Lisbon School of Economics & Management
Universidade de Lisboa

MESTRADO

GESTÃO DE SISTEMAS DE INFORMAÇÃO

TRABALHO FINAL DE MESTRADO

DISSERTAÇÃO

O IMPACTO DO *DATA LAKE* NOS SISTEMAS DE *REPORTING* DO SETOR BANCÁRIO

MARIANA ALEXANDRA CRAVO RAMOS

OUTUBRO - 2021

Lisbon School of Economics & Management
Universidade de Lisboa

MESTRADO
GESTÃO DE SISTEMAS DE INFORMAÇÃO

TRABALHO FINAL DE MESTRADO
DISSERTAÇÃO

O IMPACTO DO *DATA LAKE* NOS SISTEMAS DE *REPORTING* DO SETOR BANCÁRIO

MARIANA ALEXANDRA CRAVO RAMOS

SUPERVISÃO:
JESUALDO FERNANDES

OUTUBRO - 2021

mens sana in corpore sano

GLOSSÁRIO

BCBS - *Basel Committee on Banking Supervision*

DQ – *Data Quality*

FSB - Conselho de Estabilidade Financeira

G-SII - Instituição de Importância Sistémica Global

IFRS - *International Financial Reporting Standards*

JSON - *JavaScript Object Notation*

O-SII - Outras Instituições de Importância Sistémica

SI – Sistemas de Informação

TI – Tecnologias de Informação

XML - *Extensible Markup Language*

RESUMO

O setor bancário é caracterizado pela forte regulamentação e supervisão, bem como pela existência de inúmeros sistemas de informação heterogéneos, os quais foram sido adquiridos e adaptados ao longo do tempo. Por outro lado, a enorme quantidade de dados, definida por *Big Data*, trouxe novos desafios ao processamento e integração dos dados e, consequentemente, à sua qualidade. O mercado está a mudar e o setor bancário não é exceção. Para dar resposta à falta de qualidade dos dados, o Basel Committee on Banking Supervision (2013), (BCBS), definiu 14 princípios, nomeados como BCBS 239, que devem ser atendidos por este setor. No entanto, estes princípios são vagos e não existe nenhuma orientação sobre os métodos a utilizar para agir em conformidade com o regulador. Por este motivo, as questões de investigação são: como é que o *Data Lake* impacta a concretização dos princípios BCBS 239? De que forma está a organização em conformidade com os princípios do BCBS 239? Quais são as vantagens e os desafios do *Data Lake* do ponto de vista da organização em estudo?

Para dar resposta, foi desenvolvido um *case study* com recurso a entrevistas semiestruturadas a colaboradores de um banco que atua no mercado português. Para a análise de dados foi utilizado o programa *MAXQDA*.

Os resultados obtidos permitem identificar que a existência de várias fontes de dados e de processos não automatizados são os desafios com maior expressão nos atuais processos de *reporting*. Além disso, é possível reconhecer que as principais vantagens do *Data Lake* advêm do facto de este ser um repositório único, potencializar a uniformização de conceitos de negócio e melhorar a eficiência organizacional. Por outro lado, a ambiguidade de conceitos é um obstáculo que tem de ser ultrapassado, de forma a garantir a existência de um modelo de governo de dados.

Em suma, o *Data Lake* é um excelente meio para cumprir o BCBS 239, em especial, no desenvolvimento da arquitetura de dados e infraestrutura de TI, melhorando, consequentemente, a qualidade dos dados.

PALAVRAS-CHAVE: Banca; BCBS 239; *Big Data; Data Lake*; Gestão da Qualidade dos Dados; *Reporting*.

CÓDIGOS DE CLASSIFICAÇÃO JEL: M15; M48.

ABSTRACT

The banking sector is characterized by strong regulation and supervision, as well as the existence of numerous heterogeneous information systems, which have been acquired and adapted over time. On the other hand, the enormous amount of data, conceptualized as *Big Data*, brought new challenges to data processing and integration processes and, consequently, its quality. The market is changing, and the banking sector is no exception. Therefore, to respond to the lack of data quality, the Basel Committee on Banking Supervision (2013), (BCBS), defined the BCBS 239, which compilates 14 principles that the companies of this sector must meet. However, these principles are vague, and there is no guidance on the methods to be used to comply with regulators. For this reason, the research questions are: how does Data Lake impact the implementation of BCBS 239 principles? How is the organization compliant with the BCBS 239 principles? What are the advantages and challenges of the Data Lake from the point of view of the organization?

