**Ciência de Dados aplicada no Anuário Estatístico da Marinha**

**Autoria**: C-ApA-IM 2022 – 013

**RESUMO**

Este trabalho buscou verificar a utilização de um ambiente de Ciência de Dados em prol do aperfeiçoamento da forma de coleta e visualização de dados do Anuário Estatístico da Marinha (ANEMAR), cuja Organização Militar responsável é a Diretoria de Administração da Marinha. Para tal, objetivou-se entender os requisitos necessários para o desenvolvimento de um ambiente de Ciência de Dados, assim como compreender o processo de elaboração do ANEMAR. Após análise da bibliografia, das documentações e de outros dados coletados, buscou-se um modelo adequado para o amadurecimento do ambiente de Ciência de Dados na Marinha do Brasil. Por conseguinte, foram detalhados alguns componentes conceituais, como as arquiteturas dos sistemas transacionais e analíticos, bem como demonstrou-se uma proposta de *Dashboard* e uma síntese da estruturação do *Data Warehouse* para o ANEMAR.

**Palavras-chave**: ANEMAR. Ciência de Dados. *Dashboard*. *Data Warehouse*.

**1 INTRODUÇÃO**

 O avanço contínuo da tecnologia da informação gera uma quantidade massiva de informações, que devem ser, portanto, armazenadas em grandes bancos de dados. Diversas tecnologias existentes, como computadores, telefones celulares e sistemas de gerenciamento de banco de dados, permitiram a proliferação de vastos repositórios de informações que necessitam de análise, processamento e transformação adequados em dados úteis, aptos a gerar conhecimento e eficiência. Nesse contexto, a capacidade de explorar de forma eficaz tais dados tornou-se um fator diferencial para as organizações, popularizando-se, assim, o termo “Ciência de Dados” (PROVOST; FAWCETT, 2013).

Nessa conjuntura, a Diretoria de Administração da Marinha (DAdM) é responsável por produzir o Anuário Estatístico da Marinha (ANEMAR), uma publicação que permite uma visão geral da Marinha do Brasil (MB), com informações estatísticas das atividades operativas e administrativas referentes a um determinado ano, cujas informações são coletadas pelas Organizações Militares (OM) que integram o Sistema Estatístico da Marinha (BRASIL, 2021).

O ANEMAR também apresenta séries históricas, que permitem a comparabilidade de informações e o acompanhamento da evolução, retratando variados períodos da História Naval. Com a identificação de padrões de dados e de comportamentos apresentados no ANEMAR, a Alta Administração Naval dispõe de uma ferramenta para subsidiar processos de tomada de decisão (BRASIL, 2021), de modo que se torna imperioso reconhecer que tal coleta e análise de dados é a base da Ciência de Dados, que possui como uma de suas finalidades justamente apoiar as tomadas de decisões (PROVOST; FAWCETT, 2013).

Outrossim, o Planejamento Estratégico Organizacional (PEO) da DAdM estabelece como um de seus objetivos estratégicos, intensificar a gestão administrativa das OM da MB. Para atingir esse objetivo, este é dividido em iniciativas estratégicas, podendo ser citado o aprimoramento da sistemática do ANEMAR, que é descrita como: simplificar o processo por meio da utilização de tecnologias, a fim de aprimorar a captação dos dados e apresentação das informações com foco no aperfeiçoamento dos processos de tomada de decisão (BRASIL, 2022).

Nesse sentido, após pesquisas nos repositórios de conhecimento da Marinha, verifica-se uma lacuna quanto à aplicação desses conceitos no âmbito do ANEMAR, especialmente no setor de pessoal, em que pese este gerencie mais de 83 mil militares (BRASIL, 2021).

Nessa lógica, o trabalho buscou estudar a seguinte problemática: Como aprimorar o Anuário Estatístico da Marinha (ANEMAR) utilizando-se da Ciência de Dados?

Assim, tem-se como objetivo geral do presente estudo identificar de que modo o ANEMAR pode ser aprimorado utilizando algumas ferramentas de Ciência de Dados.

Com o fito de subsidiar o alcance do propósito principal, o presente trabalho visa alcançar os objetivos intermediários listados a seguir:

a) Identificar os conceitos fundamentais para a aplicação de Ciência de Dados;

b) Compreender a atual estrutura de Ciência de Dados no processo do ANEMAR; e

c) Contribuir com possíveis melhorias passíveis de implementação para o desenvolvimento de um ambiente de Ciência de Dados do ANEMAR que melhor fundamente as decisões tomadas na MB.

A presente obra limita-se à identificação de oportunidades de aprimoramento que poderão ser encontradas nos estágios iniciais de aplicação dos conceitos de Ciência de Dados no ANEMAR. O estudo não busca realizar verificações de impactos quantitativos da aplicação de Ciência de Dados.

O trabalho será aplicado exclusivamente no escopo dos dados do setor de pessoal do ANEMAR, cuja Organização Militar responsável pela confecção é a Diretoria de Administração da Marinha, e a OM que toma decisões a partir dessas informações é a Diretoria do Pessoal da Marinha (DPM).

