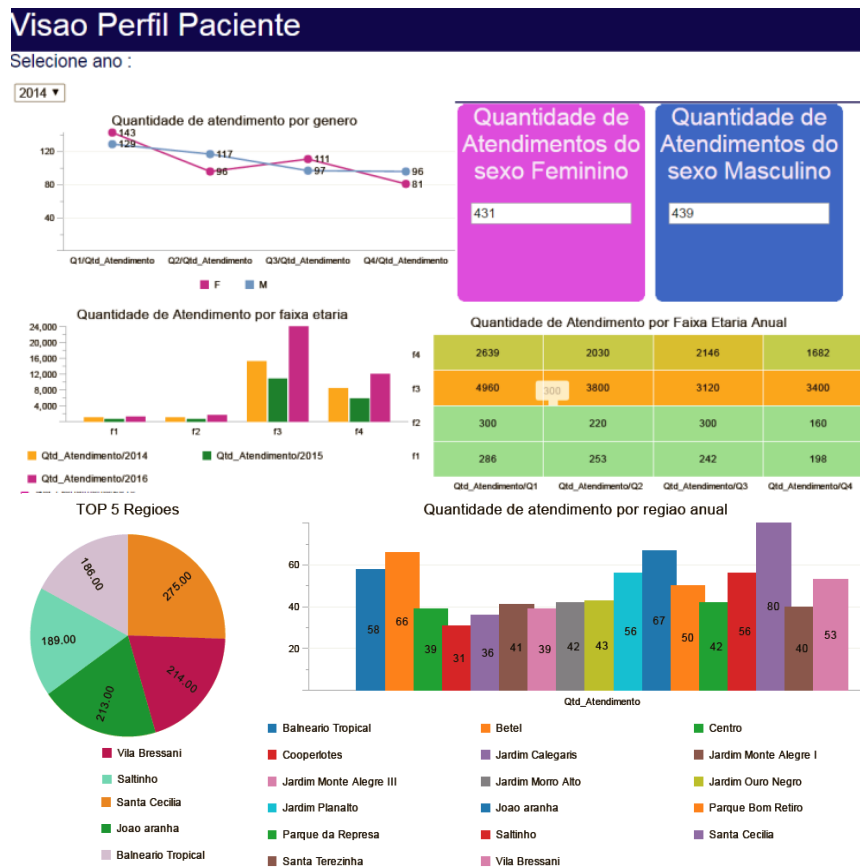
Figura 18 - *Gauges* relacionados a painel de Atendimentos com indicadores de número de atendimentos anuais e valores investidos em medicamentos dentro de uma divisão de intervalo pré-definidos.



Os dados são exibidos através de indicadores, e suas informações são apresentadas com relatórios visuais e intuitivos, conforme visto em alguns dos exemplos anteriores, observou-se que a visão gerencial poderá ser feita de forma mais rápida e simplificada auxiliando nas tomadas de decisões.

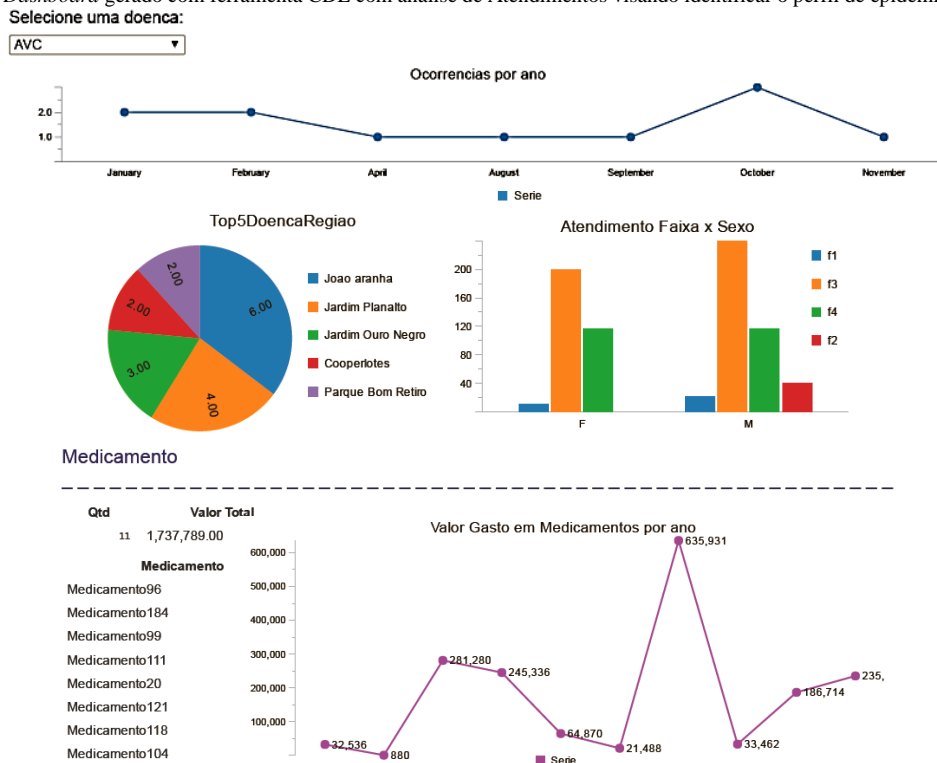
A visão completa dos *dashboards* gerados como resultado final deste projeto, contemplando as três visões gerais de perfil de paciente, mapeamento de epidemia e atendimento geral do município, podem ser visualizados respectivamente nas Figuras 19 a 21 a seguir.