A case study was developed using semi-structured interviews with current employees of a bank operating in the portuguese market to investigate this research topic. Moreover, the data analysis was conducted via the MAXQDA program.

The results obtained allowed to acknowledge that the existence of a wide range of data sources and non-automated processes is the most significant challenge in current reporting processes. Furthermore, it was possible to recognize that the main advantages of Data Lake focus on the fact that (1) it is a single repository; (2) it enhances the standardization of business concepts and consequently, (3) improves organizational efficiency. On the contrary, the present study also revealed that the ambiguity of concepts is a critical obstacle that must be overcome to guarantee the existence of a data governance model.

In short, Data Lake is a powerful way to meet BCBS 239, in particular, in developing data architecture and IT infrastructure, thereby improving data quality.

KEYWORDS: Banking; BCBS 239; Big Data; Data Lake; Data Quality Management; Reporting.

JEL CODES: M15; M48.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

AGRADECIMENTOS

Nem fazia sentido começar de outra forma. A todos os que fizeram acontecer, mesmo sem saberem, a cada um de vocês, o meu obrigada do fundo do coração!

Ao ISEG, à minha casa. Obrigada por fazeres os sorrisos crescerem! Um obrigada, em especial, ao Bernardo Reis. Obrigada por viveres e transmitires tanto o espírito iseguiano. Também a cada um que partilhou o nosso pátio comigo, obrigada!

Aos entrevistados, à organização e a todos os envolvidos no processo. Muito obrigada! Sem vocês, não teria sido possível.

Ao professor Jesualdo Fernandes. Obrigada pela disponibilidade, conhecimento, soluções e conselhos em todo o percurso. Acima de tudo, obrigada pela coordenação e dignificação do nosso mestrado.

À b.logic. À Constança, ao Mickaël e à Patrícia. Obrigada por darem sentido à palavra trabalho de equipa, obrigada pelas reuniões marcadas para depois de almoço, obrigada pela inúmeras pausas para o café.

Ao Luís Filipe. Obrigada por toda a preocupação, experiência e motivação. Obrigada por me mostrar, todos os dias, a importância dos "porquês" em todas as situações.

A todos os meus amigos. Obrigada pelos abraços, sorrisos e palhaçadas. Em especial, à Abigail, à Mara e ao Rafael, um obrigada muito grande por me fazerem acreditar que o tempo e espaço não existe para nós. Obrigada por serem os melhores amigos de sempre!

À minha família, a toda a minha família. Obrigada, são a base de tudo! Aos avós, à mãe, ao pai, ao mano, à madrinha, obrigada por acreditarem em mim. Sem vocês não estaria aqui, obrigada por tudo! À Beatriz, obrigada por estares presente em todos os momentos, sem exceção, mesmo que seja a dormir. Ao João, obrigada por ocupares o sofá todos os dias, assim não há tentações.

"*Sorri… que amanhã só resta a dor de não estar ao pé de ti.*" (Tuna Económicas, 2012). Vamos voltar a encontrar-nos, ISEG. Fizeste-me "*grab the future*", obrigada!

1. INTRODUÇÃO

"*Um banco internacional típico pode ter até 70.000 sistemas de informação e mais de 250.000 folhas de cálculo*" (Butler, 2017, p.51). Desta forma, o setor bancário é caracterizado pelos seus sistemas de informação heterogéneos, completamente desintegrados e com duplicação de dados, tendo em conta a sua história, nomeadamente fusões e aquisições de outros bancos (Ćurko et al., 2007).

Neste sentido, em 2013, surge o BCBS 239 com o objetivo de harmonizar os sistemas de informação e melhorar a qualidade dos dados, através de requisitos regulatórios que direcionam a gestão de dados das instituições financeiras (Butler, 2017). No entanto, o BCBS apenas define princípios e não existe nenhuma orientação sobre os métodos a utilizar para os fazer cumprir (Orgeldinger, 2018).