Quanto à organização, o trabalho foi dividido em cinco seções. Esta primeira parte trata de uma contextualização geral acerca do tema. A segunda parte, por sua vez, traz o referencial teórico, com os conceitos basilares a este estudo. Por seu turno, a terceira parte aborda a metodologia utilizada na condução do estudo, enquanto a quarta parte refere-se ao estudo de caso, em conjunto com sua análise e solução proposta. Por fim, a quinta e última parte expõe as considerações finais.

**2 REFERENCIAL TEÓRICO**

2.1 CIÊNCIA DE DADOS

 Segundo Provost e Fawcett (2013), a Ciência de Dados combina os campos da estatística, método científico, inteligência artificial (IA) e análise de dados para extrair valor dos mesmos, além de contemplar sua preparação para análise, incluindo limpeza, agregação e manipulação, de modo a possibilitar a realização de análises avançadas de dados. Desta forma, os cientistas de dados podem revisar os resultados para descobrir padrões e permitir que os líderes de negócios obtenham subsídios para tomada de decisão fundamentada em dados (ORACLE, 2022).

 Segundo Sharda, Delen e Turban (2019), para que boas decisões sejam tomadas, faz-se necessário que os dados sigam um ciclo, que pode ser explicitado da seguinte maneira: De início, os dados encontram-se em estado bruto, impuros, com diversos erros e inconsistências. Então, são extraídos, transformados e carregados (ETL) para serem analisados, com o objetivo de observar fatos e padrões. Feito isso, a informação poderá ser melhor compreendida com maior esclarecimento e domínio. Ressalta-se, entretanto, que esses dados precisam ser consistentes, relevantes e confiáveis, bem como estarem contextualizados e disponíveis (SERVICES, 2015).

Um dos grandes motivos para o destaque da Ciência de Dados, na atualidade, é o aumento do volume, variedade e velocidade dos dados gerados e armazenados pelas organizações, fenômeno também conhecido como Big Data (AGRAWAL, 2011). Segundo Laney (2001), o termo Big Data pode ser definido como um conjunto de dados de grande volume, adquiridos e processados em alta velocidade e com informações em uma extensa variedade de formatos. Contemporaneamente, tal conceito evoluiu, de modo que mais elementos foram adicionados à definição original: valor e veracidade (DAVENPORT, 2014).

Segundo os Ghavami (2020), as particularidades do Big Data demandam uma abordagem personalizada em relação à maneira como tipicamente os dados de interesse das organizações são tratados. Para esse fim, é necessário que sejam utilizados métodos e empregadas tecnologias próprias para análise e armazenamento de grandes volumes de dados, a fim de obter valor por meio de *insights* obtidos a partir da manipulação desses dados.

2.2 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS (DCBD)

 Com origem no termo em inglês KDD (*Knowledge Discovery in Databases*), Descoberta de Conhecimento em Base de Dados (DCBD) foi formalizado em 1989 por Piatetsky-Shapiro como um processo de procurar conhecimentos úteis a partir de grandes volumes de dados (HAN; PEI; KAMBER, 2012). Trata-se de um conceito amplo, que se concentra no desenvolvimento de métodos e técnicas para dar sentido aos dados.

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), tal processo é constituído de cinco fases:

a) Seleção e limpeza: removem-se os dados irrelevantes ou inconsistentes;

b) Pré-processamento: combinam-se múltiplas fontes de dados;

c) Transformação: consolidam-se dados apropriadamente para mineração;

d) Mineração de dados: principal atividade, pela qual se utilizam algoritmos para extrair padrões a partir dos dados obtidos; e

e) Apresentação e avaliação de conhecimentos: utilizam-se técnicas de visualização para apresentar os conhecimentos aos usuários.

**Figura 1 - Fases do KDD**



Fonte: Adaptado de Fayyad, Piatetsky-Shapiro, Smyth (1996).

Destarte, pode-se afirmar que o KDD trata-se tanto de um processo iterativo, pois poderá ser repetido quantas vezes forem necessárias para a busca de melhores resultados, como também um processo interativo, dada a participação de outros agentes, como o usuário final (HAN; PEI; KAMBER, 2012).

**2.2.1 Mineração de Dados**

Enquanto KDD refere-se ao processo geral de descoberta de conhecimento útil a partir de dados, mineração de dados é uma das etapas desse processo em que são utilizados métodos específicos para a extração de padrões (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

 Segundo Han, Pei e Kamber (2012), a mineração de dados pode ser organizada em seis fases, ou etapas, que compõem o chamado CRISP-DM (*Cross-Industry Standard Process of Data Mining*):

a) Compreensão dos Negócios: o conhecimento sobre o negócio e sobre os seus processos inerentes é de fundamental importância para que se definam os objetivos da mineração de dados;

b) Entendimento dos Dados: os dados devem ser descritos de maneira clara e objetiva, sempre explicitando as diversas fontes de obtenção e eventuais comportamentos de interdependência entre as variáveis;

c) Preparação dos Dados: análises preliminares dos dados, com eventuais tratamentos sobre os erros e inconsistências, podem ser de grande utilidade para que os métodos de *data mining* sejam aplicados corretamente;

d) Modelagem: diversas técnicas podem ser aplicadas, como a elaboração de métodos de mineração de dados, sempre com base nos objetivos propostos;

e) Análise dos Resultados: nesta etapa, é de fundamental importância que haja a participação dos conhecedores do negócio, de estatísticos e especialistas nos dados em questão, a fim de que sejam elaboradas avaliações sobre os achados na etapa anterior, a partir da análise de testes e validações; e

f) Divulgação dos Resultados: após modelagem e análise, é necessário que todos os envolvidos tomem ciência dos resultados encontrados, de modo a possibilitar a implantação de procedimentos de gestão.

