INSTITUTO UNIVERSITÁRIO MILITAR

DEPARTAMENTO DE ESTUDOS PÓS-GRADUADOS

CURSO DE PROMOÇÃO A OFICIAL SUPERIOR

2021/2022, 2ª EDIÇÃO

PROJETO DE INVESTIGAÇÃO - TIFC

GESTÃO E ESTRUTURAÇÃO DE *BIG DATA*

O TEXTO CORRESPONDE A TRABALHO FEITO DURANTE A FREQUÊNCIA DO CURSO NO IUM SENDO DA RESPONSABILIDADE DO SEU AUTOR, NÃO CONSTITUINDO ASSIM DOUTRINA OFICIAL DAS FORÇAS ARMADAS PORTUGUESAS OU DA GUARDA NACIONAL REPUBLICANA.

António Guerreiro Pacheco

PRIMEIRO-TENENTE ST-EINF

INSTITUTO UNIVERSITÁRIO MILITAR
DEPARTAMENTO DE ESTUDOS PÓS-GRADUADOS

GESTÃO E ESTRUTURAÇÃO DE *BIG DATA*

1TEN ST-EINF António Guerreiro Pacheco

Trabalho de Investigação Individual do CPOS 2021/2022 – 2ª Edição

Pedrouços 2022

INSTITUTO UNIVERSITÁRIO MILITAR
DEPARTAMENTO DE ESTUDOS PÓS-GRADUADOS

GESTÃO E ESTRUTURAÇÃO DE *BIG DATA*

1TEN ST-EINF António Guerreiro Pacheco

Trabalho de Investigação Individual do CPOS 2021/2022 – 2ª Edição

Orientador: CFR EN-AEL Pedro Luís Araújo Costa

Pedrouços 2022

Declaração de compromisso Antiplágio

Eu, **António Guerreiro Pacheco**, declaro por minha honra que o documento intitulado **"Gestão e estruturação de *Big Data*"** corresponde ao resultado da investigação por mim desenvolvida, enquanto auditor do **Curso de Promoção a Oficial Superior, 2021/2022, 2ª edição,** no Instituto Universitário Militar, e que é um trabalho original, em que todos os contributos estão corretamente identificados em citações e nas respetivas referências bibliográficas.

Tenho consciência que a utilização de elementos alheios não identificados constitui grave falta ética, moral, legal e disciplinar.

Pedrouços, 08 de julho de 2022.

Assinado por: **ANTÓNIO GUERREIRO PACHECO**
Num. de Identificação: 11204255
Data: 2022.07.27 11:38:40+01'00'

António Guerreiro Pacheco
1TEN ST-EINF

Agradecimentos

O desenvolvimento deste trabalho, enriquecedor na perspetiva do conhecimento, e motivador, pela minha experiência passada e pela temática potencialmente inovadora e contribuidora para a melhoria das Tecnologias de Informação e Comunicação na Marinha Portuguesa, não é um ato isolado, mas sim a incorporação do contributo e do saber de diversas fontes, a quem reitero o meu sincero agradecimento.

Ao meu orientador, Capitão-de-fragata Araújo Costa, pela disponibilidade que demonstrou ao longo de todo o trabalho e apoio durante a estruturação e delimitação do tema.

Ao Capitão-tenente Gonçalves Deus pelo seu aconselhamento técnico que tanto contribuiu para o desenvolvimento deste trabalho.

A todos os entrevistados e àqueles que com o seu conhecimento contribuíram para o trabalho, uma palavra de apreço e um muito obrigado pela transmissão das vossas experiências que permitiram valorizar e robustecer as conclusões e resultados apresentados.

À minha família, pilar de suporte fundamental, que me apoiou ao longo deste percurso. À minha esposa, aos meus filhos, em particular, fonte de orgulho e de motivação, e à minha mãe, a quem devo o que hoje sou, e de quem tenho muitas saudades, esperando que onde estiver seja para ela motivo de orgulho redobrado.

Índice

Índice de Apêndices

Índice de Figuras

Índice de Gráficos

Índice de Quadros

Resumo

O crescimento exponencial de dados, provenientes de múltiplas fontes e a velocidades elevadas, que contribuem para "alimentar" o sistema de *Big Data* da Marinha Portuguesa, e posteriormente utilizados para a execução de processos e apoio à tomada de decisão no âmbito do Conhecimento Situacional Marítimo, torna premente a edificação de um sistema de suporte robusto, escalável e de alta *performance*.

É, assim, objetivo deste estudo propor um modelo de gestão e estruturação de *Big Data* no contexto do Conhecimento Situacional Marítimo, no que se refere a dados recebidos via *Automatic Identification System (AIS)*, seguindo um raciocínio indutivo, uma estratégia de investigação qualitativa e um desenho de pesquisa do tipo de estudo de caso.

Através da análise documental, do desenvolvimento de um protótipo com aplicação de métricas de avaliação, e do conteúdo dos dados das entrevistas semiestruturadas realizadas a 10 especialistas de reconhecido mérito de entidades públicas e privadas de renome nacional e internacional, foi proposto um modelo integrador. Modelo este, organizado por vetores de edificação de capacidades tecnológicas, tendentes à incorporação de uma infraestrutura de NoSQL para gestão e estruturação de *Big Data* relativamente a dados AIS, sendo que este tipo de solução ainda não se encontra implementada na Marinha Portuguesa.

Palavras-Chave: Gestão e estruturação, *Big Data*, NoSQL, *Automatic Identification System*, Conhecimento Situacional Marítimo.

Abstract

The exponential growth of data, from multiple sources and at high speeds, which contribute to "feeding" the Portuguese Navy's Big Data system, and subsequently used to execute processes and support decision-making within the scope of Maritime Situational Knowledge , makes it imperative to build a robust, scalable and high-performance support system.

It is, therefore, the objective of this study to propose a model for the management and structuring of Big Data in the context of Maritime Situational Knowledge, with regard to data received via the Automatic Identification System (AIS), following an inductive reasoning, a qualitative research strategy and a case study type research design.

Through document analysis, the development of a prototype with the application of evaluation metrics, and the content of data from semi-structured interviews carried out with 10 experts of recognized merit from public and private entities of national and international renown, an integrative model was proposed. This model, organized by vectors for building technological capabilities, aimed at the incorporation of a NoSQL infrastructure for the management and structuring of Big Data in relation to AIS data, and this type of solution is not yet implemented in the Portuguese Navy.

Keywords: *Managing and structuring*, *Big Data*, NoSQL, *Automatic Identification System*, *Maritime Situational Knowledge*.

Lista de abreviaturas, siglas e acrónimos

A

ACID	Atomicidade, Consistência, Isolamento e Durabilidade
AIS	*Automatic Identification System*
APRAM	Administração dos Portos da Região Autónoma da Madeira

B

BASE	Basicamente disponível, Estado leve e Eventualmente consistente
B	*Byte*
BD	Base de Dados
BI	*Business Intelligence*

C

CDD	Centro de Dados da Defesa
CN	Comando Naval
CPOS-M	Curso de Promoção a Oficial Superior – Marinha
CPU	*Central Processing Unit*
CSM	Conhecimento Situacional Marítimo

D

DAGI	Direção de Análise e Gestão da Informação
DCSI	Direção de Comunicações e Sistemas de Informação
DN	Defesa Nacional

I

ICSM	Indicadores de Conhecimento Situacional Marítimo
IMO	*International Maritime Organization*
IUM	Instituto Universitário Militar

M

MB	*MegaByte*
MP	Marinha Portuguesa
MRCC	*Maritime Rescue Coordination Centre*
MSSIS	*Maritime Safety and Security Information System*

N

NOSQL	*Not Only SQL*

O

OE	Objetivo Específico
OG	Objetivo Geral

P

PB	*PetaByte*

Q

QC	Questão Central
QD	Questão Derivada

R

RAM	*Random Access Memory*

S

SI	Sistema de Informação
SIG-DN	Sistema Integrado de Gestão da Defesa Nacional
SGBD	Sistema de Gestão de Base de Dados
SGBDOR	Sistema de Gestão de Base de Dados Objeto-Relacional
SGBDR	Sistema de Gestão de Base de Dados Relacionais
SQL	*Structured Query Language*
SRR	*Search and Rescue Region*
SSD	*Solid State Drives*

T

TB	*TeraByte*
TIC	Tecnologias de Informação e Comunicação
TII	Trabalho de Investigação Individual

V

VHF	*Very High Frequency*
VMS	*Vessel Monitoring System*
VTS	*Vessel Traffic Service*

Y

YB	*YottaByte*

Z

ZB	*ZettaByte*

1. Introdução

De acordo com os elementos estruturantes da doutrina, a Marinha Portuguesa (MP) contribui para o uso do mar através do desempenho das funções de defesa militar e apoio à política externa, de segurança e autoridade do estado e no desenvolvimento económico, científico e cultural (Marinha Portuguesa, 2011). O desempenho eficaz nestas três funções implica uma dinâmica transformacional contínua, focalizada na obtenção dos produtos e serviços que a organização deve gerar e proporcionar (Marinha Portuguesa, 2011). Desta forma, a criação e a exploração de Conhecimento Situacional Marítimo (CSM), resultando na sua génese da vigilância do domínio marítimo, concorre diretamente para a prossecução das funções da MP e é um elemento basilar do planeamento das operações navais, sustentadas que são por linhas de ação genéticas, estruturais e operacionais (Marinha Portuguesa, 2011).

Uma das principais fontes de dados que alimenta o CSM é o *Automatic Information System* (AIS) em que as mensagens podem ser recebidas a partir de três fontes distintas:

- *Maritime Safety and Security Information System* (MSSIS)
- Rede de Antenas VHF da Marinha (continente, Açores e Madeira)
- *European Maritime Safety Agency* (EMSA)

Estas 3 fontes geram cerca de 200 a 300 mensagens por segundo, sendo que apenas metade são georreferenciadas[1] e são estas que interessam no contexto do CSM, o que origina diariamente cerca de 10 milhões de mensagens AIS, totalizando cerca de 300 GB por ano, com tendência crescente (Marinha, 2021).

Em 1970, Edgar Frank Codd apresenta o modelo relacional como forma de gerir a informação numa base de dados (BD). Esta inovação permitiu migrar sistemas hierárquicos que eram baseados em ficheiros para uma BD relacional com tabelas que contêm os dados (Codd, 1970).

Este modelo facilitou em muito a gestão da informação, o que contribuiu para que as organizações adotassem este modelo que, consequentemente, proporcionou-lhes conseguir melhores resultados. No entanto, E.F.Codd não conseguiu prever que ao longo do tempo fosse necessário armazenar grandes dimensões de informação, para além do elevado número de pedidos feitos às bases de dados. Esta situação evoluiu de forma exponencial com a propagação da internet que veio conectar todas as pessoas, de toda a parte do mundo, que possuem um qualquer dispositivo eletrónico. Assim sendo, quanto mais pessoas se

[1] Dos 27 tipos de mensagens AIS apenas as mensagens do tipo 1, 2, 3, 18 e 19 são georreferenciadas

conectam, mais informação circula, e essa mesma informação tem de ser guardada e gerida da forma mais eficiente possível (Espinosa, 2019).

No que se refere à gestão e estruturação de grandes volumes de dados, as bases de dados relacionais começaram a apresentar lacunas na capacidade de gestão e *performance* no tratamento dos dados (Espinosa, 2019).

Nos últimos anos surgiu uma nova arquitetura de BD, o NoSQL, que mais não são do que bases de dados não relacionais e mais flexíveis, que visam ultrapassar ou minimizar as limitações das bases de dados relacionais aquando do tratamento de grandes volumes de dados (Espinosa, 2019).

Esta investigação assume particular relevância, no sentido de propor a definição de um sistema de gestão e estruturação de grandes volumes de dados oriundos do AIS, para posterior análise no âmbito do CSM[2] na MP, dada a importância da disponibilidade, *performance* e capacidade de *analytics* sobre os dados armazenados.

O presente estudo tem então por objeto de investigação a gestão e estruturação de grandes volumes de dados oriundos do AIS no contexto do CSM e foi delimitado (Santos & Lima, 2019):

- Relativamente ao âmbito, à utilização das plataformas PostGreSQL e "Mongo DB Atlas Database", utilizando as versões mais atuais destas plataformas que são a versão 13.4 do PostGreSQL e a versão 5.0 do Mongo DB;

- Temporalmente, à atualidade (período de 01 janeiro de 2020 a 31 dezembro de 2021), conseguindo assim utilizar dados atualizados e em volume suficiente para estar no contexto de *Big Data*;

- Espacialmente, aos dados AIS que são provenientes de 3 fontes distintas, sendo privilegiado a fonte de dados das antenas VHF da Marinha Portuguesa, cuja cobertura abrange a costa continental portuguesa, Açores e Madeira, dada a sua importância no CSM na MP;

- Em conteúdo, às mensagens AIS de tipo 1, 2, 3, 18 e 19 que são as únicas que apresentam informação georreferenciada. Dos 5 V's do *Big Data* serão analisados o **v**olume, **v**elocidade, **v**eracidade e **v**alor, não sendo analisada a **v**ariedade, pois das fontes de dados que contribuem para o CSM apenas serão analisados os dados AIS.

[2] Os dados que caracterizam o ambiente marítimo no âmbito do CSM englobam posições de navios provenientes dos sistemas AIS e VMS, dados provenientes de radar, dados meteorológicos, relatos da atividade de fiscalização marítima, etc.

Neste enquadramento, esta investigação tem como objetivo geral (OG) *selecionar contributos para otimizar a gestão e estruturação de dados AIS, na MP, no contexto do CSM na era do Big Data*, e específicos (OE):

OE1: Comparar os sistemas de bases de dados relacionais e não relacionais no contexto de *Big Data;*

OE2: Analisar a adequação de ambientes de base de dados não-relacionais, em sistemas de *Big Data*, no que se refere a gestão, estruturação, análise e *performance* de dados AIS, no contexto do CSM na MP;

OE3: Analisar a influência da capacidade da infraestrutura na *performance* do tratamento de *Big Data*, no contexto CSM na MP.

A questão central (QC) de investigação *Como tornar um sistema de Big Data mais eficiente ao nível da gestão e estruturação de dados AIS, no contexto CSM?* encontra-se alinhada com estes objetivos.

Estruturalmente, este documento encontra-se organizado em oito capítulos. Este primeiro, que introduz o tema. Um segundo, respeitante à revisão da literatura e apresentação do modelo de análise proposto. O terceiro, concernente à metodologia e ao método utilizados. Os quarto, quinto, sexto e sétimo, relativos à análise dos dados, discussão dos resultados e resposta às questões de investigação. O oitavo, e último, ancorado nas conclusões, contributos para o conhecimento, limitações, proposta de estudos futuros e recomendações de ordem prática.

2. Enquadramento teórico e concetual

Neste capítulo são estudados, em termos de revisão da literatura, os conceitos estruturantes para o trabalho e apresentado o modelo de análise adotado.

2.1. Estado da arte e conceitos estruturantes

2.1.1. Big Data

O rápido crescimento da *internet* potenciou a migração de sistemas que antes se encontravam isolados em redes locais de cada uma das empresas e, que a partir desse momento, ficaram alojados num novo ambiente crescente e comum: a internet. Esta passagem para aplicações web torna o seu acesso mais fácil, uma vez que os utilizadores acedem todos a um único ponto onde se encontra armazenada a aplicação e a informação (Espinosa, Kaisler, Armour, & Money, 2019). Toda esta informação fica armazenada em bases de dados e foi então que surgiu a necessidade da aplicação de conjuntos matemáticos apresentados pelo matemático E. F. Codd, onde se criam regras para a representação dos dados de cada empresa tornando, assim, mais fácil a sua gestão bem como a interpretação da informação (Espinosa, Kaisler, Armour, & Money, 2019). Estas bases de dados relacionais são ideais para dados estruturados onde se pretende uma integridade dos dados e facilitar o desenvolvimento de aplicações em torno deste tipo de BD. Existem BD relacionais, umas em código aberto (p. ex. MySQL, MariaDb, PostgreSQL) e outras corporativas (p. ex. Microsoft SQL Server, Oracle), sendo as últimas mais voltadas para o mundo empresarial de larga escala. A necessidade de armazenar informação tem vindo a ser cada vez maior, o que colocou um problema a este tipo de representação de informação (modelo relacional), uma vez que não foi criado com o intuito de guardar um grande volume de informação (Espinosa, Kaisler, Armour, & Money, 2019).