Por esta razão, o presente estudo procura estudar o desenvolvimento do *Data Lake* como um método específico, dando resposta às seguintes questões de investigação: como é que o *Data Lake* impacta a concretização dos princípios BCBS 239? De que forma está a organização em conformidade com os princípios do BCBS 239? Quais são as vantagens e os desafios do *Data Lake* do ponto de vista da organização em estudo? Para conseguir encontrar respostas às questões, foram definidos os seguintes objetivos de investigação: (1) analisar a situação atual dos processos de *reporting*; (2) identificar as dificuldades sentidas atualmente; (3) compreender as motivações para a implementação do *Data Lake*; (4) identificar os objetivos da implementação do *Data Lake* na organização; (5) identificar as expetativas para a implementação do *Data Lake*; (6) estudar vantagens e desvantagens do *Data Lake*; e, (7) relacionar as vantagens do *Data Lake* com os princípios do BCBS 239;

Assim, aplicou-se a metodologia de *case study* num banco, recorrendo a entrevistas para a recolha de dados. A investigação está estruturada em 6 capítulos: (1) a introdução, onde se contextualiza o tema e onde se identificam as questões e os objetivos de investigação; (2) a revisão de literatura aprofunda os temas de *reporting*, aplicados à banca, e do *Data Lake*; (3) a metodologia, na qual se pormenoriza a recolha e análise dos dados; (4) a apresentação dos resultados obtidos, através de gráficos e frases proferidas pelos entrevistados; (5) a discussão dos resultados com base na literatura existente; e, por último, (6) a conclusão do estudo, bem como as suas limitações e investigações futuras.

2. REVISÃO DE LITERATURA

2.1. Reporting

O mercado está a mudar e o setor bancário não é exceção. Os bancos são constantemente pressionados por clientes à procura de novos produtos e soluções, mas também pelas entidades reguladoras à procura de respostas (Ćurko et al., 2007).

A crise financeira de 2008 foi o mote para o crescimento da regulação bancária, ao demonstrar que as Tecnologias de Informação (TI) e Arquitetura de Dados não são adequadas para a gestão do risco financeiro (Basel Committee on Banking Supervision, 2013). A definição de normas internacionais de reporte financeiro, as IFRS (*International Financial Reporting Standards*), por exemplo, permitem a comparação fidedigna entre dois bancos. No setor bancário, os principais objetivos da regulação são diminuir os riscos a que os credores e o sistema financeiro estão expostos, evitando, assim, uma crise financeira (Barth & Landsman, 2010).

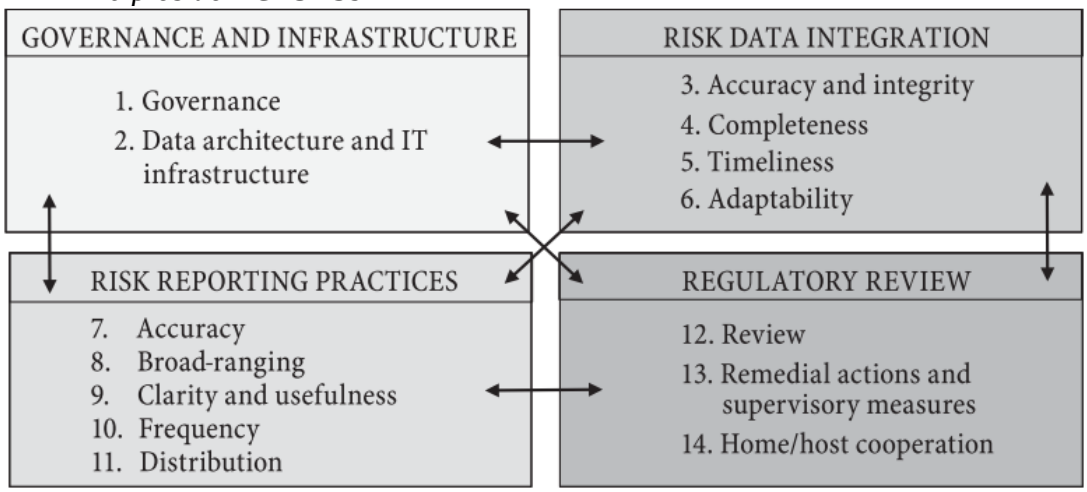
O Conselho de Estabilidade Financeira (FSB) acredita que uma melhoria da capacidade de agregação dos dados, leva a benefícios na disponibilidade de informação, facilitando a tomada de decisão; à redução da probabilidade de má gestão de risco e a melhor qualidade no planeamento estratégico. Ciente destas consequências e com o objetivo de fortalecer a capacidade de agregação dos dados, o FSB disponibiliza diversas iniciativas para apoiar e desenvolver a estabilidade financeira, motivando o desenvolvimento do BCBS 239, "Príncipios para a agregação eficaz de dados de risco" (Basel Committee on Banking Supervision, 2013).