 Considerando este modelo proposto por Fayyad, Piatetsky-Shapiro, Smyth (1996) e corroborado por Han, Pei e Kamber (2012), pode-se ratificar que a mineração de dados é capaz de gerar conhecimentos a partir dos mesmos. À vista disso, faz-se necessário que a organização possua dados à sua disposição. Assim, serão também esclarecidos alguns conceitos de um ambiente adequado para a realização da análise de dados.

2.3 BANCOS DE DADOS E *DATA WAREHOUSES*

 Segundo Elmasri e Navathe (2018), os bancos de dados tradicionais possuem suporte para o processamento de transações *online* (OLTP, *Online Transaction Processing*), possibilitando inserções, atualizações e exclusões, além de terem suporte para algum nível de consulta de informações. Desta forma, os OLTP, por possuírem estrutura normalizada e por serem otimizados para o processamento de muitas transações de inserções, atualizações e exclusões, não são adequados para consultas gerenciais, de maneira que para essa forma de análise faz-se necessário outro tipo de modelagem, e, portanto, não podem ser otimizados para OLAP (*Online Analytical Processing*), ou mineração de dados (INMON, 1997). É nesse contexto que se apresenta o *Data Warehouse*.

De modo geral, um *Data Warehouse* pode ser descrito como uma coleção de tecnologias de apoio à decisão, visando subsidiar os agentes de uma organização a tomarem decisões melhores e mais rápidas (ELMASRI; NAVATHE, 2018). Segundo Kimball e Ross (2013), um *Data Warehouse* é utilizado para armazenar informações relativas às atividades de uma organização de forma consolidada, além disso, possibilita a análise de grandes volumes de dados que são coletados a partir dos sistemas transacionais (OLTP).

Um *Data Warehouse*, comumente, trata-se de um depósito de dados integrados de múltiplas fontes, processados para armazenamento em um modelo multidimensional, sendo capazes de apoiar a análise de série temporal e tendência, ambas exigindo mais dados históricos do que usualmente é mantido nos bancos de dados transacionais (ELMASRI; NAVATHE, 2018). Assim, os *Data Warehouses* são modelados justamente para possibilitar a extração, o processamento e a apresentação eficientes para fins analíticos e de tomada de decisão (INMON, 1997).

**Figura 2 - Arquitetura geral de um *Data Warehouse***



Fonte: Adaptado de Han, Pei, Kamber (2012)

O esquema acima resume o processo integral de um *Data Warehouse*, incluindo a possível limpeza e reformatação dos dados antes que sejam carregados, processo esse tratado por ferramentas conhecidas como ferramentas de ETL (do inglês *Extraction, Transformation, Load* - extração, transformação e carga) (KIMBALL; ROSS, 2013). Nesse cenário, observa-se que os sistemas OLAP e a mineração de dados podem produzir não apenas novas informações relevantes, mas também podem incluir outras entradas de dados, como arquivos, por exemplo (ELMASRI; NAVATHE, 2018).

Nessa perspectiva, cabe ressaltar algumas características relevantes dos *Data Warehouses*, segundo Elmasri e Navathe (2018):

a) Organização: organizar por assunto utilizando metadados para tal, além de conter somente as informações mais importantes para o apoio à decisão;

b) Consistência: a codificação dos dados é altamente consistente porque estes são padronizados;

c) Variedade de tempo: armazena os dados em média de 5 (cinco) a 10 (dez) anos, de modo a possibilitar a realização de análises históricas;

d) Não-volatilidade: dados servem somente para leitura;

e) Estrutura relacional;

f) Arquitetura cliente-servidor: tem-se maior facilidade de acesso pelo usuário, possibilidade de acesso via intranet.

Ampliando a compreensão sobre o assunto, Kimball e Ross (2013), definem alguns requisitos fundamentais para um *Data Warehouse* duradouro:

a) o *Data Warehouse* deve fazer com que as informações do negócio sejam facilmente acessadas, seu conteúdo deve ser compreensível;

b) o *Data Warehouse* deve apresentar as informações do negócio de modo estável;

c) o *Data Warehouse* deve ser adaptável e flexível a mudanças;

d) o *Data Warehouse* deve ser seguro, as informações que forem confidenciais devem ser protegidas;

e) o *Data Warehouse* deve servir de base para melhores tomadas de decisões;

f) o *Data Warehouse*, para ser considerado um sucesso, deve ser utilizado ativamente pela comunidade do negócio.

Dentre as diferentes formas de implementação de sistemas analíticos (OLAP), ressaltam-se duas fundamentais: os *Data Warehouses* e os *Data Marts*. Segundo Elmasri e Navathe (2018), ambos têm o objetivo de organizar dados úteis para o apoio à análise e posteriormente a tomada de decisões, além de ambos utilizarem-se da mesma tecnologia de sistemas gerenciadores de bancos de dados.