De forma a poder colmatar a fraca capacidade de armazenamento da informação, e baseado no modelo relacional, criaram-se as *datawarehouses*, que conseguem guardar uma grande quantidade de dados sem perder a integridade, fiabilidade e com grande disponibilidade no acesso aos mesmos. Desta forma, as *datawarehouses* solucionam o problema de armazenamento de um grande volume de dados para uma representação de dados bem estruturada (respeitando as regras do modelo relacional). No entanto, as *datawarehouses* podem não ser uma solução viável uma vez que, em casos onde a representação da informação esteja mal estruturada ou até mesmo sem qualquer estrutura, estas não solucionam o problema de armazenamento. Para além disso, uma das grandes

desvantagens destas *datawarehouses* é a sua longa e dispendiosa implementação no sistema organizacional das empresas (Espinosa, Kaisler, Armour, & Money, 2019).

Segundo Vieira et al. (2012), *Big Data* pode ser resumidamente definido como "uma coleção de bases de dados tão complexa e volumosa que se torna muito difícil e complexo fazer algumas operações simples (ex: remoção, ordenação, sumarização) de forma eficiente utilizando Sistemas de Gestão de Bases de Dados tradicionais". Também acrescenta que engloba o processamento, de forma eficiente e escalável, de grandes volumes de dados complexos produzidos por diversas aplicações. Segundo Espinosa (2019), o sistema *Big Data* assenta nos chamados 5V's:

- **Volume**: A necessidade de lidar, de forma sustentada, com o crescimento exponencial de dados gerados;

- **Velocidade**: A necessidade de processar os dados em tempo real, devido à importância que isso tem para as organizações atuais;

- **Variedade**: O facto dos dados se apresentarem nos mais variados formatos, alguns deles dificilmente acomodáveis em estruturas rígidas como as disponibilizadas pelo modelo relacional;

- **Veracidade**: Devido ao elevado volume de dados há necessidade de separar os verdadeiros dos falsos;

- **Valor**: Necessidade de gerar informação de valor acrescentado através dos dados.

2.1.2. Conhecimento Situacional Marítimo (CSM)

No âmbito do CSM, como parceiro indispensável para a ação do Estado no Mar (CEMA, 2011), a Marinha Portuguesa reconheceu ser essencial "deter superioridade de informação no ambiente marítimo" por via do conhecimento/vigilância desse espaço. Assim, e através de comportamentos singulares ou atípicos, seria possível obter indicadores de potenciais ameaças "à segurança, ao exercício da autoridade do estado, ao ambiente e/ou aos recursos económicos" (Marinha Portuguesa, 2011). O processo é concretizado pela aquisição, executada através de sensores e do elemento humano, i.e., vigilância, e pelo controlo, processo correspondente à análise de dados à criação de conhecimento e à sua partilha. Estes dois patamares, aquisição e controlo, sustentam um terceiro correspondente à intervenção, i.e., a exploração operacional do conhecimento desenvolvido nos dois primeiros.

O objetivo geral do CSM é obter uma compreensão efetiva das atividades no domínio marítimo, que permita aos decisores e à comunidade operacional atuar de forma oportuna,

precisa e eficaz, possibilitando ao mesmo tempo a respetiva avaliação dos efeitos da ação e assim ajustar em conformidade, com o propósito de ultrapassar desafios, minimizando os riscos e rentabilizando o emprego de recursos (Marinha Portuguesa, 2011). Como resultado da análise desses dados e da observação documental, obter-se-ão padrões de atuação/ocorrências que facilitarão e sustentarão, nos centros de comando e controlo e coordenação, a decisão e atuação pretendidas. A sistematização de todo este processo, o qual inclui a difusão e partilha da informação gerada, promove e otimiza a intervenção no momento adequado e com os meios disponíveis, inclusive através do envolvimento de outras agências com quem existam protocolos estabelecidos (Marinha Portuguesa, 2011).

A construção de Indicadores de Conhecimento Situacional Marítimo (ICSM) a partir das posições de navios através de dados AIS constitui um dos primeiros requisitos identificados pelo Comando Naval (CN) que possibilitam conhecer a dinâmica da navegação em áreas de interesse nacional. Um dos requisitos em termos de informação de natureza estatística está em estimar o número de monitorizações efetuadas ao longo de um ano pelos *Maritime Rescue Co-ordination Centre* (MRCC). Por "monitorizações" entende-se o número de navios distintos que os MRCC (MRCC Lisboa e MRCC de Ponta Delgada e MRCC Funchal) acompanham diariamente através de sistemas de informação como o OVERSEE, SEAVISION ou TV32 e outros sistemas de informação e comunicação. Estas monitorizações traduzem um esforço por parte da organização em acompanhar os trânsitos destes navios ao longo das *Search and Rescue Region* (SRR) nacionais. Também esta informação, após adequado tratamento estatístico, permite conhecer a dinâmica destas áreas em termos da sua densidade de navegação e principais rotas praticadas por tipo de navio (Marinha Portuguesa, 2011).

2.1.3. *Automatic Information System* (AIS)

O *Automatic Identification System* (AIS) é um sistema de monitoração de curto alcance utilizado em navios e *Vessel Traffic Service* (VTS). O sistema foi desenvolvido por militares, porém a tecnologia foi transferida para o setor civil sem grandes modificações (IMO, 2002).

O sistema AIS serve para identificar e localizar embarcações por intermédio da troca eletrónica de dados com outros navios e estações VTS, permitindo obter informações tais como identificação, posição, rota e velocidade. O sistema AIS destina-se a auxiliar os oficiais das embarcações e permitir que as autoridades navais rastreiem e monitorizem os deslocamentos das embarcações (IMO, 2002).

O sistema AIS integra um sistema transcetor VHF padrão tal como LORAN-C ou recetor GPS, juntamente com outros sensores de navegação, eletrónicos ou não, tais como girobussola, indicador de velocidade e indicador de velocidade de rotação e de direção (IMO, 2002).

Existem 27 tipos diferentes de mensagens AIS, descrito em Apêndice B – Lista de mensagens AIS.

Os tipos de mensagem mais comuns a ser transmitidos são dois: relato de posição e dados do navio e da viagem (IMO, 2002). Os relatos de posição contêm, entre outros dados: posição, velocidade, rumo e proa. A IMO estabelece que a periodicidade de envio deve ser ajustada automaticamente de acordo com a velocidade do navio. Por exemplo, se o navio estiver a mais de 14 nós, eles devem ser enviados a cada 2 segundos. Se estiver ancorado, deve enviar a cada 3 minutos. Já a mensagem que transmite os dados do navio e da viagem, deve ser enviada a cada 6 minutos. Essa mensagem contém, entre outros dados: nome do navio, tipo do navio, comprimento, boca, calado e destino (IMO, 2002).

Os dados AIS provenientes da Rede de antenas VHF da Marinha, APRAM, Portos dos Açores e VTS são integrados num único serviço que os disponibiliza internamente da Rede de Comunicações da Marinha. Este serviço é designado por AIS nacional.

O quadro abaixo apresenta estimativas para o nº de mensagens recebidas por dia, para o nº médio de *bytes* que estas ocupam por dia e para a percentagem de mensagens dinâmicas e estáticas por dia.

Quadro 1 – Nº de mensagens AIS (diárias)

Fonte	MSSIS				Nacional				SatAIS				Total			
Ano	Horas	GB	MB/Hora	Ano	Horas	GB	MB/Hora	Ano	Horas	GB	MB/Hora	Ano	Horas	GB	MB/Hora	Ano
2016	8549	140	16,76921	143,5536	8625	38,7	4,594643	39,33266					17174	178,7	21,36386	182,8863
2017	8763	148	17,29453	148,0507	8759	31,1	3,635849	31,12485					17522	179,1	20,93038	179,0529
2018	8027	137	17,47702	149,6128	8543	38,4	4,602786	39,40236					16570	175,4	22,0798	188,8858
2019	7814	137	17,95342	153,6911	8759	33,7	3,93981	33,72693	4608	30,4	6,755555556	57,83125	21181	201,1	28,64878	245,0814
2020	6021	107	18,19764	155,7818	6582	28,6	4,449468	38,08988	6786	58,07	8,762699676	75,01350133	19389	193,67	31,40981	268,7011
Mínimo	6021	107	16,76921	143,5536	6582	28,6	3,635849	31,12485	4608	30,4	6,755555556	57,83125	16570	175,4	20,93038	179,0529
Média	7834,8	133,8	17,53836	**150,138**	8253,6	34,1	4,244511	**36,33534**	5697	44,235	7,759127616	**66,42237566**	18367,2	185,594	24,88653	**212,9215**
Máximo	8763	148	18,19764	155,7818	8759	38,7	4,602786	39,40236	6786	58,07	8,762699676	75,01350133	21181	201,1	31,40981	268,7011
												Média 2019/2020:	20285	197,385	30,0293	**256,8912**

Fonte: Direção de Análise e Gestão da Informação (2022)

A partir dos valores apresentados na tabela acima, verifica-se que será necessário armazenamento mínimo de 300 GB por ano só para os dados "raw", contando já com algum crescimento nas fontes Nacional e Sat-AIS.

2.2. Modelo de análise

Esta investigação foi desenvolvida conforme o modelo apresentado no Apêndice B.

3. Metodologia e método

Este estudo enquadra-se nas Ciências Militares, designadamente na área de Técnicas e Tecnologias Militares, subárea de Comando, Controlo, Comunicações, Computadores e Informação (C4i), apresentando-se neste capítulo a metodologia e o método que o nortearam.

3.1. Metodologia

Por forma a responder às questões de investigação enunciadas e, consequentemente atingir o objetivo proposto, a filosofia de pesquisa adotada foi o pragmatismo, abordando diferentes perspetivas, consideradas relevantes, e enfatizando as suas consequências práticas. Neste sentido aplicar-se-á um raciocínio indutivo e pensamento crítico (Santos & Lima, 2019). Adotar-se-á uma estratégia qualitativa, sem invalidar a apresentação de dados quantitativos para ilustrar alguns aspetos, quando considerado pertinente. A presente investigação efetua-se com base num desenho de pesquisa de estudo de caso (Santos & Lima, 2019), conforme Figura 1.

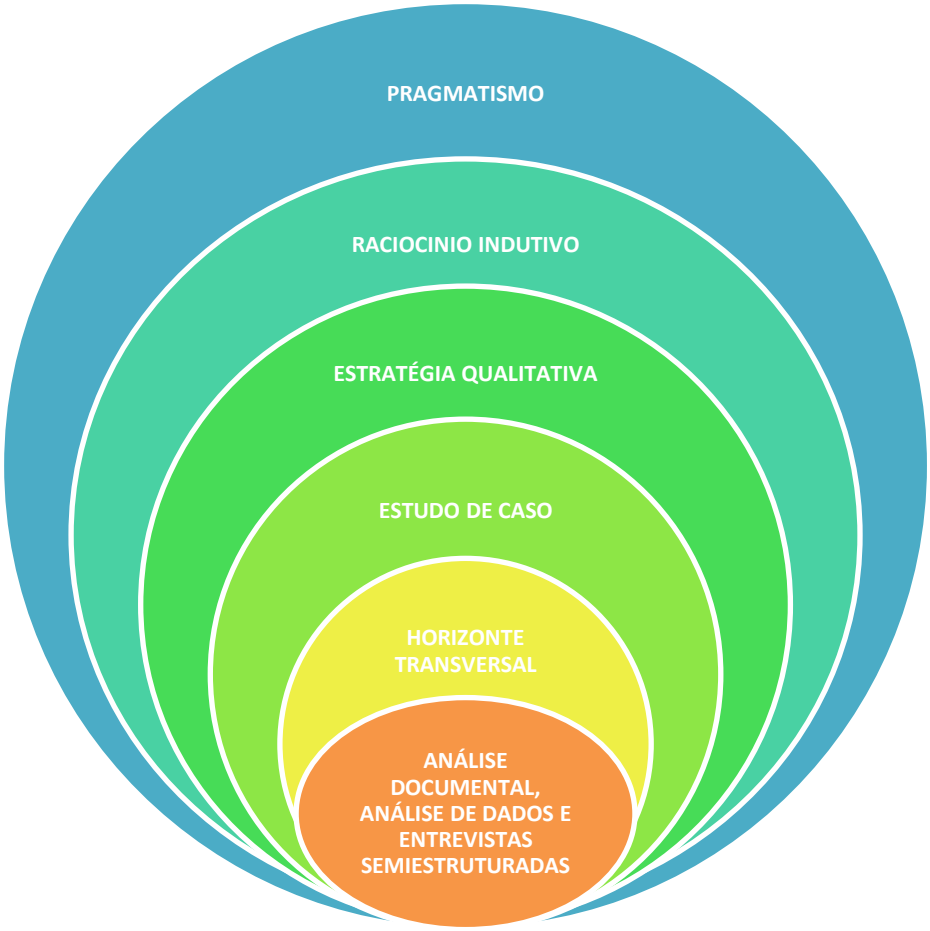

Figura 1 - Estratégia e metodologia na investigação

3.2. Método

3.2.1. Participantes e procedimento

- Participantes: Integram esta investigação 10 especialistas de reconhecido mérito: 2 militares e 8 civis, conforme Apêndice D.

- Procedimento: As entidades entrevistadas foram previamente contactadas por telefone ou email, para averiguar da sua disponibilidade para integrar o estudo. Após confirmação, foi agendada a entrevista (presencialmente, por email, vídeo ou audioconferência, conforme disponibilidade e restrições associados à situação de pandemia vigente). Foram asseguradas as garantias de anonimato e confidencialidade da informação prestada, da qual todas as entidades abdicaram.

3.2.2. Instrumento(s) de recolha de dados

Foi construído um guião de entrevista semiestruturada (Apêndice E) destinado a entidades militares e civis para aferir a situação atual relativa ao armazenamento e processamento de dados, e o seu posicionamento quanto a requisitos *Big Data* e adoção de *cloud computing*.

3.2.3. Técnicas de tratamento de dados

Foi efetuada uma análise qualitativa do conteúdo das entrevistas (Apêndice F), cujo resultado foi conjugado com a análise documental, revisão da literatura e aplicação de métricas em testes realizados num protótipo de forma a permitir responder às questões de investigação.

4. Caracterização de sistemas de bases de dados relacionais e não relacionais no contexto de *Big Data*

4.1. Bases de Dados relacionais

4.1.1. Características

Em 1970, Edgar Frank Codd publicou um artigo com o tema "Modelo de dados relacional para grandes bases de dados partilhadas", onde este definiu um modelo relacional baseado na teoria dos conjuntos matemáticos. Aqui, foi apresentada uma nova proposta de arquitetura, com o objetivo de conseguir armazenar, gerir e relacionar os dados numa BD. Este novo modelo aliviou a carga de trabalho das pessoas que fazem desenvolvimento de aplicações, sendo que estas não precisam de saber detalhes sobre os dados que vão ser geridos. Este marco importante delineou uma metodologia, onde se vai combinar a álgebra e o cálculo relacional, para permitir o armazenamento e recuperação de grandes quantidades de informação. Esta arquitetura tornou-se, assim, a base do modelo relacional.