2.1.1. BCBS 239

Os 14 príncipios definidos pelo BCBS (*Basel Committee on Banking Supervision*) podem ser agregados em quatro áreas mais abrangentes: *governance* e infraestrutura, integração dos dados de risco, práticas do *reporting* de risco e revisão regulatória (Orgeldinger, 2018) e são caracterizados por:

Figura 2.1

Princípios do BCBS 239



Fonte: Orgeldinger (2018, p.58)

1. Governance – Os recursos de agregação de dados devem ser consistentes com outros príncipios e orientações do BCBS. A estrutura de dados deve estar definida em conformidade com as políticas da empresa sobre a confidencialidade, integridade e disponibilidade dos dados, ao mesmo tempo que as práticas de *reporting* devem ser documentadas e validadas, garantindo que os processos correspondem ao perfil de risco do banco. Por outro lado, a administração do banco deve procurar e compreender as causas que impedem a total agregação dos dados, deve avaliar o impacto que uma tomada de decisão, por exemplo, o desenvolvimento de um produto novo ou a compra/venda de entidades, tem nos recursos de agregação dos dados, e, por último, deve identificar quais são os dados críticos para agregação dos dados de risco (Basel Committee on Banking Supervision, 2013).

2. Arquitetura de Dados e Infraestrutura de TI – A arquitetura de dados e a infraestrutura de TI devem ser planeadas, construídas e mantidas com o objetivo de dar resposta às necessidades de agregação dos dados e aos restantes príncipios, independentemente da situação atual, seja ela normal ou de crise. O banco deve estabelecer taxonomias, caracterizar os dados, usar identificadores únicos, integrar os dados de todo o grupo bancário e garantir que existem controlos adequados para todo o ciclo de vida de um dado, desde a sua correta inserção no *front office*, até à sua atualização (Basel Committee on Banking Supervision, 2013).

3. Exatidão e Integridade – Um banco deve ser capaz de gerar dados precisos e confiáveis. Para minimizar a probabilidade de erros, os dados devem ser agregados numa base de dados automatizada e devem ser comparados com os dados contabilísticos, de forma a assegurar que os dados são precisos. No entanto, a intervenção humana pode ser apropriada para processos que necessitem de análise e julgamento profissional, motivando a controlos eficazes nestes casos. Por último, os supervisores esperam que o banco documente e explique todos os seus processos, descrevendo soluções alternativas, a criticidade da precisão da agregação dos dados e planos de ação para dados com baixa qualidade (Basel Committee on Banking Supervision, 2013).

4. Completude – Um banco deve ser capaz de capturar e agregar todos os dados de risco do seu grupo, com o nível de detalhe requerido pelo reporte em questão. Os dados de risco não têm de ser expressos em métricas comuns, mas os recursos de agregação devem ser os mesmos, independentemente dos sistemas de informação (SI) utilizados. Neste sentido, os supervisores esperam que os dados reportados sejam materialmente completos (Basel Committee on Banking Supervision, 2013).

5. Tempestividade – Um banco deve ser capaz de agregar dados atualizados em tempo útil, sem descuidar dos princípios relativos à exatidão, integridade e adaptabilidade. O tempo útil depende de vários fatores, como a volatilidade do risco, o perfil de risco e a frequência do reporte. A frequência da necessidade dos dados aumenta em situações de crise e os recursos de agregação devem conseguir dar resposta atempada para estas possíveis situações (Basel Committee on Banking Supervision, 2013).

6. Adaptabilidade – Um banco deve ser capaz de agregar dados de risco de maneira a responder às diversas solicitações de reporte, sejam estas causadas por mudanças nas necessidades internas, por situações de crise ou por consultas da supervisão. Por este motivo, os processos de agregação devem ser flexíveis e personalizáveis, conforme a necessidade do utilizador, bem como a incorporação de novos produtos, mudanças nos requisitos regulatórios e/ou outros fatores que impactem no perfil de risco do banco (Basel Committee on Banking Supervision, 2013).

7. Exatidão – Os relatórios devem transmitir os dados agregados de maneira exata. Por este motivo, os diversos relatórios devem ser reconciliados e validados. Os processos devem ser capazes de identificar, descrever e explicar erros nos dados ou na sua

integridade. Para que tal aconteça, é importante existir um inventário com a descrição das relações que devem ser validadas (Basel Committee on Banking Supervision, 2013).