Ainda segundo os autores supracitados, a diferença entre *Data Warehouse* e *Data Mart* consiste na finalidade de seus usos. Enquanto um *Data Warehouse* contempla a consolidação de dados e análises de toda a organização, acumulando dados de diferentes origens e para diferentes finalidades, um *Data Mart* restringe-se a uma determinada área de assunto e apresenta informações mais detalhadas sobre o mercado ou departamento em questão.

Assim, um *Data Mart* pode ser formado de duas maneiras, segundo Han, Pei, Kamber (2012):

a) Capturando dados diretamente de sistemas transacionais, de maneira que cada *Data Mart* busca as informações relevantes para o seu mercado;

b) Capturando dados de todos os sistemas transacionais em um *Data Warehouse* central, que, por sua vez, alimenta todos os *Data Marts*.

**2.3.2 Modelagem de dados para *Data Warehouses***

De acordo com Elmasri e Navathe (2018), modelo de dados pode ser caracterizado como:

[...] uma coleção de conceitos que podem ser usados para descrever a estrutura de um banco de dados — oferece os meios necessários para alcançar essa abstração. Com estrutura de um banco de dados, queremos dizer os tipos, os relacionamentos e as restrições que se aplicam aos dados. A maioria dos modelos de dados também inclui um conjunto de operações básicas para especificar recuperações e atualizações no banco de dados. (ELMASRI; NAVATHE, 2018, p. 30).

Segundo Han, Pei, Kamber (2012), um *Data Warehouse* é modelado em uma arquitetura multidimensional, chamada de cubo de dados, em que cada dimensão corresponde a um atributo da estrutura dos dados, e cada célula guarda o valor de alguma medida de agregação, como uma soma ou contagem. Para dados que se seguem à formatação dimensional, o desempenho da consulta nas matrizes multidimensionais pode ser muito melhor que no modelo de dados relacional, permitindo a visualização de dados de interesse de uma forma muito mais rápida (ELMASRI; NAVATHE, 2018).

 Segundo Elmasri e Navathe (2018), o modelo multidimensional é composto por dois tipos de tabelas, quais sejam as tabelas de dimensão e as tabelas de fatos. Estas contém os dados de interesse do negócio e os dados que identificam as dimensões, enquanto aquelas consistem em linhas com os atributos da dimensão. Cita-se como um dos esquemas mais comuns de modelos multidimensionais o esquema em estrela (SHARDA; DELEN; TURBAN, 2019).

**Figura 3 - Esquema em estrela**



Fonte: Sharda, Delen e Turban (2019).

 De forma a possibilitar a construção desses modelos e aquisição de dados para o *Data Warehouse*, Elmasri e Navathe (2018) elencam as seguintes etapas a serem seguidas:

a) Os dados devem ser extraídos de diversas fontes, como por exemplo, bancos de dados ou outras entradas de dados, tais quais as que contenham dados de pessoal ou dados de pagamentos.

b) Os dados necessitam ser formatados de modo coerente dentro do *Data Warehouse*, isto é, as inconsistências de formatos devem ser corrigidas.

c) Os dados devem ser limpos para que possam ser asseguradas sua validade e confiabilidade. Essa limpeza é um processo complexo e complicado e comumente considerado como a parte do processo que exige mais trabalho na construção de um *Data Warehouse*.

d) Os dados devem ser adequados ao modelo de dados do *Data Warehouse*. Os dados que se originaram de diversas fontes devem ser retratados no modelo de dados do *Data Warehouse*, de modo que por vezes haverá a necessidade de sua conversão de bancos de dados relacionais (OLTP) para um modelo multidimensional.

e) Os dados devem ser carregados no *Data Warehouse*. O grande volume de dados faz com que essa seja uma tarefa significativa. Para isso, são necessárias ferramentas de monitoramento para as cargas.

**3 METODOLOGIA DA PESQUISA**

3.1 CLASSIFICAÇÃO DA PESQUISA

Segundo Prodanov e Freitas (2013), a Pesquisa Científica visa conhecer cientificamente um ou mais aspectos de determinado assunto. Para tanto, o autor define que uma pesquisa pode ser classificada quanto a sua natureza, seu objetivo e o procedimento a ser utilizado.

Nesse cenário, foi realizada uma pesquisa aplicada, uma vez que objetiva gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos (PRODANOV; FREITAS, 2013). Para atingir o objetivo deste trabalho, foi realizada uma pesquisa de caráter descritivo.

De acordo com Gil (2008), quanto aos procedimentos técnicos utilizados, trata-se de uma pesquisa bibliográfica, pois foi realizada por fontes secundárias, como livros e artigos, pesquisa documental, caracterizada pela utilização de dados primários, como normas e portarias internas da MB, estudo de caso, que tem o objetivo de descrever e explorar uma situação real da MB, além de entrevistas.

Referente à forma de abordagem utilizada neste estudo, esta é classificada como uma pesquisa qualitativa, com o objetivo de elaborar considerações sobre o tema em análise, intencionando a qualificação dos dados coletados. (PRODANOV; FREITAS, 2013).