As vantagens do modelo relacional são o tratamento da derivabilidade, redundância e consistência das inter-relações. A indexação dos dados tem vantagens e desvantagens, como por exemplo melhora o desempenho das pesquisas e nas atualizações dos dados, mas podem prejudicar nas operações de inserção e remoção dos mesmos, como também a sua redundância. As possibilidades de acesso aos dados são representadas em estruturas de árvore ou num modelo em rede, mas muitas aplicações falham se alguma destas estruturas for alterada. (Codd, 1970)

4.1.2. ACID

No geral quando uma BD relacional é modelada, para que se possa assegurar a integridade dos dados é uma premissa que a BD mantenha nas suas transações quatro propriedades conhecidas pela sigla ACID, descritas a seguir (Johnsen, 2019):

- A (Atomicidade): todas as transações devem ser atómicas, ou seja, só podem ser consideradas efetivadas se executadas na sua totalidade, logo caso aconteça alguma falha no decorrer do processo significa que toda transação será invalidada.

- C (Consistência): tudo deve ser consistente desde o início até o final de determinada transação, caso a transação não aconteça a BD garante a integridade dos dados retornando ao estado consistente anterior.

- I (Isolamento): as transações só podem acontecer de maneira isolada, nenhuma transação pode interferir com outra.

- D (Durabilidade): por último deve ser durável, uma vez que a transação foi concluída, os dados consequentes da mesma não podem ser perdidos.

O uso das propriedades ACID trouxe ao modelo relacional, segurança e eficiência, mas em contrapartida tornou-o de certa forma inflexível, pois este conjunto de propriedades é demasiadamente restritivo para ambientes de processamento distribuídos de grande porte e acaba inviabilizando soluções que necessitam de maior flexibilidade.

4.1.3. Limitações

A estrutura das bases de dados relacionais tem vindo a ser a preferida devido ao seu *layout*, tabelas de nomes fixos e tipos de colunas. No entanto, o volume de dados tem crescido exponencialmente em certas organizações, como por exemplo o Facebook, que já conseguiu atingir o nível de *petabytes* de informação. (Zuckerberg, s.d.)

No caso deste tipo de organizações, a utilização de Sistemas de Gestão de Bases de Dados (SGBD) relacionais têm-se mostrado uma problemática e não tão eficiente, apresentando limitações, tais como:

- Escalabilidade Vertical – Os utilizadores podem escalar uma BD relacional, executando num computador com vários *Central Processing Unit* (CPU) que compartilham a *Random Access Memory* (RAM) e discos rígidos. Deste modo, podem ser adicionados mais processadores e memórias para aumentar o desempenho de um sistema. No entanto, esta abordagem torna-se limitada e normalmente cara (Leavitt, 2010). Consegue-se escalar até um certo ponto e fazer uma distribuição por vários servidores. Além do mais, estas bases de dados não foram pensadas para trabalhar com particionamento de dados. (Leavitt, 2010)

- Complexidade – Com estas BD, os utilizadores podem converter todos os dados em tabelas. Se os dados, por acaso, não se conseguirem encaixar numa tabela, terão que ser criadas mais tabelas para esses dados e isso pode tornar a estrutura mais complexa, difícil e mais lenta para se trabalhar. (Leavitt, 2010)

- SQL – Usar SQL é eficaz para dados estruturados, porém o uso desta linguagem com outro tipo de dados é difícil, porque foi projetada para trabalhar com Bases de Dados (BD) estruturadas, organizadas e com os dados em tabelas. O SQL pode gerar grandes quantidades de código complexo se for necessário fazer junções entre várias tabelas e também não funciona bem para o desenvolvimento moderno. (Leavitt, 2010)

4.2. Bases de Dados não relacionais

4.2.1. O que é o NoSQL?

O NoSQL nasceu de uma necessidade premente em conseguir-se trabalhar com um grande volume de dados armazenados, dados esses com informação de utilizadores, produtos, etc. Com a frequência a que estes mesmos dados estão a ser acedidos, existe uma grande necessidade de ter um bom desempenho de processamento e, para isso, teve de existir uma grande mudança no paradigma das bases de dados. Devido a este novo tipo de modelo de dados, foram desenvolvidas novas plataformas para se conseguir trabalhar com um grande volume de dados e ter um grande desempenho. (Blokdyk, 2020)

O NoSQL, que se pode traduzir por "*Not Only SQL*", é um termo para definir SGBD's não relacionais, que não se baseiam no modelo relacional, e que não seguem as suas regras. Tornam-se mais flexíveis porque não aplicam as propriedades ACID utilizadas no modelo relacional. Este tipo de flexibilidade é necessário devido aos requisitos da alta escalabilidade para a gestão de grandes volumes de dados, assim como também da sua disponibilidade. Para aplicar este tipo de sistemas tem de existir um conjunto de características, por exemplo, ser não relacional, distribuído, que o esquema seja flexível e por fim aplicar a escalabilidade horizontal. (Blokdyk, 2020)

O modelo de BD não relacional implicou uma mudança nas propriedades em relação ao modelo relacional, o NoSQL possui propriedades BASE (*Basically Available, Soft state, Eventual consistency*), enquanto o relacional possui propriedades ACID (*Atomicity, Consistency, Isolation and Durability*). (Blokdyk, 2020)

4.2.2. Características

O NoSQL tem algumas características que se distinguem dos SGBD's relacionais. Essas características são importantes para o armazenamento ajustado a grandes volumes de dados e quando não se encontram estruturados ou semiestruturados. Dentro dessas características podemos enaltecer as seguintes:

- Escalabilidade horizontal

Uma característica importante para sistemas distribuídos é ser capaz de escalar horizontalmente, isto é, adicionar nós ao sistema (replicar e particionar os dados em diferentes servidores). Assim, permite que as operações de leitura e escrita possam ser executadas de forma muito eficiente. (Blokdyk, 2020)

Por princípio, como se pode adicionar mais nós ao sistema, este deve escalar linearmente. A título de exemplo, se dobramos o número de nós num sistema, este deveria

ser capaz de suportar o dobro da taxa de transferência máxima, mas ao analisarmos a Lei de Amdahl, pode verificar-se que isso não é verdade. Os sistemas distribuídos são limitados pela quantidade de computação em paralelo e os mecanismos envolvidos nessa distribuição apresentam alguma sobrecarga, o que degrada o desempenho. (Blokdyk, 2020)

Como exemplificado na Figura 2, a escalabilidade horizontal é alusiva com a funcionalidade na distribuição de dados e da carga por diversos servidores, sem precisar de partilhar memória ou disco. Esta abordagem permite o uso de *hardware* mais barato e comum. Os SGBDs que são voltados para sistemas de *datawarehouse* fornecem escalabilidade horizontal, porém as consultas podem ser complexas como, por exemplo, vários *joins* a tabelas diferentes. (Blokdyk, 2020)

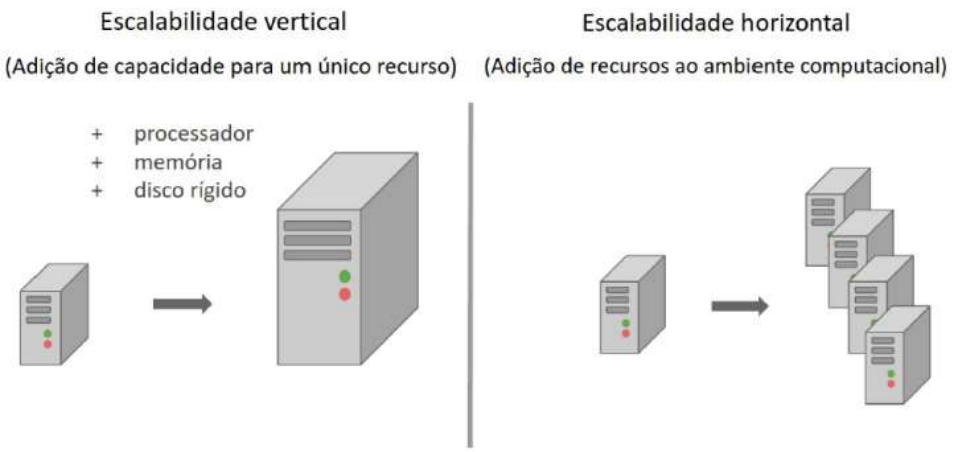

Figura 2 – Escalabilidade vertical Vs escalabilidade horizontal

Fonte: adaptado a partir de Blokdyk, (2020)

- Schema free

Um *schema* na área das BD é um conjunto de fórmulas, relações e constrangimentos complexos aquando da criação das mesmas. (Rybinski, 1987)

Este tipo de estrutura é o resultado de um processo chamado normalização, uma simplificação e organização dos campos que existem nas tabelas para assim reduzir a redundância dos dados. (Codd, 1970)

Com este tipo de esquemas e normalizações, à medida que a quantidade de dados vai aumentado numa tabela, o acesso a esses dados pode ficar mais demorado. Para melhorar o desempenho na leitura dos dados, abdica-se da normalização, mas cria-se um problema na redundância de dados. (Bock & Schrage, 2002)

Esta é uma das características mais relevantes dos sistemas NoSQL: a ausência completa ou quase total de um esquema que defina a estrutura do modelo de dados. Com essa ausência de um esquema definido, fica mais fácil aplicar a escalabilidade e aumenta a disponibilidade. (Blokdyk, 2020)

Torna-se também, mais fácil a inserção de dados, já que estes não têm de obedecer a regras de um esquema pré-definido. Assim, passa a existir uma BD que armazena dados não estruturados, em qualquer formato, sem precisar de normalizar os mesmos, ficando, desta forma, com um bom desempenho. No entanto, devido a essa ausência de um esquema definido, não existe garantia de integridade dos dados porque passam a estar menos organizados. (Edlich, 2018)

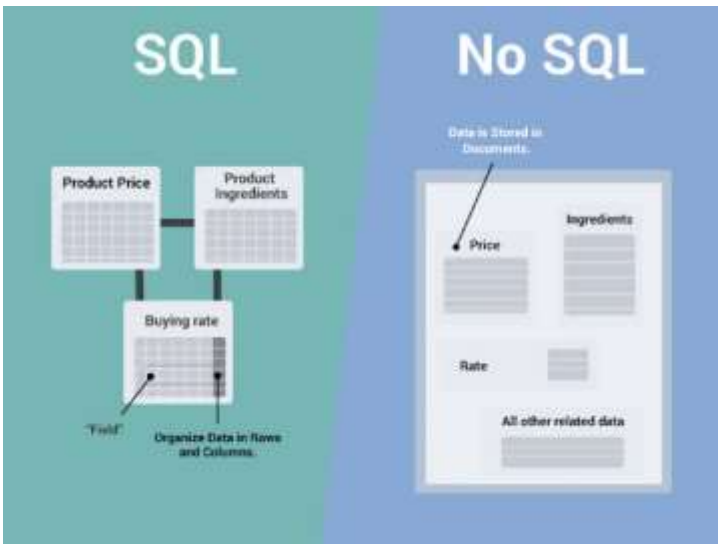

Figura 3 – Schema inclusivo (SQL) VS Schema Free (NoSQL)

Fonte: adaptado a partir de Edlich (2018)

- API simples

Um dos principais objetivos dos sistemas NoSQL é proporcionar que o acesso aos dados seja feito de uma forma rápida, para assim conseguir oferecer uma alta disponibilidade e uma boa escalabilidade. Para se conseguir este objetivo, foram desenvolvidas API com um conjunto de funções para facilitar o acesso à informação, permitindo, deste modo, que qualquer aplicação possa ter acesso aos dados de uma BD de uma forma rápida e eficiente. (Edlich, 2018)

- Suporte nativo à replicação

Outra forma de fornecer escalabilidade é através da replicação. Autorizar a replicação de forma nativa, diminui o tempo gasto para recuperar as informações. Existem duas abordagens para a replicação:

- Master-Slave: Este tipo de replicação cria um nó com uma copia oficial que manipula as escritas enquanto os escravos fazem a sincronização com o seu mestre e também pode lidar com as leituras; (Harrison, 2015)

- Peer-to-Peer: É uma replicação que permite gravar em qualquer nó, onde estes mesmos nós se coordenam para sincronizar as suas cópias dos dados. (McCreary & Kelly, 2014)

4.2.3. Teorema de CAP

O teorema de CAP de Eric Brewer explica que é preciso escolher entre a consistência forte (*Consistency*), alta disponibilidade (*Availability*) e tolerância no particionamento (*Partition tolerance*) para um sistema ser distribuído.

- Consistência forte: Todos os nós de um sistema têm de ter os mesmos dados ao mesmo tempo, portanto, qualquer utilizador que utilize o sistema receberá a mesma cópia independentemente de qual nó responde ao seu pedido;

- Alta disponibilidade: Implementação de um sistema de modo que seja garantido que este fica ativo durante um determinado tempo e quando é solicitado para algum tipo de operação;

- Tolerância no particionamento: Capacidade de um sistema continuar a trabalhar sobre a circunstância de acontecer uma falha na rede ou alguma perda de dados. Isto significa que vai garantir que as operações vão ser concluídas, mesmo que elementos individuais não estejam disponíveis. Uma falha em um nó não deve criar uma debilidade no sistema. (Brewer, 2012)

Este teorema de CAP contém estas três propriedades que funcionam como atributos de qualidade.

Figura 4 - Modelo do Teorema CAP

Fonte: adaptado a partir de Brewer (2012)

Neste teorema, Eric Brewer demonstrou que entre estas três propriedades, somente duas podem ser garantidas ao mesmo tempo num sistema de dados partilhados. Por isso temos três opções:

- **Consistência + Disponibilidade:** Com esta possibilidade, os dados estão consistentes entre os nós, desde que estejam todos ativos, onde vai ser capaz de realizar leituras e escritas em qualquer nó e certificar que os dados são os mesmos. Esta situação não sabe lidar com uma possível falha de uma partição. Caso ocorra, o sistema inteiro pode ficar indisponível até o elemento ser retomado. (Brewer, 2012)

Figura 5 - Sistema com Consistência e Disponibilidade

Fonte: adaptado a partir de Brewer (2012)

- **Consistência + Particionamento:** Os sistemas, ao adotarem estas duas propriedades, inevitavelmente perderão um pouco da disponibilidade. Os dados são consistentes entre todos os nós, evitando que fiquem dessincronizados, mas podem ficar indisponíveis se em algum dos nós ocorrer uma falha. (Brewer, 2012)

Figura 6 - Sistema com Consistência e Particionamento

Fonte: adaptado a partir de Brewer (2012)

- Disponibilidade + Particionamento: Existem sistemas que nunca podem ficar *off-line*, por isso não desejam sacrificar a disponibilidade. Para ter alta disponibilidade mesmo com tolerância a particionamento, vai ser preciso abrir mão da consistência. Aqui, a ideia é os sistemas aceitarem escritas e sincronizarem os dados depois. (Brewer, 2012)

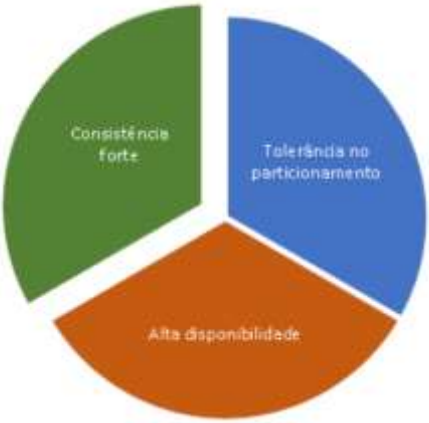

Figura 7 - Sistema com Disponibilidade e Particionamento

Fonte: adaptado a partir de Brewer (2012)

Muitos clientes NoSQL baseiam-se nestas três propriedades do teorema de CAP e como o próprio Eric Brewer indica, só se consegue garantir duas das três propriedades a trabalhar ao mesmo tempo. Assim, cada cliente de NoSQL seleciona as duas propriedades que mais se adequam à sua realidade. (Brewer, 2012)

Na Figura 8 são mostrados alguns clientes NoSQL distribuídos pelas propriedades deste teorema.