Figura 19 – Dashboard gerado com ferramenta CDE com análise de Atendimentos visando identificar o perfil de pacientes atendidos.



Fonte: Acervo do Autor.

Figura 20 – Dashboard gerado com ferramenta CDE com análise de Atendimentos visando identificar o perfil de epidemias frequentes.



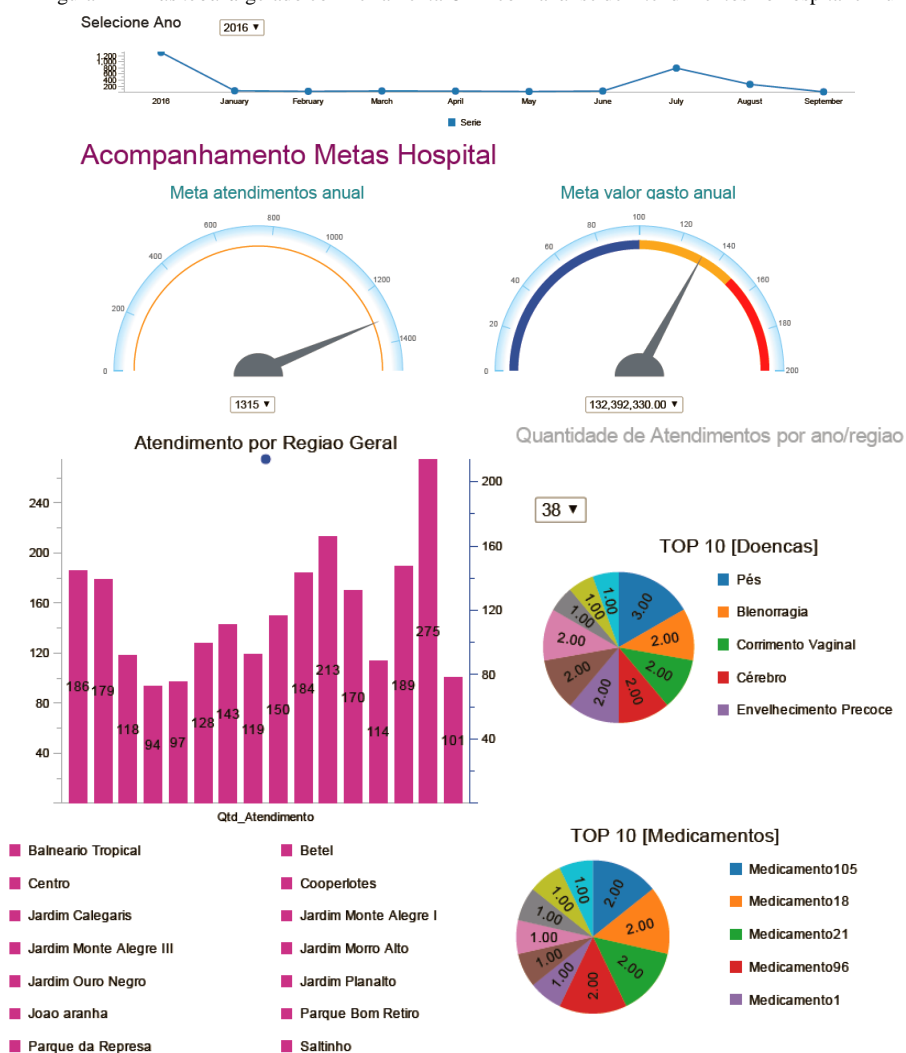
Fonte: Acervo do Autor.