De forma a buscar compreender o processo do ANEMAR, foi realizada pesquisa documental em normas e publicações da MB sobre a temática, bem como entrevistas junto aos agentes envolvidos na confecção do ANEMAR. Ademais, para obtenção de conceitos teóricos e casos de aplicação prática, foi realizada uma pesquisa bibliográfica em livros e artigos, de forma a incrementar os entendimentos acerca do objeto de estudo.

Referente ao caso estudado, foram entrevistados os agentes responsáveis pela confecção do ANEMAR na DAdM, bem como os agentes responsáveis pela coleta de dados em um Centro de Coleta de Dados (CCD), nesse caso a DPM, com o propósito de subsidiar uma melhor compreensão acerca do processo de confecção do ANEMAR desde o início da obtenção dos dados na origem até a sua disponibilização para visualização e fundamentação das tomadas de decisões. Tais entrevistas foram realizadas de forma não estruturada para melhor explorar o assunto (LAKATOS; MARCONI, 2017).

No que tange ao tratamento dos dados, foi analisado o processo de confecção do ANEMAR a partir das entrevistas, sob a ótica dos conceitos de Ciência de Dados compreendidos a partir da pesquisa bibliográfica e documental. Em seguida, a partir do estudo de caso junto a DAdM e DPM, foram analisados quais dados e a forma que se utilizam os mesmos para tomada de decisão. Por fim, foram comparadas essas informações, de modo a obter-se um entendimento acerca de como o ANEMAR pode ser aprimorado a partir do caso estudado e alinhado aos conceitos teóricos de Ciência de Dados.

3.2 ESTUDO DE CASO

**3.2.1 Confecção do ANEMAR**

 A DAdM é uma Diretoria Especializada subordinada à Secretaria-Geral da Marinha, com área de atuação atinente à superintendência das atividades de administração geral da MB. Dentre suas missões, essa Diretoria é responsável por “supervisionar e executar as atividades pertinentes ao Sistema Estatístico da Marinha” (BRASIL, 2022, p. 3). Nesse contexto, emerge o ANEMAR, cujo Departamento de Estatística é responsável pelo cumprimento de seu processo.

 Assim, foram realizadas entrevistas com os profissionais da área de Estatística da DAdM, além de visitas *in loco*, objetivando observar a estrutura e compreender o funcionamento do processo de confecção do ANEMAR, cujas etapas podem ser ilustradas da seguinte forma: (i) coleta dos dados; (ii) tratamento e organização dos dados; e (iii) disponibilização para visualização dos dados.

 Para o processo de coleta, a DAdM possui modelos de Mapas de Apuração (planilhas em formato de *Excel*), que são planilhas padronizadas disponibilizadas aos Centro de Coleta de Dados (CCD) – OM, que centralizam os dados por área e também são responsáveis pela distribuição e compilação das informações dos mapas pelas suas OM subordinadas. Tais planilhas são preenchidas de forma manual pelas diversas OM e enviadas aos CCD, que as recebem posteriormente e verificam as informações que lhes foram encaminhadas com base em alguns controles. Por fim, os CCD supramencionados enviam as planilhas preenchidas à DAdM.

 Já quanto ao processo de tratamento e organização dos dados, é responsável o setor de estatística da DAdM. Os dados enviados pelas CCD são conferidos a fim de verificar quaisquer inconsistências perceptíveis, como por exemplo erros de digitação ou de valores de somatórios. Em seguida, esses dados são corrigidos ou encaminhados de volta às CCD para correção a depender do tipo de inconsistência. Não havendo mais erros perceptíveis, os arquivos que foram recebidos são agrupados por CCD.

No processo de visualização dos dados, os arquivos dos CCD são compilados em um PDF por meio de um macro que une os arquivos no formato de *Excel* por assuntos que serão exibidos na publicação do ANEMAR. Tal visualização é padronizada por alguns tipos de análises com dimensões pré-definidas, formando-se, assim, as tabelas e os gráficos que constam no ANEMAR.

**3.2.2 Coleta de dados na DPM**

 Ainda sobre o caso do ANEMAR, com o intuito de verificar a origem dos dados, aprofundou-se no caso específico da Diretoria do Pessoal da Marinha (DPM), que é a OM responsável por planejar, coordenar e controlar as atividades relacionadas com o pessoal da MB.

 Essa diretoria possui diversos sistemas transacionais, que servem para auxiliar no cumprimento de suas diversas atribuições, de maneira que os dados necessários para subsidiar a confecção do ANEMAR estão contidos nesses sistemas.

Assim, considerando que as informações solicitadas pela DAdM para confeccionar o ANEMAR são padronizadas, a DPM criou um sistema que coletasse os dados automaticamente de seus diversos sistemas transacionais. No entanto, há um sistema apenas para a coleta de dados pertinentes ao ANEMAR nesta OM, de tal maneira que o sistema não é utilizado para gerenciamento ou armazenagem dos dados, mas tão somente para automatizar a coleta e a compilação dos dados de pessoal.