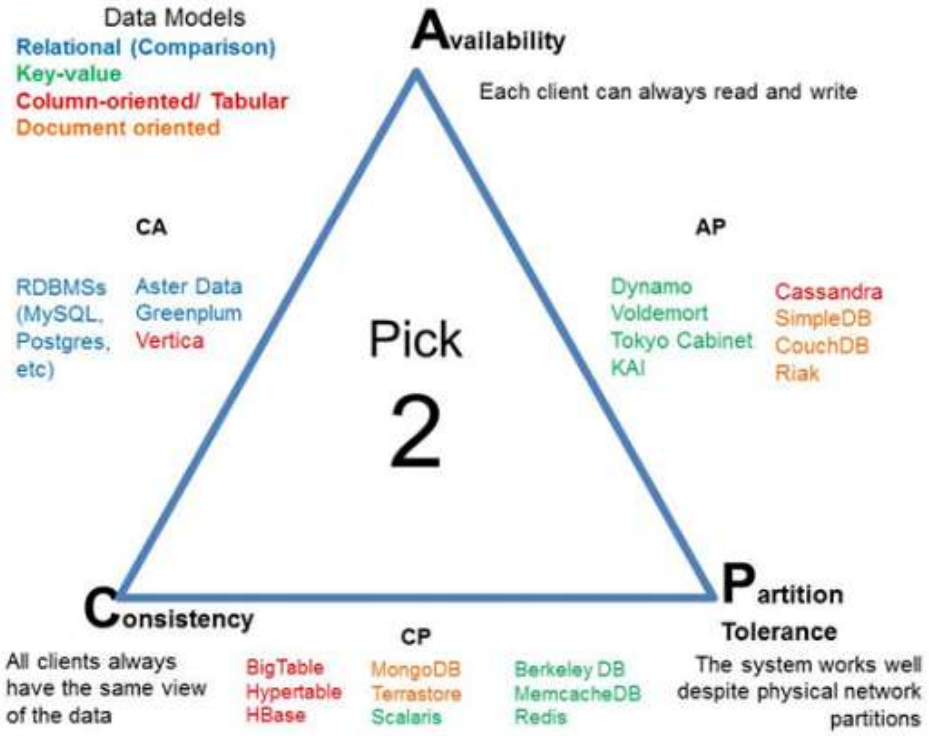

Figura 8- Clientes NoSQL seguindo o teorema CAP

Fonte: adaptado a partir de Guémar (2015)

A partir deste teorema de CAP foram criadas as propriedades BASE (*Basically Available, Soft-state, Eventually consistency*), que são o oposto das propriedades ACID, que os clientes NoSQL utilizam para o controlo da consistência dos dados (Johnsen, 2019). A consistência eventual (transações que são normalmente assíncronas) é uma forma de consistência mais fraca, que permite melhorar a velocidade e a disponibilidade. As propriedades ACID fornecem uma consistência forte (operações síncronas) e para bases de dados particionadas dificultam a disponibilidade (Johnsen, 2019). Enquanto o ACID é pessimista e requer consistência no final de cada operação, o BASE é otimista e aceita que a consistência de uma BD seja feita num momento mais tarde. A consistência eventual é simplesmente um reconhecimento da existência de um atraso não vinculado na propagação de uma alteração feita em uma máquina para todas as outras cópias, em que estas podem

conter dados desatualizados. Por exemplo, um sistema distribuído mantém cópias dos dados divididos em várias máquinas num *cluster* para garantir uma alta disponibilidade. Quando os dados são atualizados num *cluster*, pode haver algum intervalo de tempo durante o qual algumas cópias serão atualizadas e outras não. Eventualmente, as alterações vão ser propagadas por todas as máquinas restantes e é por isso que dá pelo nome de consistência eventual. Este tipo de consistência nada tem a ver com os sistemas de um único nó, já que estes não necessitam de propagação (Johnsen, 2019).

As propriedades BASE são características que existem nas bases de dados NoSQL (Johnsen, 2019). Nomeadamente:

- **Basicamente disponível:** Esta propriedade afirma que o sistema garante a disponibilidade dos dados como é referido no teorema de CAP. Vai existir uma resposta a qualquer pedido, mesmo que essa resposta possa ter uma falha a obter os dados solicitados, como também dados inconsistentes. (Johnsen, 2019)

- **Estado leve:** O estado de um sistema pode mudar ao longo do tempo, mesmo durante períodos sem inserção de dados. O sistema não tem de estar consistente todo o tempo, por isso o seu estado tem de ser *soft* para conseguir suportar modificações; (Johnsen, 2019)

- **Eventualmente consistente:** O sistema eventualmente vai convergir para um estado consistente. Os dados vão ser propagados por todos os nós disponíveis, mas se continuar a receber dados e se a consistência destes não for verificada em cada transação, passado algum tempo pode não existir a consistência desejada. (Johnsen, 2019)

Um sistema está essencialmente sempre a funcionar, mas não tem de estar consistente todo esse tempo, porque irá ficar consistente no momento devido. As propriedades que o BASE perde para o ACID, algumas delas podem ser tratadas a nível aplicacional, tratando os dados e validando os mesmos para irem consistentes aquando da sua inserção.

4.2.4. Limitações

Os sistemas NoSQL também causam certas preocupações e dúvidas, e isso implica limitações no uso deste tipo de sistemas, tais como as seguintes:

- **Sobrecarga e complexidade** – Os sistemas NoSQL não trabalham com SQL. Requerem programação manual para fazer consultas, o que pode ser rápido para tarefas simples, mas demorado em algumas tarefas mais complexas. Aliás, programar consultas complexas para estas bases de dados é difícil; (Johnsen, 2019)

- Confiabilidade – As bases de dados relacionais suportam nativamente as propriedades ACID, enquanto as bases de dados não relacionais não suportam, por isso o nível de confiança dos dados é baixo. Se os programadores quiserem adicionar restrições ACID para um conjunto de dados, terá de ser feita programação adicional para o efeito; (Johnsen, 2019)

- Consistência – Como os sistemas NoSQL não suportam nativamente as propriedades ACID, isso pode comprometer a consistência dos dados, a menos que seja fornecido um suporte manual. Como não oferece consistência, isso ajuda a um melhor desempenho e escalabilidade, mas é um problema para vários tipos de *software*, como por exemplo os do sector da banca; (Johnsen, 2019)

- Familiaridade com esta tecnologia – Muitas das organizações ainda não se encontram familiarizadas com o NoSQL, por isso torna-se mais difícil escolher o melhor sistema para resolver um problema; (Johnsen, 2019)

- Ecossistema limitado – Contrariamente aos sistemas de base de dados comerciais, muitos dos clientes NoSQL, ainda não oferecem apoio ao cliente, para além da falta de interfaces para ajudar os programadores. (Leavitt, 2010)

4.2.5. ACID vs BASE

No Quadro 2 fica uma breve comparação entre o ACID e o BASE, para mostrar o que as bases de dados relacionais têm de diferente relativamente às bases de dados não relacionais a nível de controlo transacional.

Quadro 2 - BASE vs. ACID

ACID (Relacional)	BASE (Não relacional)
Consistência forte	Fraca consistência
Isolamento	Disponibilidade primeiro
Foca-se no "*commit*"	O melhor esforço
Transações encapsuladas	Respostas aproximadas
Menos disponibilidade	Simples e rápido
Conservador (pessimista)	Agressivo (otimista)
Evolução difícil	Evolução fácil

Fonte: adaptado a partir de Cook (2019)

4.3. Síntese conclusiva e resposta à Questão Derivada 1

Na possibilidade de utilizar um sistema NoSQL, ao contrário de um que use o modelo relacional, é preciso levar algumas questões em consideração, como os critérios de escalabilidade, a consistência e a disponibilidade dos dados.

A escalabilidade é essencial e neste aspeto os sistemas NoSQL possuem grande vantagem em relação aos SGBD relacionais, pois basicamente foram criados com esse propósito. A partir do momento em que um sistema relacional está a ser acedido por um grupo grande de utilizadores, a escalabilidade vertical passa a não ser suficiente. Resolver este problema consiste em escalonar o próprio sistema de BD, de forma a distribuir a BD por várias máquinas, particionando os dados, onde este processo tem a definição de escalabilidade horizontal. Este tipo de escalabilidade é muito complexo para ser implementado num SGBD relacional, devido à dificuldade em se adaptar em toda a sua estrutura lógica do modelo relacional. (Johnsen, 2019)

Em relação à disponibilidade, é uma questão que preocupa muitas organizações. Neste aspeto, os sistemas NoSQL destacam-se pela sua maior eficiência na disponibilidade, rapidez nas consultas, paralelismos na atualização de dados e maior grau de concorrência. (Johnsen, 2019)

Há, também, que destacar os diferentes paradigmas utilizados tanto nos sistemas relacionais como nos NoSQL. No que diz respeito aos relacionais, a nível de transações, utilizam as propriedades ACID, que força a consistência no final de cada operação. Já as propriedades BASE são usadas pelo NoSQL, que permitem que o sistema fique eventualmente consistente, ou seja, o sistema só se torna consistente no devido momento. (Johnsen, 2019)

No Quadro 3 está representado um resumo de comparação entre os sistemas relacionais e o NoSQL.

Quadro 3 – Sistema Relacional vs NoSQL

	Sistemas relacionais	NoSQL
Modelo de armazenamento	Tabelas com colunas e linhas fixas	Documentos JSON, Chave-Valor e outros tipos
Histórico	Desenvolvido nos anos 70, com foco em redução de dados duplicados	Desenvolvido em 2000 com o foco em escalabilidade e mudança rápida de desenvolvimento
Exemplos	Oracle, MySQL, Microsoft SQL Server, e PostgreSQL	• **Documento:** MongoDB e CouchDB, • **Chave-Valor:** Redis e DynamoDB, • **Wide-column:** Cassandra e HBase, • **Graph:** Neo4j e Amazon Neptune
Esquemas	Rígidos	Flexíveis
Escalabilidade	Vertical (Com mais poder de processamento na mesma máquina)	Horizontal (Escala distribuindo em duas ou mais máquinas)
Transações	Suportado	A maioria não suporta, no entando o MongoDB sim
Joins	Normalmente necessário	Normalmente não é necessário
Mapeamento de Dados para Objetos	Requer um ORM (object-relational mapping)	Pode não precisar de um ORM. Os documentos no MongoDB mapeiam diretamente para dados de estrutura das maiorias das linguagens
Controlo Transacional	ACID	BASE
Escalabilidade	Possível, mas complexo. Devido à sua natureza estruturada, a adição de forma dinâmica e transparente de novos nós a uma tabela não é realizada naturalmente.	Não possui um esquema pré-definido, fazendo com que este tipo de modelo seja flexível. Este modelo contém uma maior flexibilidade, o que favorece a inclusão de outros elementos.
Consistência	Ponto mais forte desde modelo. As regras da consistência presentes são bastante rigorosas quanto à consistência das suas informações.	É realizada eventualmente no modelo: garante apenas se não existir nenhuma atualização dos dados, todos os acessos aos mesmos, será devolvido o ultimo valor que foi atualizado.
Disponibilidade	Por não conseguir trabalhar de forma eficiente com a distribuição de dados, este modelo pode não suportar uma solicitação muito grande de informações de uma BD.	É um ponto forte no sucesso deste modelo. O alto grau de distribuição dos dados, propicia que um maior número de requisições aos dados seja atendido por parte do sistema e que esse fique o menos tempo indisponível.
Isolamento	Sim	Sim
Replicação	Sim	Sim
CAP	Consistência (C) + Disponibilidade (A)	Disponibilidade (A) + Particionamento (P)

Fonte: adaptado a partir de Johnsen (2019)

Com base na análise documental e resumida no Quadro 3, em resposta à QD1 *Quais as principais diferenças entre a utilização de um sistema de base de dados relacional e de um não relacional no contexto de Big Data?*, conclui-se que as principais diferenças entre um SGBD estruturado e um NoSQL assentam no facto dos primeiros terem uma estrutura rígida, serem muito focados na integridade dos dados através de um controlo transacional ACID, o que provoca bastante latência no sistema e escalabilidade vertical. Em suma, permite um sistema robusto ao nível da integridade dos dados, mas com dificuldades em lidar com grandes volumes de dados. Por sua vez, os sistemas NoSQL, têm uma estrutura flexível e proporcionam alta disponibilidade dos dados, no entanto ignoram a integridade dos dados através de um controlo transacional BASE. Finalmente, apresentam a possibilidade de escalabilidade horizontal, que permite alta *performance* no tratamento de grandes volumes de dados.

5. Adequabilidade de uma estrutura NoSQL, na gestão e estruturação de dados AIS, no âmbito do CSM

5.1. Enquadramento dos testes

Os testes realizados têm como objetivo implementar métricas que permitam medir e comparar a *performance* da consulta dos mesmos dados AIS numa infraestrutura relacional (PostGres) e numa NoSQL (MongoDB). As arquiteturas ao nível das capacidades de *hardware* são similares, no que se refere ao número de máquinas, à capacidade de processamento, à capacidade e velocidade de disco e de memória RAM, conforme Quadro 4.

Quadro 4 – Capacidades das infraestruturas de testes

	Arquitetura PostGres	Arquitetura MongoDB
Nº de máquinas	1	1
Processador	Intel® Xeon® CPU E7 – 4850 @ 2.00GHz (64 bits) – 4 cores	Intel® Xeon® CPU E7 – 4850 @ 2.00GHz (64 bits) – 4 cores
Disco	Rack de discos de SSD	Rack de discos de SSD
Memória RAM	64 GB	64 GB

No que se refere ao volume de dados, em ambos os sistemas foram carregados os dados AIS (tipo 1, 2, 3, 18 e 19) recebidos entre 01 de janeiro de 2020 e 31 de dezembro de 2021, o que compreende um volume de dados de 3,06 biliões de registos e que totalizam 590GB.

5.2. Estruturas de dados

Na BD relacional (PostGres) foram desenhadas duas tabelas, a "*information*" (conforme Quadro 5) e a "*position*" (conforme Quadro 6) ligadas entre si (conforme Figura 9).