FERNANDES, V.; BATISTA, E. A.

Automação Implementação de Data Warehouse para área de saúde

©Revista Ciência e Tecnologia, Campinas, v. 20, n. 37, p. 105 - 119, jul./dez. 2017 - ISSN: 2236-6733

Figura 21 – Dashboard gerado com ferramenta CDE com análise de Atendimentos no hospital e município.



Fonte: Acervo do Autor.

VI. CONCLUSÕES

Após a implementação e visualização dos dashboards criados, pode-se concluir que a aplicação dos conceitos de BI na área da saúde, assim como em outras áreas, torna-se vantajoso para uso dos gestores da área médica.

A centralização dos dados e seu devido tratamento facilitam as análises e expõem os principais pontos de atenção de todo o sistema. A *Suit Pentaho* também mostrou-se eficiente, sendo capaz de atender as necessidades levantadas no início do projeto.

A ferramenta fornece, além dos componentes utilizados nesta versão, outros componentes e extensões que poderão implementar análises de acordo com as necessidades futuras.

Portanto, este trabalho mostrou que a utilização de ferramentas livres é viável e plenamente aplicável em projetos de BI. Este projeto também mostrou que é

possível a implementação de BI em hospitais e isto poderá trazer melhorias nas análises dos dados, fornecendo apoio à tomadas de decisões dos gestores hospitalares, como foi proposto no início deste projeto.

No desenvolvimento do projeto pode-se observar a existência de pontos de evolução. Um primeiro item a ser considerado é a integração de dados dos médicos nos atendimentos ocorridos. Com a adição dos dados referentes aos médicos torna-se possível também a combinação de agendamentos de plantões com períodos de maior sobrecarga de atendimento. Outro ponto importante para a melhoria deste projeto é contabilização do tempo de espera e consulta; esta integração auxiliaria, assim, na análise de qualidade dos atendimentos.

REFERÊNCIAS

FERNANDES, V.; BATISTA, E. A.

Automação Implementação de Data Warehouse para área de saúde

©Revista Ciência e Tecnologia, Campinas, v. 20, n. 37, p. 105 - 119, jul./dez. 2017 - ISSN: 2236-6733

- [1] UOL. **IBGE revela como anda a saúde do Brasil.** São Paulo, 2015. Disponível em: <http://noticias.uol.com.br/saude/ultimas-noticias/redacao/2015/06/02/ibge-revela-como-anda-a-saude-do-brasil.htm>. Acesso em 01 de out. de 2016.
- [2] LIMA, S. Serviço público de saúde é ruim ou péssimo para 61%, diz pesquisa. Disponível em: <http://glo.bo/xmVQkt>. Acesso em 01 de outubro de 2016.
- [3] OLIVEIRA, C.H.P. **SQL Curso Prático.** São Paulo: Novatec Editora, 2002.
- [4] HEUSER, C.A. **Projeto de Banco de Dados.** Porto Alegre: Editora Sagra Luzzatto, 2001.
- [5] ALVES, W.P. **Banco de dados, Teoria e desenvolvimento,** São Paulo: Editora Érica, 2009.
- [6] MACHADO, F. N. R. **Tecnologia e projeto em Data Warehouse.** São Paulo: Editora Érica, 2007.
- [7] KIMBAL, R. **Data Warehouse Toolkit.** Rio de Janeiro: Makron Books, 1998.
- [8] INMON, W.H. **Como Construir o Data Warehouse.** Rio de Janeiro: Campus, 1997.
- [9] TECHNET MICROSOFT. **Sobre Cubos OLAP.** Disponível em <[https://technet.microsoft.com/pt-br/library/hh916536\(v=sc.12\).aspx](https://technet.microsoft.com/pt-br/library/hh916536(v=sc.12).aspx)>. Acesso em 24 de março de 2016.
- [10] PENTAHO. **Página oficial da plataforma.** Disponível em <www.pentaho.com>. Acesso em 28. de março de 2016.