 Este sistema acessa os já citados sistemas transacionais da DPM na busca dos dados que são necessários para subsidiar o ANEMAR, gerando um arquivo que contém as planilhas que irão compor o Mapa de Apuração da DPM.

**4 ANÁLISE DOS DADOS E PROPOSTA DE SOLUÇÃO**

4.1 ANÁLISE DO AMBIENTE ATUAL

 O estudo de caso focou em como aprimorar o processo realizado pelo Setor de Estatística da DAdM, considerando desde a etapa de coleta dos dados até a disponibilização para a visualização.

 Inicialmente, verificaram-se as compatibilidades entre o caso estudado e a literatura acerca do tema. Dado que (i) o ANEMAR propõe-se a coletar dados históricos para subsidiar a tomada de decisão e que (ii) um *Data Warehouse* serve para armazenar dados para posterior consulta por meio de um OLAP, que subsidiará a tomada de decisão, foi então sintetizada uma maneira de estruturar conceitualmente um sistema em que fosse possível a execução de tais tarefas de forma aprimorada e capaz de suportar um maior volume de dados.

Analisando o ambiente atual da DAdM, tem-se que os dados são obtidos a partir da compilação de planilhas enviadas pelos CCD, que, por sua vez, compilam os subsídios das OM subordinadas. Tal procedimento, por vezes, acaba sendo demorado e incorrendo em risco de informações incorretas ou inconsistentes.

Conclui-se, assim, que a DAdM não possui uma ferramenta de ETL para a extração e tratamento dos dados dos sistemas OLTP. Igualmente, a DAdM não possui um ambiente OLAP para análise desses dados e tampouco utiliza um *Dashboard* como ferramenta de visualização mais facilitada e interativa das informações obtidas. Nesse contexto, serão abordados cada um desses tópicos a seguir: estrutura OLTP, estrutura OLAP, ETL, *Dashboard* interativo e proposta de arquitetura de *Data Warehouse* para o ANEMAR.

4.2 OLTP

 A partir das informações obtidas junto a DAdM e DPM, pode-se observar que as fontes primárias dos dados que compõem o ANEMAR são, em sua maior parte, dados que se encontram nos sistemas transacionais da MB, que são modelados a partir de estruturas relacionais normalizadas.

 Desta forma, com o intuito de acessar tais dados diretamente para a utilização das demais ferramentas de Ciência de Dados, como o *Data Warehouse* e o *Dashboard*, foi elaborado um modelo conceitual para exemplificar como se encontram os dados na fonte para que o setor de pessoal compile e encaminhe para a DAdM por meio de planilhas. Tal modelo encontra-se a seguir:

**Figura 4 - Modelo relacional de um OLTP**



Fonte: Elaborado pelo autor (2022).

4.3 OLAP

Conforme explicitado anteriormente, para originar os dados que compõem o ANEMAR utilizam-se planilhas, em que pese haja potencial de serem obtidos diretamente dos sistemas transacionais da MB ou de eventuais *Data Marts* dos CCD ou de outras OM. Foi estruturada uma modelagem para suportar essas diversas origens, de maneira que, no caso estudado, foram utilizados os dados da DPM como exemplo de um dos esquemas em estrela para compor o *Data Warehouse* do ANEMAR.

 O esquema em estrela proposto é um modelo relacional, que, no entanto, diferencia-se dos esquemas relacionais normalmente utilizados num OLTP por ser desnormalizado. Assim, pode-se afirmar que o modelo em análise é uma forma que possibilita a consulta de dados mais simples e ágil, bem como uma interpretação acessível das informações, com maior flexibilidade (SHARDA; DELEN; TURBAN, 2019)

Outrossim, o foco principal é a tabela fato, por meio da qual serão extraídos os objetos cuja análise é pretendida e, os valores em si, que são cruzados na tabela e complementados com dados das tabelas dimensões. Para o caso exemplificado, construiu-se um modelo em estrela que possibilitasse consultas que suprissem a maior parte dos dados que comumente são exibidos no ANEMAR, na seção de pessoal. Tal modelo encontra-se a seguir:

**Figura 5 - Modelo em estrela do *Data Warehouse* proposto**



Fonte: Elaborado pelo autor (2022).

4.4 ETL

 De modo a coletar esses dados de forma mais confiável e eficiente, propõe-se também o uso de uma ferramenta de ETL, com o fim de diminuir a inserção manual dos dados que possuem o potencial de obtenção diretamente de sistemas transacionais ou mesmo de outros *Data Warehouses* ou *Data Marts*. A ferramenta em questão possibilita a aquisição de dados em periodicidades mais frequentes do que a anual, caso seja de interesse, podendo acompanhar a evolução mensal dos dados, por exemplo.

 Para este estudo foram utilizadas as próprias funções de ETL do *software Power BI*® da Microsoft.

4.5 *DASHBOARD*

 Dando sequência na lógica da estrutura de uma nova sistemática para o ANEMAR, tem-se o *Data Warehouse* formado a partir dos dados coletados das diversas fontes, tratados por meio do ETL. Neste momento, o processo retorna valor para a instituição, na medida em que os dados são apresentados e analisados e deles podem ser extraídos conhecimentos (GHAVAMI, 2020).