Quadro 5 – Tabela "*information*"

Campo	Tipo de Dados
id (chave primária)	bigint
instant	timestamp
mmsi	integer
ais_version	smallint
imo	integer
callsign	character(7)
name	character(20)
vessel_type	smallint
length_to_bow	smallint
length_to_stern	smallint
width_to_port	smallint
width_to_starboard	smallint
epfd	smallint
eta_month	smallint
eta_day	smallint
eta_hour	smallint
eta_minute	smallint
draught	real

destination	character(20)
dte_flag	boolean
spare	smallint
nmea	character

Quadro 6 – Tabela "*position*"

Campo	Tipo de Dados
Id (chave primária)	bigint
instant	timestamp
mmsi	integer
lonlat	geometry(Point,4326)
navigational_status	smallint
rate_of_turn	smallint
speed_over_ground	real
position_accuracy	boolean
course_over_ground	real
true_heading	smallint
utc_second	smallint
maneuver_indicator	smallint
spare	smallint
raim_flag	boolean
radio_status	integer
nmea	character
info_fk	bigint

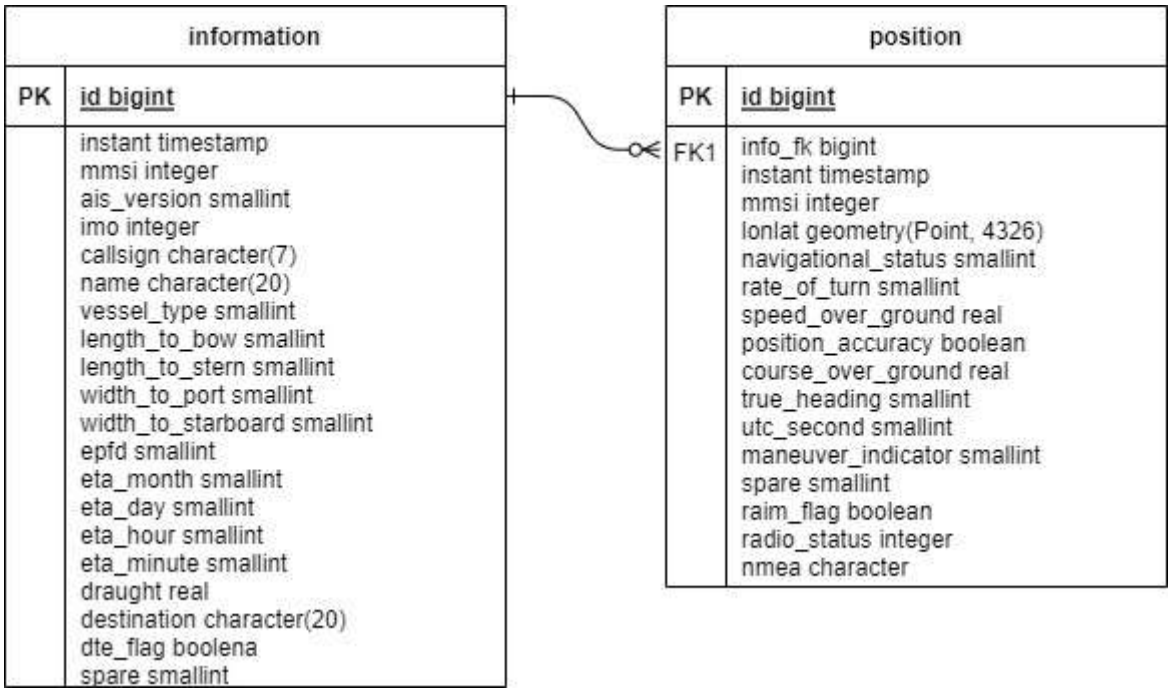

Figura 9 – Relação entre as tabelas "*information*" e "*position*"

Fonte: Autor (2022)

No ambiente não relacional MongoDB foi criado uma estrutura de documento para suportar todos os campos conforme Figura 10.

```
{
    "instant": "",
    "mmsi": "",
    "ais_version": "",
    "imo": "",
    "callsign": "",
    "name": "",
    "vessel_type": "",
    "length_to_bow": "",
    "length_to_stern": "",
    "width_to_port": "",
    "width_to_starboard": "",
    "epfd": "",
    "eta_month": "",
    "eta_day": "",
    "eta_hour": "",
    "eta_minute": "",
    "draught": "",
    "destination": "",
    "dte_flag": "",
    "spare": "",
    "nmea": "",
    "poisitons":
    [
        {
            "lonlat": "",
            "navigational_status": "",
            "rate_of_turn": "",
            "speed_over_ground": "",
            "position_accuracy": "",
            "course_over_ground": "",
            "true_heading": "",
            "utc_second": "",
            "maneuver_indicator": "",
            "spare": "",
            "raim_flag": "",
            "radio_status": ""
        },
        {
            ...
        }
    ]
}
```

Figura 10 – Estrutura do documento em MongoDB

5.3. Métricas

Foram realizadas as seguintes consultas em ambos os ambientes, PostGres e MongoDB, registando o tempo despendido para obtenção das respostas às seguintes consultas:

- Dados (todos os campos) recebidos no dia 01 de janeiro de 2020
- Dados (todos os campos) recebidos no dia 01 de agosto de 2021
- Dados (todos os campos) recebidos em janeiro de 2020
- Dados (todos os campos) recebidos em agosto de 2021
- Dados (todos os campos) recebidos em 2020
- Dados (todos os campos) recebidos em 2021
- Dados (todos os campos) recebidos de 01 janeiro de 2020 a 31 dezembro de 2021

Cada consulta foi efetuada 3 vezes e o valor registado é a média aritmética dos tempos obtidos, mitigando assim a possibilidade de registos de tempos condicionados por alguma menor disponibilidade do sistema naquele exato momento.

5.4. Resultados

No Quadro 7 apresentam-se os resultados obtidos nas 7 consultas efetuadas.

Quadro 7 – Resultados dos testes realizados nas 7 consultas

Consulta: Dados(todos os campos) recebidos em:	Nº de registos	Arquitetura PostGres (em segundos)	Arquitetura MongoDB (em segundos)
01 de janeiro de 2020	3.822.028	192	1,9
01 de agosto de 2021	4.123.567	199	2.1
janeiro de 2020	119.416.557	7680	4,1
agosto de 2021	127.737.528	8255	4,3
2020	1.360.641.968	Indefinido	18,3
2021	1.675.902.567	Indefinido	18,8
01 janeiro de 2020 a 31 dezembro de 2021	3.036.544.535	Indefinido	22,4

Com base nos resultados obtidos pode-se constatar que o tempo de resposta aumenta exponencialmente em função do incremento do número de registos como se pode verificar no Gráfico 1. De salientar que nas consultas que são realizadas com um volume de registos superior a 1 bilião, a arquitetura PostGres deixou de responder ou respondeu em tempo não útil (duração de mais de 1 dia em execução para devolução de resultados), enquanto a arquitetura MongoDB, na consulta com mais registos (01 janeiro de 2020 a 31 dezembro de 2021) devolveu resultados em 22,4 segundos conforme Gráfico 1.

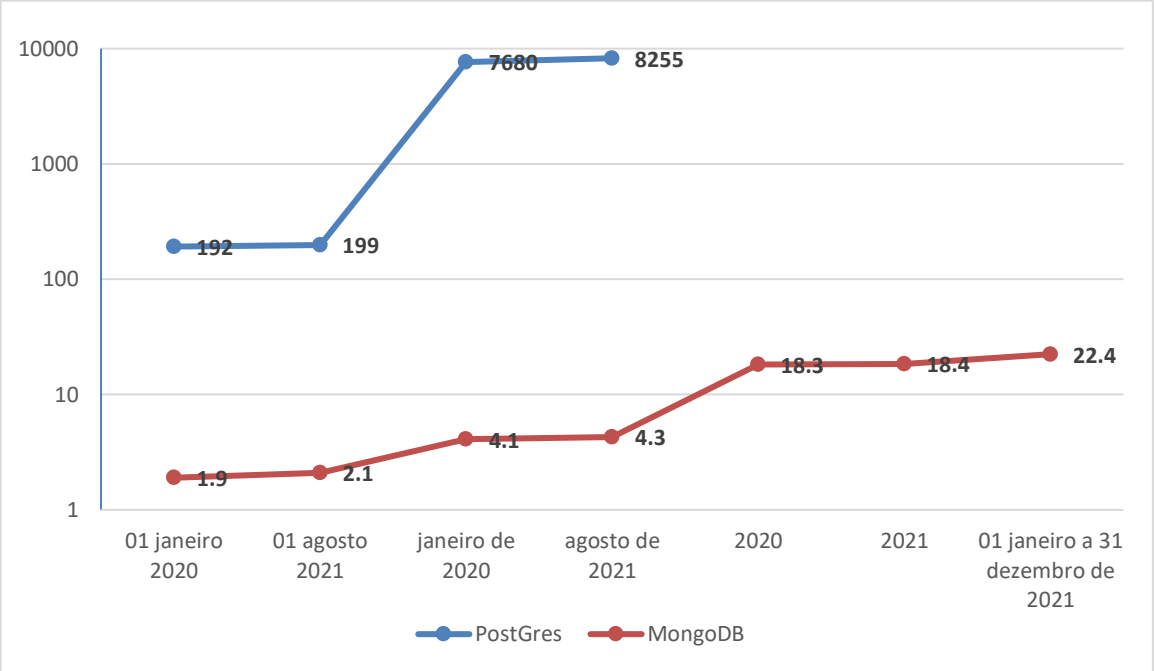

Gráfico 1 – Análise dos tempos obtidos nos testes efetuados 7 consultas (em segundos)

5.5. Síntese conclusiva e resposta à Questão Derivada 2

A infraestrutura de suporte dos dados AIS no âmbito do CSM tem necessariamente de estar adaptada a receber centenas de registos/minuto e cerca de 700 MB/dia, sendo estes dados oriundos de várias fontes e com diferentes tipos de dados, validados quanto à integridade e de valor enorme para os sistemas de apoio à decisão relacionados com o CSM. Assim sendo, os dados AIS compreendem os 5 V´s (Velocidade, Variedade, Volume, Veracidade e Valor), logo estamos perante um ambiente de *Big Data* com uma enorme tendência de crescimento do volume de dados.

Sendo a fonte de dados o AIS, a preocupação com a integridade dos dados é secundária, pois as mensagens transmitidas chegam estruturadas num padrão "rígido", evitando assim "anomalias" nos dados recebidos, o que leva a dispensar aquelas que são as principais vantagens de um sistema de base de dados relacional que assenta num modelo de transações ACID. Optando-se então por uma infraestrutura NoSQL, a escalabilidade horizontal com a utilização de vários nós de dados é algo que se faz com relativa facilidade, com vista a haver uma distribuição da carga associada à gestão e estruturação dos dados pelos diferentes nós. No âmbito do CSM afigura-se mais vantajoso incidir o foco no binómio "Consistência + Particionamento" do que na "Disponibilidade", pois segundo o Teorema de CAP (Eric Brewer, 2012), a vertente "Disponibilidade" já é assegurada em níveis bastante elevados. Assim, os dados permanecem consistentes entre todos os nós, evitando que fiquem

dessincronizados e incoerentes, mas podendo ficar indisponíveis momentaneamente se em algum dos nós ocorrer uma falha. Dos SGBD NoSQL que se enquadram na combinação das propriedades "Consistência + Particionamento", conjugando com o facto do armazenamento de dados ser baseado em documentos modelados utilizando a formatação JSON, que prima pela sua versatibilidade, independência de plataformas e tecnologias, e finalmente pela alta *performance* de processamento, um SGBD NoSQL que se adequa é o MongoDB, conforme Figura 8.

Os testes realizados assentaram numa infraestrutura com um único nó no sistema "MongoDB" em comparação com um sistema relacional "PostGres" com um servidor de dados, no sentido de dotar ambas as infraestruturas de características muito similares. Com base nos resultados obtidos nos testes efetuados no protótipo desenvolvido reforça-se a potencialidade dos sistemas NoSQL, neste caso o "MongoDB", em lidar de forma flexível, dinâmica e com alta *performance* no tratamento de grandes volumes de dados. As métricas avaliadas permitem demonstrar uma considerável melhoria nos tempos de resposta a consultas similares num ambiente NoSQL (MongoDB) em comparação com um ambiente estruturado (PostGres), principalmente quando o número de registos é mais elevado, i.e. acima de 1 bilião de registos.

Com base na análise documental, nos resultados obtidos no protótipo com aplicação de métricas de avaliação, e do conteúdo dos dados das entrevistas semiestruturadas realizadas, em resposta à QD2 *Qual a adequabilidade dos ambientes de base de dados não relacionais para estruturar e gerir dados AIS, no contexto CSM, em sistemas Big Data na MP?*, conclui-se que um sistema NoSQL é o mais apropriado no armazenamento e tratamento de *Big Data*, tal como os dados AIS, pelas razões que se apresentam em seguida:

- Possibilidade de armazenar dados não estruturados, estrutura dinâmica;
- Alto desempenho e escalabilidade horizontal;
- Desenvolvimento relativamente fácil para utilização de dados voláteis;
- Assente num controlo transacional BASE;
- Possibilidade de armazenamento em serviços *cloud*, facilitando a progressão horizontal de recursos conforme o crescente volume de dados assim o exija.

6. Efeito da capacidade da infraestrutura na *performance* do tratamento de *Big Data*, no âmbito do CSM

6.1. Enquadramento dos testes

Os testes desenvolvidos têm como objetivo implementar métricas que permitam medir e comparar a *performance* da consulta dos mesmos dados AIS numa infraestrutura relacional (PostGres com escalabilidade vertical) e numa NoSQL (MongoDB com escalabilidade horizontal). As arquiteturas ao nível das capacidades de *hardware* iniciais são similares (conforme Quadro 4), no que se refere à capacidade de processamento, à capacidade e velocidade de disco e de memória RAM, no entanto foram efetuadas escalabilidades vertical e horizontal conforme Quadro 8.

Quadro 8 – Escalabilidade vertical e horizontal

	Arquitetura PostGres – Escalabilidade Vertical (1 único servidor)	Arquitetura MongoDB – Escalabilidade Horizontal (Nº de servidores)
1º teste	RAM: 64GB Processador (Nº de *Cores*): 4	1
2º teste	RAM: 128GB Processador (Nº de *Cores*): 8	2
3º teste	RAM: 256GB Processador (Nº de *Cores*): 10	4

6.2. Métricas

Em cada um dos 3 testes realizados (conforme Quadro 8) foram realizadas as seguintes consultas em ambos os ambientes, PostGres e MongoDB, registando o tempo despendido para obtenção das respostas às seguintes consultas:

- Dados (todos os campos) recebidos no dia 01 de Janeiro de 2020
- Dados (todos os campos) recebidos em Janeiro de 2020
- Dados (todos os campos) recebidos em 2021
- Dados (todos os campos) recebidos de 01 Janeiro de 2020 a 31 Dezembro de 2021

Cada consulta foi efetuada 3 vezes e o valor registado é a média aritmética dos tempos obtidos, mitigando assim a possibilidade de registos de tempos condicionados por alguma menor disponibilidade do sistema naquele exato momento.

6.3. Resultados

No Quadro 9 apresentam-se os resultados obtidos nas 4 consultas efetuadas em cada um dos 3 testes (conforme Quadro 8).

Quadro 9 – Resultados dos testes obtidos nas 4 consultas

Consulta: Dados(todos os campos) recebidos em …	Nº de registos	Testes	Arquitetura PostGres (em segundos)	Arquitetura MongoDB (em segundos)
01 de janeiro de 2020	3.822.028	1º teste	192	1,9
		2º teste	124	1,5
		3º teste	56	1,2
janeiro de 2020	119.416.557	1º teste	7680	4,1
		2º teste	3455	2,8
		3º teste	1790	1,3
2021	1.675.902.567	1º teste	Indefinido	18,8
		2º teste	Indefinido	12,4
		3º teste	8345	5,8
01 janeiro de 2020 a 31 dezembro de 2021	3.036.544.535	1º teste	Indefinido	22,4
		2º teste	Indefinido	18,5
		3º teste	Indefinido	8,9

Com base nos resultados obtidos pode-se constatar que na arquitetura PostGres (escalabilidade vertical), à medida que se aumentou o número de *cores* do servidor e a memória RAM, os tempos de resposta tiveram um decréscimo significativo, no entanto quando temos mais de 1 bilião de registos o sistema continuou a não responder em tempo útil. Na arquitetura MongoDB (escalabilidade horizontal), à medida que aumentamos o nº de nós (servidores), o tempo de resposta diminuiu consideravelmente, conseguindo obter tempos a baixo dos 10 segundos mesmo com mais de 3 biliões de registos, conforme Gráfico 2.

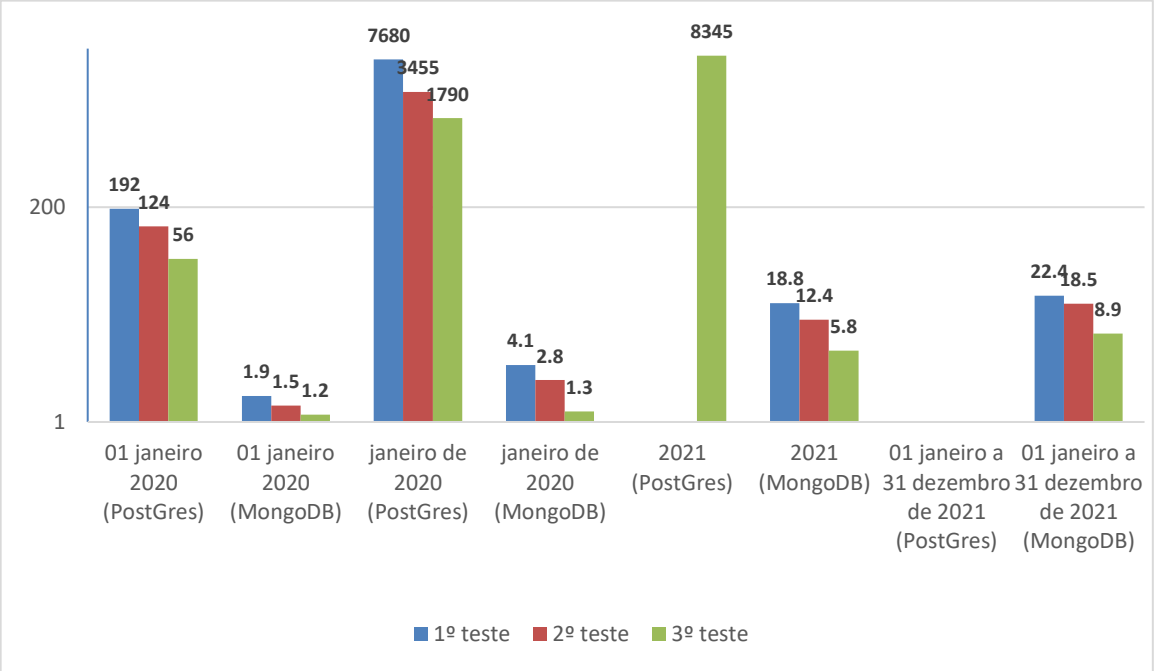

Gráfico 2 – Análise dos tempos obtidos nos testes às 4 consultas (em segundos)

6.4. *Cloud* Vs "*On premisses*"

Um dos dilemas no momento de implementar uma arquitetura de *Big Data* passa por decidir sobre um sistema "*On premisses*" ou "*Cloud*". Um sistema "*On premisses*" é montado na íntegra em máquinas físicas ou virtuais ligadas à infraestrutura interna da organização. Os sistemas "*Cloud*" utilizam uma infraestrutura pré-configurada, com recursos variáveis em função das necessidades do momento e alojados numa infraestrutura externa à organização (Johnsen, 2019), ao contrário do ambiente "*On premisses*" em que o controlo do sistema e dos dados está na íntegra sob a alçada da própria entidade, pois toda a gestão do sistema é realizada internamente e sobre a própria infraestrutura tecnológica (Johnsen, 2019).

Com base nas entrevistas semiestruturadas realizadas a diversas entidades públicas e privadas de reconhecido mérito nacional e internacional (conforme Apêndice F), pode-se constatar que a EDP, o Millenium BCP e o Leroy Merlin têm arquiteturas *Big Data* em *Cloud* e apresentam como mais valias, por exemplo: a disponibilidade dos sistemas, a flexibilidade e atualização imediata de recursos de *hardware* (memória RAM, capacidade de processamento e disco) em função das necessidades do momento com possibilidade de escalabilidade horizontal, e finalmente a otimização de custos em *hardware*, *software* e recursos humanos.

O Banco BIG, a NOS Telecomunicações e a Sonae têm soluções "*On premisses*" e estão em fase de migração ou tencionam migrar para um ambiente *Cloud* no sentido de poderem usufruir das mais valias mencionadas anteriormente. A MP no âmbito do CSM tem em desenvolvimento um grupo de trabalho para estudo da possibilidade de criação de uma arquitetura *Big Data* na *Cloud* (C. Roque, entrevista por e-mail, 9 de junho de 2022).

6.5. Síntese conclusiva e resposta à Questão Derivada 3

Após os testes realizados aplicando as métricas sobre os ambientes "PostGres" (estruturado) e "MongoDB" (NoSQL), conclui-se que a escalabilidade horizontal do "MongoDB" é significativamente mais eficaz que a escalabilidade vertical do "PostGres".

Os testes efetuados no "PostGres" com o aumento de capacidade de processamento e de memória, melhorou significativamente os tempos de resposta, no entanto quando o volume de dados ultrapassa 1 bilião de registos, o sistema raramente consegue responder e quando consegue, devolve os dados em tempo não útil, ou seja, com tempos de espera de várias horas. Em contraponto, o sistema NoSQL conseguiu devolver resultados em menos de 10 segundos, mesmo com mais de 3 biliões de registos.

Com base na análise documental, nos resultados obtidos no protótipo com aplicação de métricas de avaliação, e do conteúdo dos dados das entrevistas semiestruturadas realizadas, em resposta à QD3 *De que forma a capacidade da infraestrutura influencia a performance do tratamento de Big Data, no contexto CSM na MP?*, conclui-se que a escalabilidade horizontal no ambiente "MongoDB" mostrou-se a mais eficiente, pois permite aumentar o número de nós da infraestrutura, de forma a distribuir a carga, proporcionando assim uma *performance* elevada, flexível, disponível e em tempo real, mesmo quando o volume de dados vai crescendo a um ritmo elevado. A escalabilidade vertical aplicada nos SGBD's estruturados permite um incremento de *performance*, mas este torna-se insuficiente quando estão envolvidos mais de 1 bilião de registos.

Um sistema *Cloud* para além de proporcionar a otimização de custos em *hardware*, *software* e recursos humanos pode trazer algumas questões de confidencialidade quanto aos dados, devido ao facto de estarem a ser geridos numa infraestrutura externa (Johnsen, 2019), no entanto no contexto do AIS não existe esta condicionante pois estamos perante dados "não confidenciais" (Marinha, 2012). O sistema *Cloud* como base de um ambiente NoSQL permite uma escalabilidade horizontal de uma forma ágil, imediata e sem quebras de serviço.

7. Proposta de definição de um sistema de gestão e estruturação de grandes volumes de dados oriundos do AIS, no âmbito do CSM na MP, e resposta à Questão Central

Pelo até aqui analisado, e em resposta à QC, *Como tornar um sistema de Big Data mais eficiente ao nível da gestão e estruturação de dados AIS, no contexto CSM?*, apresenta-se no Quadro 10, a proposta para um modelo de gestão e estruturação de *Big Data* refletindo a adoção de serviços *Cloud*.

Quadro 10 – Proposta para um modelo eficiente de gestão e estruturação de dados AIS (*Big Data*) no contexto do CSM

Ponto decisor	Opção	Vantagens
Arquitetura estruturada VS NoSQL	NoSQL	Conforme ponto 5.5 (resposta à QD2).
Controlo Transacional	BASE	O BASE permite flexibilidade, simplicidade e alto desempenho do sistema. Apesar de proporcionar uma fraca consistência, nos dados AIS não se torna um constrangimento, pois estes são recebidos através de uma estrutura rígida que não permite erros, e para além disso poderá sempre existir um *schema* de validação desenhado programaticamente no momento da importação dos dados. Conforme ponto 4.3 (resposta à QD1).
Escalabilidade	Horizontal	A escalabilidade horizontal permite um acréscimo de *performance* da infraestrutura em função do crescimento do volume de dados. Conforme ponto 6.5 (resposta à QD3).
CAP	Consistência + Particionamento	"Consistência + Particionamento" em detrimento da "Disponibilidade".
SGBD	MongoDB (ou SGBD com características idênticas)	O MongoDB é um SGBD baseado em documentos e que dá um reforço extra na "Consistência + Particionamento", sendo que a "Disponibilidade" por si só, num ambiente NoSQL já é bastante elevada. Assim sendo, consegue-se alta capacidade nas propriedades CAP, no entanto a "Disponibilidade" é a propriedade menos preponderante num sistema AIS, pois é de extrema importância o sistema estar consistente e com alta *performance* em detrimento de poder não estar disponível num muito curto espaço de tempo. Conforme ponto 6.5 (resposta à QD3).
Licenciamento	Sem custos	Não apresenta custos de licenciamento para aquisição e *upgrades* de versão.
Cloud VS "*On premisses*"	Cloud	Facilidade de incremento de capacidade de hardware através de escalabilidade horizontal em tempo real. Otimização de custos de *hardware*, *software* e recursos humanos. Os dados AIS são "não reservados" pelo que a questão da possível falta de confidencialidade dos dados em ambientes *cloud* não é um problema. Conforme pontos 6.4 e 6.5 (resposta à QD3).
Departamento dedicado	2 ou 3 pessoas	Conforme ponto 6.4 e análise das entrevistas (Apêndice F) um ambiente NoSQL em *Cloud* não necessita de um elevado número de recursos humanos especializados, pois os encargos de implementação e manutenção passam para o lado do fornecedor do serviço.

8. Conclusões

O desenvolvimento tecnológico verificado nos últimos anos potenciou a geração, processamento e armazenamento de grandes volumes de dados em formatos diferentes. Dados estes que devem ter um tratamento quase em tempo real de forma a disponibilizá-los com uma *performance* adequada para a finalidade a que se destinam.

É na prossecução deste paradigma que surge a exigência de uma gestão e estruturação de *Big Data* no âmbito do CSM, que permita a disponibilização dos dados AIS no momento exato, no local certo e de forma atempada.

A pesquisa teve como base um estudo de caso, seguindo um raciocínio indutivo com uma estratégia qualitativa, sem invalidar a apresentação de dados quantitativos para ilustrar alguns aspetos, quando considerado pertinente, definindo-se como questão central a identificação de um modelo que torne um sistema *Big Data* mais eficiente ao nível da gestão e estruturação de dados AIS no contexto do CSM. Foram identificados diferentes vetores: qual o tipo de arquitetura (estruturada ou NoSQL), o SGBD mais adequado, o modo de escalabilidade, a adequabilidade de um sistema *Cloud* e a dimensão de um departamento de suporte.

A recolha dos dados no sentido de permitir atingir os OE e responder às QD, fundou-se, qualitativamente, através da análise documental e das entrevistas formalizadas, e quantitativamente na aplicação de métricas em testes realizados sobre um protótipo desenvolvido. Assim, foi possível concluir em resposta à QD1, que se torna vantajoso a implementação de um sistema NoSQL para armazenamento e tratamento de *Big Data*, pois permite alto desempenho e escalabilidade, um desenvolvimento relativamente fácil, tem uma estrutura dinâmica, suporta uma alta volatilidade de dados, é assente num controlo transacional BASE e permite o armazenamento dos dados em serviços *Cloud*. Um sistema de gestão de bases de dados relacional, é mais rígido, garante a integridade dos dados, mas tem muita dificuldade em lidar de forma eficiente com dados voláteis e em grandes volumes.

No que respeita à QD2, sendo os dados AIS considerados *Big Data* pelos 5 V's (Volume, Velocidade, Variedade, Veracidade e Valor), considera-se que o sistema MongoDB (ou outro SGBD NoSQL de características similares) é adequado para sustentar uma estrutura de dados AIS, pela sua flexibilidade, capacidade de escalabilidade horizontal, alta *performance* com biliões de registos e por dar importância máxima à combinação das propriedades "Consistência + Particionamento", pois a "Disponibilidade" já é um ponto forte dos sistemas NoSQL.

Em relação à QD3, o aumento da capacidade de processamento e de memória, através de escalabilidade vertical (SGBD´s estruturados) e o incremento do número de servidores através da escalabilidade horizontal (NoSQL) foi notório, através dos testes realizados, que ambos os sistemas melhoraram a *performance* em função do aumento de recursos. No entanto a escalabilidade horizontal mostrou-se muito mais capaz de acompanhar um grande volume de dados de AIS, apresentando *performances* muito aceitáveis, mesmo executando pesquisas com biliões de dados, ao contrário do que a escalabilidade vertical proporcionou no sistema estruturado, em que mesmo com o aumento de recursos não foi possível a obtenção de resultados em tempo útil. Assim sendo, conclui-se que a escalabilidade horizontal, possível nos sistemas NoSQL, é adequada para a gestão e estruturação dos dados AIS e que por sua vez podem ser geridos numa arquitetura *Cloud* com administração dos recursos em tempo real.

Posto isto, em resposta à QC, verifica-se que para tornar um sistema de *Big Data* mais eficiente no contexto do CSM na MP, é vantajoso a utilização de um SGBD NoSQL, dispondo de um sistema transacional do tipo BASE, com escalabilidade horizontal, dando importância à "Consistência + Particionamento", alojado num ambiente *Cloud*, gerido por um departamento interno e dedicado de 2 ou 3 pessoas.

Decorrentes deste estudo, assinala-se como principal contributo para o conhecimento a identificação de um conjunto de evidências metodológicas e cientificamente validadas para propor um modelo de gestão e estruturação de *Big Data* na realidade da MP.

Identifica-se como limitação a este estudo, que lhe é alheia e não condicionou as conclusões apresentadas, a condição ainda pouco maturada da estratégia *Cloud* para a Administração Pública, refletida em mecanismos, ferramentas e normativos ainda parcamente desenvolvidos, bem como a impossibilidade de implementar o protótipo de testes num ambiente *Cloud* devido aos custos associados.

Concernente a estudos futuros, afigura-se pertinente analisar o tratamento diferenciado de informação classificada, quer em modelos *Cloud* públicos, quer privados, no que se refere a sistemas *Big Data*, assim como a formação e constituição de equipas especializadas em gestão e estruturação de sistemas de grandes volumes de dados.

Como recomendação de ordem prática, sugere-se à MP, a realização de um estudo prévio com vista à implementação das medidas apresentadas no presente estudo, contribuindo também para uma melhor estruturação da estratégia para a *Cloud*.

Referências Bibliográficas

Alexandre, J., & Cavique, L. (2013). *NoSql no suporte à análise de grande volume de dados*. Revista de Ciências de Computação [Página *online*]. Retirado de http://repositorioaberto.uab.pt/handle/10400.2/3091

Amdahl, G. (1967). *Validity of the single processor approach to achieving large scale computing capabilities*. AFIPS.

Berryhill, J., Heang, K., Clogher, R., & McBride, K. (2019). *Hello, World: Artificial intelligence and its use in the public sector*. OECD Working Papers on Public Governance, 36, OECD Publishing.

Blokdyk, G. (2020). *NoSQL Databases A complete Guide*. 5StarCooks.

Bock, B. & Schrage, J. (2002). *Denormalization guidelines for base and transaction tables*. ACM SIGCSE Bulletin, Volume 34.

Brewer, E. (2012). *CAP Twelve Years Later: How the "Rules" Have Changed* [Página *online*]. Retirado de https://ieeexplore.ieee.org/document/6133253

Cattell, R. (2010). *Scalable SQL and NoSQL Data Stores* [Versão PDF]. Retirado de http://www.cattell.net/datastores/Datastores.pdf

CEMA. (2011). *Diretiva de Política Naval 2011*. Diretiva. Lisboa: Marinha Portuguesa.

Centro de Computação Gráfica – Investigação & Desenvolvimento Tecnológico. (2019). *Cloud computing vs fog computing vs. edge computing na era da internet das coisas industrial* [Página *online*]. Retirado de https://www.ccg.pt/cloud-computing-vs-fog-computing-vs-edge-computing-na-internet-das-coisas-industrial/

Codd, E. (1970). *A Relational Model Of Data for Large Shared Data Banks*. Volume 13.

Cook, J. (2019). *ACID versus BASE for database transactions* [Página *online*]. Retirado de http://www.johndcook.com/blog/2019/07/06/brewer-cap-theorem-base/

Diana, M., & Gerosa, M. (2010). *NOSQL na Web 2.0: Um Estudo Comparativo de Bancos*.