 Isto posto, para possibilitar a devida análise dos dados conforme a necessidade do tomador de decisão, a apresentação dos mesmos torna-se ainda mais acessível e vantajosa quando esses dados podem ser filtrados e cruzados a partir da necessidade do cliente. Para este estudo foi utilizado o *software* *Power BI*®, com o intuito de exemplificar o potencial de aplicação de ferramentas análogas, em que pese tal ferramenta de análise de dados possa ser encontrada no mercado em diversas outras opções.

 Ressalta-se que para a construção deste modelo, não foram utilizados os dados reais individualizados por militar pela impossibilidade de sua obtenção junto a DPM devido a questões de proteção de dados. Entretanto, com o intuito de demonstrar a possibilidade de uso da ferramenta em dados com as mesmas características e estrutura, foram utilizados dados simulados, porém respeitando o tipo de relacionamento entre eles, conforme estabelecido no item 4.2.

Dessa forma, foram elaborados modelos de tabelas que expressassem, pelo menos, as mesmas informações que o ANEMAR já se dispunha a publicar. No entanto, com o *Dashboard* é possível gerar diversas outras visualizações:

**Figura 6 - Modelo de *Dashboard* proposto**



Fonte: Elaborado pelo autor (2022).

Nesse *Dashboard* proposto tem-se:

a) tabela de quantidade de militares por posto e área de localização;

b) tabela de quantidade de militares por Setor de Distribuição de Pessoal e posto; e

c) mapa de quantidade de militares por área de localização

No *Dashboard* apresentado, pode-se gerar qualquer consulta que seja de interesse apenas clicando em seus filtros, conforme demonstra-se a seguir:

**Figura 7 - Exemplo de consulta no *Dashboard***



Fonte: Elaborado pelo autor (2022).

Neste exemplo de consulta, foram filtrados apenas os militares do 1º Distrito Naval (área de localização) do sexo masculino. Nesse sentido, nota-se que as duas tabelas e o mapa adaptam-se à consulta solicitada.

Assim, é possível notar nos painéis a possibilidade de obtenção de um panorama geral da MB nos mais diversos setores, assim como a especificação das visualizações por buscas em grupos de interesse específicos. Essas visualizações são possíveis devido à estrutura de um OLAP e pelo fato de seus dados serem imutáveis e se concatenarem por meio de uma tabela fato, de forma que suas consultas se fazem de maneira muito mais ágil.

4.6 PROPOSTA DE ARQUITETURA DE *DATA WAREHOUSE*

 De modo a possibilitar a descoberta de conhecimentos e mineração de dados nas bases de dados do ANEMAR, faz-se necessária uma estrutura de dados robusta, modelada e projetada, que seja capaz de realizar a armazenagem e o gerenciamento de extenso volume de dados coletados das variadas fontes que compõem o ANEMAR. Assim, apresenta-se a proposta de modelo para um *Data Warehouse* personalizado ao caso estudado abaixo:

**Figura 8 - Arquitetura do sistema de *Data Warehouse* proposto**



Fonte: Elaborado pelo autor (2022).

 Por todo o exposto, fica evidente que o ANEMAR poderia ser aprimorado com a utilização de um *Data Warehouse* que compilasse os diversos dados produzidos pela MB. Esse *Data Warehouse* poderia ser carregado com dados obtidos de algumas formas, como por exemplo:

i) junto aos sistemas transacionais (OLTP) da MB;

ii) junto aos *Data Warehouses* ou *Data Marts* de outras OM;

iii) alimentado manualmente por encaminhamento de planilhas. Ressalta-se que esse método de carregamento manteria a forma de coleta atualmente praticada e possuiria como vantagem sua especial utilidade nos casos de transição, em que não há a disponibilização do acesso aos dados pelas formas anteriores.

 Assim, a DAdM pode possuir um *Data Warehouse* para armazenar as informações relativas a MB como um todo. Por conseguinte, os dados seriam não apenas mais confiáveis, mas obtidos de maneira mais ágil, quando adquiridos diretamente de sistemas transacionais ou outros *Data Warehouses* ou *Data Marts*.

Ressalta-se, ainda, que por meio do modelo proposto subsistiria ainda a possibilidade de continuação da forma de coleta de dados por meio de planilhas enviadas pelos CCD, além de tornar possível a comparação dos dados históricos em diferentes períodos de interesse com um *Data Warehouse,* como demonstrado no *Dashboard*.

**5 CONSIDERAÇÕES FINAIS**

O trabalho buscou elucidar os conceitos de Ciência de Dados e contribuir com proposições de aperfeiçoamento dos subsídios à Alta Administração Naval para tomada de decisões fundamentadas em dados. Além disso, buscou-se cumprir o objetivo deste trabalho em alinhamento com o PEO da DAdM.

 Assim, o presente estudo teve por fito a compreensão dos requisitos necessários para o uso de e a possibilidade de aplicação de um ambiente de Ciência de Dados em prol de auxiliar a DAdM no processo do ANEMAR. Para que este objetivo fosse alcançado, foram realizadas algumas etapas. De início, foi verificada a relevância do estudo do tema, percebendo-se assim que havia uma carência de estudos afetos a essa área, especialmente quanto a suas aplicações no ANEMAR.