Dubois, A., & Gadde, L. (2002). *Systematic Combining: An abductive approach to case research. Journal of Business Research*.

Edlich, P. (2018). *NoSQL* [Página *online*]. Retirado de http://nosql-database.org/

Espinosa, J., Kaisler, S., Armour, F., & Money, W. (2019). *Big Data Redux: New Issues and Challenges Moving Forward*. Proceedings of the 52nd Hawaii International Conference on System Sciences. Maui, Hawai.

Fachada, C. P. A., Ranhola, N. M. B., Marreiros, J. P. R., & Santos, L. A. B. (2020). *Normas de Autor no IUM* (3.ª Ed., revista e atualizada). *IUM Atualidade*, 7. Lisboa: Instituto Universitário Militar.

Fidge, C. (1988). *Timestamps in Message-Passing Systems That Preserve the Partial Ordering.* Theoretical Computer Science - TCS.

Freixo, M. (2009). *Metodologia Científica* (1ª ed.). Lisboa: Instituto Piaget.

Gudivada, V., Rao, D., & Raghavan, V. (2014). *NoSQL Systems for Big Data Management. Services (SERVICES)*, 2014 IEEE World Congress.

Guémar, H. (2015). *Cours NoSQL* [Página *online*]. Retirado de https://hguemar.fedorapeople.org/nosql/nosql01/#1

Harrison, G. (2015). *Next Generation Databases*. Apress.

Haughian, G. (2014). *Benchmarking replication in nosql data stores*. Dissertação, Imperial College London.

Henderson, K. (2000). *The Guru's Guide to Transact-SQL*. Addison-Wesley Professional.

IMO (1998). *Resolution MSC.74, Annex 3. Recommendation on Performance Standards for an Universal Shipborne Automatic Identification Systems (AIS)*. International Maritime Organization [Versão PDF]. Retirado de https://wwwcdn.imo.org/localresources/en/OurWork/Safety/Documents/AIS/Resolution%20MSC.74(69).pdf

IMO (2002*). Resolution A.917. Guidelines for the onboard operational use of a Shipborne Automatic Identification System*s (AIS). International Maritime Organization [Versão PDF]. Retirado de https://wwwcdn.imo.org/localresources/en/KnowledgeCentre/IndexofIMOResolutions/AssemblyDocuments/A.917(22).pdf

Information Systems Audit and Control Association. (2015). *ISACA Glossary of Terms - Portuguese* (3rd Edition). Information Systems Audit and Control Association [Página *online*]. Retirado de https://www.isaca.org/resources/glossary

Johnsen, F. (2019). *Towards Big Data in the Tactical Domain*. *NATO Science and Technology Organization.*

Kuznetsov, S., & Poskonin, A. (2014). *NoSQL data management systems. Journal Programming and Computing Software*, Volume 40, pp. 323-332.

Lam, P. (2011). *Hadoop in Action*. 1ª ed. Manning Publications.

Leavitt, N. (2010). *Will NoSQL Databases Live Up to Their Promise?*. Computer, Volume 43, pp. 12 - 14.

Marinha Portuguesa (2011). *Conceito de conhecimento situacional marítimo*. IOA114. Lisboa: CEMA e AMN.

Marinha Portuguesa (2021). *Projeto APEC SIFICAP*. Lisboa: DAGI.

McCreary, D., & Kelly, A. (2014). *Making Sense of NoSQL: A guide for managers and the rest of us.* Manning Publications Co.

NATO (2014). *NATO Cyber Defence Taxonomy and Definitions (AC/322-N(2014)0072)*. Brussels: Consultation, Command and Control (C3) Board.

Nayak, A., Poryia, A., & Poojary, D. (2013). *Type of NOSQL databases and its comparison with relational databases.* International Journal of Applied Information Systems.

Nurseitov, N., Paulson, M., Reynolds, R., & Izurieta, C. (2009). *Comparison of JSON and XML Data Interchange Formats: A Case Study*. San Francisco, California.

Okcan, A., & Riedewald, M. (2011). *Processing Theta-Joins using MapReduce*. SIGMOD 2011, Issue Conference: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 949-960.

PostgreSQL. (2021). *PostgreSQL - Manual Archive* [Página *online*]. Retirado de http://www.postgresql.org/docs/manuals/archive

Quivy, R., & Campenhoudt, L. (1998). *Manual de Investigação em Ciências Sociais* (2ª Ed.). Lisboa: Gradiva.

Ramakrishnan, R., & Gehrke, J. (2003). *Database Management Systems*. 3ª Edição.

Rybinski, H. (1987). *On First-Order-Logic Databases*. ACM Transactions on Database Systems, Volume 12, pp. 325-349.

Santos, L.A.B., & Lima, J.M.M. (Coord.) (2019). *Orientações metodológicas para a elaboração de trabalhos de investigação* (2.ª ed., revista e atualizada). Cadernos do IUM, 8. Lisboa: IUM.

Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research Methods for Business Students* (8ª ed.). Harlow, UK: Pearson Education Ltd.

Vieira, M. (2012). *Bancos de Dados NoSQL: Conceitos, Ferramentas, Linguagens e Estudos de Casos no Contexto de Big Data*.

Vieira, M., Figueiredo, J., Liberatti, G., & Viebrantz, A. (2012). *Bancos de Dados NoSql: conceitos, ferramentas, linguagens e estudos de casos no contexto de Big data* [Versão PDF]. Retirado de http://data.ime.usp.br/sbbd2012/artigos/pdfs/sbbd_min_01.pdf

Welling, L., & Thompson, L. (2003). *PHP and MySql Web Development*. 2ª Edição

Yin, R. K. (2018). *Case Study Research and Applications: Design and Methods* (6ª ed.). Thousand Oaks, USA: SAGE Publications Inc.

Zuckerberg, M. (2021). *The face of Facebook* [Página *online*]. Retirado de https://www.newyorker.com/magazine/2010/09/20/the-face-of-facebook

Apêndice A – Corpo de Conceitos

ACID – Propriedades dos Sistemas de Gestão de Bases de Dados relacionais: Atomicidade, Consistência, Isolamento e Durabilidade (Johnsen, 2019).

BASE – Propriedades dos Sistemas de Gestão de Bases de Dados NoSQL: Basicamente disponível, Estado leve e Eventualmente consistente (Johnsen, 2019).

Ciclo de vida da informação – Planeamento, recolha, criação ou geração de informação, organização, recuperação, utilização, acessibilidade, transmissão, armazenamento, proteção e disposição (NATO, 2004).

Computação em nuvem (*Cloud Computing*) – pode ser definida como um modelo de disponibilização e utilização de Tecnologias de Informação e Comunicação, que permite o acesso remoto, através da internet, a um leque de recursos de computação partilhados em forma de serviços (Centro de Computação Gráfica - Investigação & Desenvolvimento Tecnológico, 2019).

Dados – elementos discretos, não organizados, compostos por números, palavras, sons ou imagens independentes, que podem ser facilmente estruturados, mas que por si só não conduzem à compreensão dum facto ou situação (NATO, 2014).

Informação – Qualquer comunicação ou representação de conhecimento (factos, dados ou opiniões) utilizando qualquer meio ou forma (audiovisual, narrativa, cartográfica, numérica, gráfica, textual) (NATO, 2007).

Infraestrutura *cloud* - Todos os recursos de hardware e software que permitem disponibilizar serviços de acordo com as características do modelo *cloud* (NATO, 2014). Compreende uma camada física, que engloba servidores e componentes de armazenamento e de rede, e uma camada de abstração, composta por software implementado sobre a componente física (NATO, 2014).

Infraestrutura TIC – Conjunto de recursos de *hardware*, *software* e *networking* que permitem a uma organização disponibilizar soluções de tecnologias de informação e comunicações aos seus utilizadores (NATO, 2014).

Machine learning – Conjunto de técnicas que permitem uma aprendizagem automatizada, sem intervenção humana, por parte de máquinas, com base em padrões e inferências, que pode ser efetuada através de treino, fornecendo à máquina um conjunto de situações corretas para assimilação, ou por situações de tentativa/erro, interpretando regras fornecidas (Berryhill et al., 2019). É considerada uma subárea da IA (Berryhill et al., 2019).

Memória RAM - é um tipo de memória que permite a leitura e a escrita, utilizada como memória primária em sistemas eletrónicos digitais. A RAM (*Random Access Memory*) é um componente essencial não apenas nos computadores pessoais, mas em qualquer tipo de computador, pois é onde basicamente ficam armazenados os programas básicos operacionais. Por mais que exista espaço de armazenamento disponível, na forma de um HDD (*Hard Disk Drive*) ou memória flash, é sempre necessária uma certa quantidade de RAM (Information Systems Audit and Control Association, 2015).

Sistema de Apoio à Decisão (DSS) - Os Sistemas de Apoio à Decisão são sistemas de informação interativos e flexíveis, baseados em computador, para gerir a tomada de decisão, no caso de problemas semiestruturados (Information Systems Audit and Control Association, 2015).

Unidade Central de Processamento (CPU) - também conhecida como processador, é a parte de um sistema computacional, que realiza as instruções de um programa de computador, para executar a aritmética básica, lógica, e a entrada e saída de dados (Information Systems Audit and Control Association, 2015).

VHF - termo referido para *Very High Frequency*, ou seja, trata-se de uma Frequência Muito Alta, determinada para faixas de radiofrequência de 30 a 300 MHz (IMO, 2018).

Apêndice B – Modelo de análise

Quadro 11 – Modelo de análise

Gestão e estruturação de *Big Data*			
Filosofia	Pragmatismo	**Raciocínio**	Indutivo
Estratégia	Qualitativa com suporte em dados quantitativos	**Desenho de pesquisa**	Estudo de Caso
Horizonte temporal	01 janeiro de 2020 a 31 dezembro de 2021	**Técnicas de recolha e análise**	Análise documental e Entrevistas semiestruturadas
Objeto de investigação			
Gestão e estruturação de grandes volumes de dados (*Big Data*) oriundos do AIS no contexto do CSM			
Objetivo geral	Objetivos específicos	Questão central	Questões derivadas
Selecionar contributos para otimizar a gestão e estruturação de dados AIS, na MP, no contexto do CSM na era do *Big Data*	**OE1:** Comparar os sistemas de bases de dados relacionais e não relacionais no contexto de *Big Data.* **OE2:** Analisar a adequação de ambientes de base de dados não-relacionais, em sistemas de *Big Data*, no que se refere a gestão, estruturação, análise e *performance* de dados AIS, no contexto do CSM na MP. **OE3:** Analisar a influência da capacidade da infraestrutura na *performance* do tratamento de *Big Data*, no contexto CSM na MP.	Como tornar um sistema de *Big Data* mais eficiente ao nível da gestão e estruturação de dados AIS, no contexto CSM?	**QD1:** Quais as principais diferenças entre a utilização de um sistema de base de dados relacional e de um não relacional no contexto de *Big Data*? **QD2:** Qual a adequabilidade dos ambientes de base de dados não relacionais para estruturar e gerir dados AIS, no contexto CSM, em sistemas *Big Data* na MP? **QD3:** De que forma a capacidade da infraestrutura influencia a *performance* do tratamento de *Big Data*, no contexto CSM na MP?

Apêndice C – Tipos de mensagem AIS

Quadro 12 – Tipos de mensagem AIS

#	Tipo de mensagem
1	Scheduled position report (Class A shipborne mobile equipment)
2	Position reportAIS Message
3	Position reportAIS Message
4	Base station reportAIS Message
5	Static and voyage related dataAIS Message
6	Binary addressed messageAIS Message
7	Binary acknowledgementAIS Message
8	Binary broadcast messageAIS Message
9	Standard SAR aircraft position reportAIS Message
10	UTC/date inquiryAIS Message
11	UTC/date responseAIS Message
12	Addressed safety related messageAIS Message
13	Safety related acknowledgementAIS Message
14	Safety related broadcast messageAIS Message
15	InterrogationAIS Message
16	Assignment mode commandAIS Message
17	DGNSS broadcast binary messageAIS Message
18	Standard Class B equipment position reportAIS Message
19	Extended Class B equipment position reportAIS Message
20	Data link management messageAIS Message
21	Aids-to-Navigation reportAIS Message
22	Channel managementAIS Message
23	Group assignment commandAIS Message
24	Static data reportAIS Message
25	Single slot binary messageAIS Message
26	Multiple slot binary message with Communications StateAIS Message
27	Position report for long range applications

Fonte: IMO (2002)

Apêndice D – Identificação dos entrevistados

Quadro 13– Identificação dos entrevistados

Cargo	Titular	Fase: Exploratório (E) ou "Campo" (C)	Área de *expertise*
Superintendente das Tecnologias de Informação	Comodoro Cancela Roque	C	Marinha
Diretor da Direção de Sistemas de Informação	Coronel Carlos Passos	C	Secretaria-Geral do Ministério da Defesa Nacional
DBA Analista de Sistemas	Drª Ana Marta Grade	C	Banco BIG
Scrum Master in Big Data	Eng. Marco Loureço	C	EDP
Big Data Project Manager	Dr Ricardo Figueiredo	C	LeRoy Merlin
Chief Techical Officer	Eng Miguel Gaspar	C	Meanify
IT Director - Architecture & Transformation	Eng. Nuno Reis	C	Millenium BCP
Big Data Engineer	Eng. José Andrade Santos	C	Natixis
Data Engineer	Eng. Ralph Venâncio	C	NOS Telecomunicações
Administrador de Sistemas Big Data	Eng. Timóteo Pereira	C	Sonae

Apêndice E – Guião de entrevista semiestruturada

INSTITUTO UNIVERSITÁRIO MILITAR
DEPARTAMENTO DE ESTUDOS PÓS-GRADUADOS
CURSO PROMOÇÃO A OFICIAL SUPERIOR
2021/2022 – 2.ª Edição

GUIÃO DE ENTREVISTA SEMIESTRUTURADA

O Instituto Universitário Militar (IUM), no âmbito das suas atribuições, leciona anualmente o Curso de Promoção a Oficial Superior (CPOS), frequentado por auditores dos diferentes Ramos das Forças Armadas, da Guarda Nacional Republicana (GNR) e dos Oficiais de Países Amigos.

O auditor da Marinha, Primeiro-Tenente, ST-EINF, António Guerreiro Pacheco, frequenta a 2.º edição do CPOS 2021/2022, encontrando-se a desenvolver um Trabalho de Investigação Individual, com o tema "Gestão e estruturação de *Big Data*", sob a orientação do Senhor Capitão-de-fragata EN-AEL, Araújo Costa.

De acordo com o projeto de investigação apresentado, o auditor tem como objetivo propor a definição de um sistema de gestão e estruturação de grandes volumes de dados oriundos do AIS, para posterior análise no âmbito do Conhecimento Situacional Marítimo na Marinha Portuguesa.

O tema proposto, está alinhado com uma necessidade premente da Marinha Portuguesa em definir uma infraestrutura de Big *Data* que permita uma utilização eficiente dos dados, com vista à sua aplicação em áreas operacionais e no apoio à decisão.

Neste sentido, a recolha de informação junto de entidades que tenham grandes volumes de dados armazenados, e necessidades de gestão e estruturação de Big Data, é extremamente relevante para sustentar a credibilidade dos argumentos que servirão de base ao trabalho.

Face ao anteriormente exposto, o contributo de V/ Exa. constitui-se como uma mais-valia para a investigação em curso e, consequentemente, para a qualidade das conclusões a alcançar e das recomendações a efetuar.

Neste sentido, solicito autorização para que as suas respostas, ou excertos das mesmas, devidamente contextualizados, sejam citados e identificados. Caso, em alternativa, não deseje ser identificado, será assegurada a salvaguarda do anonimato e confidencialidade das respostas prestadas.