Em seguida, foram realizados estudos na literatura, em busca de sintetizar os conceitos fundamentais afetos à Ciência de Dados, de modo a averiguar sua possível aplicabilidade no processo do ANEMAR, atingindo, assim, o primeiro objetivo específico.

Posteriormente, foi realizado um estudo de caso para melhor compreender os detalhes do processo de confecção do anuário estatístico, desde a coleta de dados na fonte primária até a disponibilização para visualização das informações, com o propósito de compreender a atual estrutura de Ciência de Dados no ANEMAR, cumprindo, então, o segundo objetivo específico.

Por conseguinte, a partir dos conhecimentos adquiridos com o estudo da bibliografia e o estudo de caso, prosseguiu-se com a aplicação de ferramentas de Ciência de Dados nas diversas etapas do ANEMAR. Assim, foram elaborados: (i) um modelo relacional de um OLTP compatível com a estrutura de dados encontrados na DPM; (ii) um modelo em estrela, gerado a partir da estrutura de dados de pessoal, de modo a realizar a conversão de um modelo relacional para um modelo multidimensional; (iii) uma breve descrição da fase de ETL; (iv) uma proposta de *Dashboard*, na qual fosse possível a demonstração de mais possibilidades de visualizações das informações; e (v) uma proposta de arquitetura de *Data Warehouse* que fosse adequada para a realidade da DAdM e capaz de operar com as diversas fontes de dados que subsidiam o ANEMAR. Denota-se, assim, que o terceiro objetivo foi cumprido pela proposta de possíveis melhorias passíveis de realização para que seja desenvolvido um ambiente adequado de Ciência de Dados do ANEMAR.

Finalmente, foi atingido o objetivo geral do presente trabalho: identificar como o ANEMAR pode ser aprimorado utilizando algumas ferramentas de Ciência de Dados. Como limitações da pesquisa destaca-se a impossibilidade de trabalhar com informações da base de dados reais, sendo necessária a utilização de dados simulados com estrutura semelhante.

 Por fim, como sugestões de trabalhos futuros, sugere-se:

a) A utilização da base de dados reais ou simulada para mineração de dados, utilizando *Machine Learning*;

b) A utilização de ferramentas de *Business Intelligence* no processo do ANEMAR; e

c) O aprimoramento de um *Dashboard* para o ANEMAR que contemple mais setores, isto é, para além do setor de pessoal.

**REFERÊNCIAS**

AGRAWAL, Divyakant et al. **Challenges and opportunities with Big Data 2011-1**. 2011.

BRASIL. Marinha do Brasil. Diretoria de Administração da Marinha. **Anuário Estatístico da Marinha.** 49ª edição. Rio de Janeiro, RJ, 2021.

BRASIL. Marinha do Brasil. Diretoria de Administração da Marinha. **Planejamento Estratégico Organizacional.** Rio de Janeiro, RJ, 2022.

DAVENPORT, T. H. **Big Data at work**: dispelling the myths, uncovering the opportunities. Harvard Business Review Press, 2014.

ELMASRI, Ramez; NAVATHE, Shamkant. **Sistemas de Banco de Dados**. 6. ed. São Paulo: Pearson, 2011. 770 p.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From Data Mining to Knowledge Discovery in Databases**. Ai Magazine, Providence, v. 17, n. 3, p.37-54, 1996.

GHAVAMI, P. **Big Data Analytics Methods**: Analytics Techniques in Data Mining, Deep Learning and Natural Language Processing. 2nd ed. Walter de Gruyter GmbH & Co KG; 2020.

GIL, A. C. **Métodos e Técnicas de Pesquisa Social**. 6. ed. São Paulo: Atlas, 2008.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data Mining**: Concepts and Techniques.3. ed. San Francisco, CA, USA: Elsevier, 2011. 744 p.

INMON, William H. **Como Construir o Data Warehouse**. 2a. Ed., Editora Campus, Brasil, 1997.

KIMBALL, Ralph; ROSS, Margy. **Data warehouse toolkit**: the definitive guide to dimensional modeling. 3. ed. Indiana: Wiley, USA, 2013.

LANEY, D. **Application delivery strategies**. META Group, Stamford, 2001. Disponível em: https://www.academia.edu/37216189/Application\_Delivery\_Strategies. Acesso em: 26 set. 2022.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. Atlas, 2017.

ORACLE. **O que é Ciência de Dados?** 2022. Disponível em: https://www.oracle.com/br/data-science/what-is-data-science.html. Acesso em: 26 set. 2022.

PRODANOV, C. C.; FREITAS, E. C. **Metodologia do Trabalho Científico**: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico. Novo Hamburgo: Universidade Feevale, 2013.

PROVOST, Foster; FAWCETT, Tom. **Data Science for Business**: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc., 2013.

SERVICES, EMC Education. **Data Science and Big Data Analytics**: Discovering, Analyzing, Visualizing and Presenting Data. John Wiley & Sons, 2015.

SHARDA, R.; DELEN, D.; TURBAN, E. **Business Intelligence e Análise de Dados para**

**Gestão do Negócio**. 4. ed. Porto Alegre: Bookman, 2019.