Ressalva-se, para os devidos efeitos, que os resultados da investigação terão o grau de classificação de segurança NÃO CLASSIFICADO.

Muito obrigado pela colaboração.

Perguntas:

Por infraestrutura de gestão e estruturação de dados entende-se o conjunto de recursos (infraestruturas, *hardware* e *software*) utilizados para processar e armazenar os dados de uma Organização ao longo do seu ciclo de vida.

1 - O(s) repositório(s) de dados existentes na sua Organização têm volume suficiente para poderem ser considerados *Big Data* ?

2 - Caso já exista uma infraestrutura de gestão e estruturação de Big Data como caracteriza os seguintes pontos:

2.1 - Volume de dados (aproximado) em nº de registos (ex. registos/ano) e em Gigabytes (ex. GB/ano)

2.2 - Infraestrutura assente em modelos relacionais ou NoSQL

2.3 - Infraestrutura baseada em tecnologias proprietárias ou livres de licenciamento

2.4 - *Cloud Computing*, *On premises* ou híbrido

2.4.1 – Caso a sua organização tenha serviços "*Cloud Computing*" como os caracteriza quanto a:

2.4.1.1 – Custos (em €/por ano), considerando que neste modelo serão cobrados os recursos utilizados *(pay as you use*), contrapondo com o investimento necessário à aquisição, sustentação e gestão de recursos próprios

2.4.1.2 – Vantagens/desvantagens

2.5 – Grau de satisfação com a infraestrutura implementada no que se refere a:

Gestão e estruturação de Big Data

2.5.1 – Disponibilidade

2.5.2 – Flexibilidade e escalabilidade de recursos

2.5.3 – Segurança

2.6 – Melhorias/alterações a implementar na infraestrutura existente

2.7 – Qual a solução/marca solução Big Data implementada na sua organização (ex. Azure, Seagate, Amazon, Apache Hadoop, etc) ?

2.8 – A sua organização tem um departamento dedicado à gestão e estruturação da infraestrutura de *Big Data* ? Em caso afirmativo, a equipa é constituída por quantos elementos ?

3 - Caso ainda não exista uma infraestrutura de gestão e estruturação de *Big Data*, a sua organização tem planeado a edificação de uma infraestrutura para tratamento de grandes volumes de dados ? Em caso afirmativo, quais as linhas de ação e decisões tecnológicas para os pontos 2.2, 2.3 e 2.4.

Apêndice F – Análise das entrevistas semiestruturadas

Quadro 14 – Análise de entrevistas semiestruturadas

Pergunta	Marinha	Secretaria-Geral do Ministério da Defesa Nacional	Banco Big	EDP	Meanify
1	Sendo o "volume suficiente" um valor subjetivo e difícil de determinar, considero que o volume de dados relacionados com dados do CSM na Marinha, nomeadamente do AIS, pode-se considerar como sendo suficiente para Big Data.	Sim (BD Oracle SIG DN).	Sim, contudo a volumetria do mesmo não é considerada excessivamente grande.	Sim.	Sim, trabalhamos com clientes que tem muitos Petabytes de informação que necessita ser processada.
2.1	O AIS disponível no CSM da Marinha provém de diferentes fontes, nomeadamente: estações próprias e parcerias externas (MSSIS, APRAM, Portos dos Açores, EMSA e VTS) O total de mensagens AIS recebidas pelas várias fontes é de cerca 20 mil milhões (19.938.666.055) por ano, totalizando cerca de 1,165 TB (1.163.387.985.169 bytes).	933 biliões de registos por ano de crescimento. 350gb por ano de crescimento.	Conjunto de dados dividido por 8 instancias(servidores). Cada instância consoante a sua criticidade apresenta um um conjunto. Na totalidade conseguimos reunir cerca de 2617 Gb / ~ 3Tb.	Há casos onde estamos a falar de mais de 4 milhões de registos por quarto de hora o que nos remete para a carga não de GBs/ano mas sim de TB/ano.	Na ordem dos múltiplos petabytes.
2.2	A principal infraestrutura existente no CSM da Marinha foca-se no processamento e partilha de informação, não se focando muito no armazenamento da informação. As principais razões para não existir foco no armazenamento são: a) Falta de capacidade de armazenamento (hardware) b) Falta de definição de políticas de armazenamento e de histórico	Assente em Modelo relacional.	Modelos Relacionais. Linguagem T-SQL e ORACLE	Assente em ambos os modelos para possibilitar analítica mais tradicional e também mecânicas mais evoluídas como por exemplo, machine learning.	NoSql.

	c) Falta de definição da utilização esperada para os dados armazenados (tendo em vista a otimização do armazenamento para facilitar a sua utilização).				
2.3	Toda a infraestrutura é baseada em tecnologias livres de licenciamento. A base de dados relacional usada é o Postgres SQL. Na análise é utilizada, principalmente, a ferramenta MATLAB, que é licenciada.	É baseada em tecnologias proprietárias.	Proprietárias, com licenciamentos.	Infraestrutura baseada em tecnologia cloud com serviços maioritariamente PaaS.	Ambas.
2.4	On Premisses.	On Premisses.	On Premisses.	Cloud.	On Premisses.
2.4.1.1	Não aplicável.	Não aplicável.	Não aplicável. Não temos estes serviços.	---	
2.4.1.2	Não aplicável.	Não aplicável.	No entanto ainda estamos a considerar a melhoria ao nível do SSMS Azure para implementação de Cloud Computing, visto que o departamento encarregue focado nesse âmbito é recente e pequeno.	As desvantagens estão maioritariamente ligadas à latência pois atualmente ainda se lida com realidades onde as fontes de dados estão on premisses o que leva perdas de performance.	Não aplicável.
2.5.1	A solução implementada na Marinha foi desenvolvida internamente com recurso a software de fonte aberta e de licenciamento livre, tendo vários anos de solidez comprovada e disponibilidade no que diz respeito a recolha, processamento e partilha de informação AIS. Os constrangimentos mencionados no ponto 2.2., inviabilizam realizar analises temporais superiores a 2 anos.	---	Rápido e acessível. A nível de backups utilizamos soluções que nos permitem ter a sincronização e assincronização directa com um tempo de indisponibilidade muito reduzido.	Para alguns casos específicos muito pontuais existe a necessidade de implementar outras mecânicas que aumentem a disponibilidade.	Alta disponibilidade apenas para determinados componentes da plataforma que são críticos ao negócio.

2.5.2	A escalabilidade e flexibilidade da solução implementada pode considerar-se mista, ou seja, é escalável e flexível, mas é necessário algum trabalho e know-how para o fazer. Tratando-se de uma solução on-permises acrescentar recursos é um processo demorado, verificando-se um hiato de tempo significativo entre a colocação da necessidade, a aquisição de hardware/software e a sua disponibilização.	Algo limitada.	Algo limitada.	Fácil escalabilidade e flexibilidade.	Escalabilidade limitada visto que está on premise.
2.5.3	Uma vez que toda a infraestrutura é on-premises e os dados AIS estão em alguns casos disponíveis ao público em geral por outras fontes.	---	Todos os departamentos têm um user/role que permite aceder aos schemas e base dados estritamente necessárias com permissões de leitura apenas. Temos user aplicacional para equipas de desenvolvimento com algumas limitações.	Bastante satisfeito.	Requisitos de segurança bastante apertados, relacionados com acessos, compliance e auditorias.
2.6	Aumentar a capacidade de armazenamento para processar individualmente cada uma das fontes de informação e completar a mesma com meta-dados relativos a: a) Estatísticas do fluxo de dados b) Estatísticas da localização dos contactos recebidos, cobertura da estação, etc… c) Atribuição de nível de confiança da informação Aumentar a capacidade computacional de forma a permitir o	Aumentar a flexibilidade e performance de acesso aos dados.	Implementação do SSMS AZURE. Implementação de instancia somente para utilização do SSIS para não criar sobrecarga de dados e performance. Implementação com a dinâmica crescente dos dados de Cloud Computing.	Necessário rever algumas politicas de rede e de segurança para garantir melhor performance nos serviços utilizados.	Atualmente existe uma heterogeneidade na estrutura de metadata da informação armazenada e seria útil alguma homogeneização da estrutura dos metadados.

	processamento de grandes volumes de dados (vários anos) em simultâneo, em tempo reduzido.				
2.7	A solução big data implementada usa BD relacional Postgres SQL não existindo uma solução cloud.	---	Azure.	Azure.	Solução proprietária, Pentaho Data Integration, Hitachi Content Intelligence, Hitachi Content Platform.
2.8	Neste momento não existe. Existe uma intenção de preparar uma estratégia da Marinha para a *cloud* onde este deverá ser um dos pontos a abordar.	Sim.	Equipa pequena e recente a fim de ser aumentada. Até ao fim do ano anterior era constituída por 3 elementos e passou a ser 1 elemento somente, com a saída dos outros 2.	Não, são utilizados serviços com fornecedores para isso.	9 só dedicados a manutenção.
3	No âmbito do projeto APEC-SIFICAP encontra-se em curso o desenvolvimento de uma infraestrutura Big Data na cloud para apoio às atividades de fiscalização da pesca. No futuro a Marinha, dependendo do serviço em particular, poderá utilizar de modelos de dados relacionais ou não relacionais ou mesmo mista, apontando-se para o uso de software livre de licenciamento e a implementação por decidir se será na cloud, on-premises ou mista.	Sendo considerada a atual estrutura de SIG algo a melhorar, pensamos em evoluir para SAP S/4HANA, um sistema com tecnologias inteligentes incorporadas, incluindo AI, machine learning e funções analíticas avançadas. É uma infraestrutura baseada em tecnologias proprietárias, NoSQL e será implementada em sistema Hibrido.	Não aplicável.	---	Não aplicável.

Fonte: Construído a partir de C. Roque (*op. cit.*), de C. Passos (*op. cit.*), de A. Grade (*op. cit.*), de M. Lourenço (*op. cit.*) e de M. Gaspar (*op. cit.*).

Quadro 15 – Análise de entrevistas semiestruturadas

Pergunta	Millenium BCP	Natixis	NOS	LeRoy Merlin	Sonae
1	Sim.	Sim.	Sim.	Sim, usamos um vasto repositório de dados com os assets que espelham a atividade das soluções data driven da empresa, bem como agregamos informação Analitica de ecommerce e ainda alguma atividade monotorizada das lojas agregado todo o data lake e respetivo data warehouse numa infraestrutura GCP usando a solução google BIG QUERY e outras ferramentas da google big data.	Sim, temos cerca de 117 servidores dedicados apenas à equipa de big data.
2.1	Volume de dados criado situar-se-á perto dos 20 TB/ano. Volume de registos é no plano dos muitos milhões/dia, mas não consigo precisar os valores.	Cerca 1365GB/ano apenas num processo (diariamente temos cerca de 10000 processos).	1,5 TB/ano.	Dado estarmos a usar infraestrutura em Cloud GCP usando servidores da UE, não existe uma necessidade de gestão de infra, mas sim uma optimização da utilização da solução afim de garantir uma economia na utilização das ferramentas de big data.	Temos cerca de 1500 registos ao ano e usamos cerca de 200TB/ano em dados processados.
2.2	Temos ambos. Usamos bastante NoSQL, mas acreditamos que o futuro é mais baseado em NewSQL. Aliás, o acesso a repos NoSQL é muitas vezes efetuado usando wire SQL.	Modelo relacional.	Aproximadamente 15 modelos relacionais, um modelo dimensional com 40 data marts e um data lake no modelo de lake house.	Temos por exemplo, publicação de informação usando firebase (não relacional) faz também parte das ferramentas GCP neste caso do conjunto de soluções Firebase da google.	---
2.3	A nova arquitetura de dados é baseada sobretudo em tecnologias open-source, existindo contudo a necessidade de existência de suporte corporativo às soluções e que muitas vezes implicam a utilização de soluções	Ambas, no entanto cada vez mais livre licenciamento.	Híbrido, mas maioritariamente open source.	Proprietárias.	---

	ligeiramente adaptadas (Ex: kafka – Confluent; Presto/Trino – Starburst). Também utilizamos soluções PaaS que caiem num espaço intermédio nesta questão.				
2.4	A nova arquitetura de dados é totalmente suportada em cloud. O DW "financeiro e histórico" e solução de reporting suportado em Mainframe que é on-prem.	---	Atualmente em on premise, mas iniciando projeto go to cloud.	Cloud Computing	On premisses.
2.4.1.1	Sentimos que o valor do custo da cloud, mesmo em cenários otimizados é tendencialmente superior às soluções on-prem.	---	Ainda não mensurável.	Os consumos GCP estão diretamente dependentes do storage usado, e tambem maioritariamente do processamento e eficiência das queries executadas para minimizar custos temos que ter atenção quanto à segmentação das tabelas. (ver a política de preços do BQ da google).	Em termos de custos gastamos cerca de 1Milhão / ano com as aplicações big data.
2.4.1.2	Vantagens: Agilidade, Tempo de Setup e facilidade de escala, bem como potenciação da automação. Desvantagens: Custo (Alto & incerto).	---	Ainda não mensurável.	As vantagens prendem-se com o facto da ausência de manutenção de infraestrutura, alta disponibilidade, utilização em Equipa como gestão de acessos integrada, ferramentas, self manage services, de inbound requisição e publicação de estruturas, etc...	Segurança será a nossa grande vantagem. Lentidão será a desvantagem escolhida.
2.5.1	Alta.	Satisfeitos, mesmo em caso stress (DRP onde apenas trabalhamos com metade do balanceamento)	Um pouco insatisfeito.	Infraestrutura cloud com alta disponibilidade.	Satisfeito.
2.5.2	Alta.	---	Insatisfeito.	Escalável sem limites de storage com alta eficiência.	Satisfeito.
2.5.3	Alta.	---	Um pouco satisfeito.	Garantia de segurança da infraestrutura cloud da google, garantia de autenticação dos utilizadores através de contas Gsuite	Muito satisfeito.

				e gestão de IAM dos projetos gcp com criação de service accounts para partilhas de informação e integração Machine to Machine.	
2.6	Melhoria nos processos de automação e resiliência – deployment de novos serviços, casos de falha, etc.	---	Alterações na orquestração dos processos, melhorias na modelagem do data lake e alteração do processo de ingestão de dados.	Melhorias na utilização de queries para que necessitem de menos processamento de dados e mais performantes.	Passar as nossas infrastruturas para cloud , em vez de onprem.
2.7	A cloud que usamos é Azure, no entanto, não temos uma solução, temos um conjunto de soluções suportadas nesta infraestrutura e com o aumento da maturidade a adoção de soluões IaaS assume um papel fundamental.	Apache Hadoop, Cloudera.	Apache Hadoop, Cloudera.	GCP.	Hadoop , microstrategy e IBM.
2.8	Existem múltiplos departamentos com responsabilidades neste espaço, desde a equipa de suporte à cloud, que tem uma sub-equipa focada em Data e múltiplas equipas de arquitetura e desenvolvimento que suportam a solução. Focados exclusivamente na Infraestrutura de Data temos cerca de 6 pessoas.	Sim tem, cerca 20 pessoas no entanto este número varia bastante.	Sim, 8 pessoas.	Sim, a nível de cloud ops, esta atividade é fornecida pelo Grupo de empresas adeo.	Sim, somos cerca de 4 elementos , sendo que procuramos um 5 elemento para preencher mais a equipa.
3	---	---	O Data Office estipulou um dia da semana, onde o fluxo do trabalho semanal é interrompido. Esse dia é dedicado a workouts com diferentes temas, como melhorias de processos e cloud roadmap.	Não aplicável.	---

Fonte: Construído a partir de N. Reis (*op. cit.*), de J.A. Santos (*op. cit.*), de R. Venâncio (*op. cit.*), de R. Figueiredo (*op. cit.*) e de T. Pereira (*op. cit.*).