8. Abrangência – Os relatórios devem abranger todas as áreas de risco dentro da organização. A profundidade e o âmbito dos relatórios devem ser coerentes com o tamanho e complexidade das operações e do perfil de risco do banco, bem como com os requisitos dos destinatários (Basel Committee on Banking Supervision, 2013).

9. Clareza e Utilidade – Os relatórios devem conseguir transmitir informação de uma forma clara e concisa, facilitando a sua leitura. No entanto, devem incluir informação suficiente e relevante para a tomada de decisão (Basel Committee on Banking Supervision, 2013).

10. Frequência – A administração do banco deve definir qual a frequência de recolha, agregação e distribuição de relatórios internos de gestão de risco. Os requisitos da frequência dependem de fatores como as necessidades da receção dos dados, a natureza do risco e a volatilidade do risco (Basel Committee on Banking Supervision, 2013).

11. Distribuição – Os relatórios devem ser distribuídos a todas as partes relevantes, assegurando a confidencialidade dos dados. Os supervisores esperam que os bancos confirmem regularmente se as partes interessadas estão a receber todos os relatórios no período definido (Basel Committee on Banking Supervision, 2013).

12. Revisão – Os supervisores devem rever e avaliar a conformidade dos bancos com os 11 princípios já definidos, através de solicitações ocasionais de informações, às quais o banco deve responder dentro de prazos curtos (Basel Committee on Banking Supervision, 2013).

13. Ações Corretivas e Medidas de Supervisão - Os supervisores devem usar ferramentas e recursos apropriados para exigirem ações corretivas, eficazes e oportunas, de modo a resolver deficiências nos recursos de agregação de dados de risco do banco (Basel Committee on Banking Supervision, 2013).

14. Cooperação - Os supervisores devem cooperar com os supervisores relevantes de outras jurisdições no que toca à supervisão e à revisão dos princípios definidos, bem como a implementação de qualquer ação corretiva. Os supervisores devem discutir as suas experiências de qualidade de dados com as diferentes partes do grupo, de forma a

identificar preocupações significantes e a agir proativamente na sua mitigação (Basel Committee on Banking Supervision, 2013).

Em favor dos príncipios apresentados, o Comité acredita que os investimentos executados pelos bancos serão compensados, a longo prazo, pela melhoria da capacidade de agregação dos dados. Caso a entidade envolvida alargue as sinergias adquiridas a outros processos, este benefício será sentido em maior escala (Basel Committee on Banking Supervision, 2013). Não obstante, é possível que os bancos não consigam atender à conformidade dos princípios, visto que estes são vagos e demonstram a lacuna existente entre a teoria e a prática (Orgeldinger, 2018).

Resumindo, os princípios têm como objetivo auxiliar na melhoria da intensidade e eficácia da supervisão bancária e, por outro lado, permitir uma resolução bancária menos agressiva, devido à melhor capacidade de agregação dos dados de risco (Basel Committee on Banking Supervision, 2013).

2.1.2. Big Data

O termo *Big Data* surgiu, em 2015, com o intuito de definir uma enorme quantidade de dados que não podia ser processada, devido à sua complexidade (Bedeley & Iyer, 2014). Estes dados são mais difíceis de recolher, tratar e integrar, visto que há uma grande proporção de dados não estruturados (Cai & Zhu, 2020).

Os dados não estruturados surgiram com a nova realidade e são objetos como *e-mails*, fotografias, *posts* nas redes sociais, segmentos de vídeo e/ou áudio, *etc*. Por outro lado, os dados que podem ser armazenados em forma tabular são dados estruturados, como números, caracteres e datas (Hoffer et al., 2017). O termo "dado" refere-se a factos relativos a objetos e/ou eventos que têm importância para o utilizador, enquanto que as suas características, os "atributos", identificam as várias ocorrências (Hoffer et al., 2017). Neste seguimento, os dados estruturados são associados a bases de dados mais tradicionais, como o *data warehouse*. Estas ferramentas utilizam o *schema on write*, de forma a garantir o alto nível de qualidade dos dados, através da definição do modelo de dados *a priori* (Hoffer et al., 2017). Em contrapartida, as tecnologias *Big Data* servem-se da abordagem *schema on read* para assegurar a integração de dados heterogêneos. Neste s*chema*, normalmente baseado em *JavaScript Object Notation* (JSON) ou *Extensible*

Markup Language (XML), não existe um modelo predefinido de dados, sendo este apenas desenvolvido perante a leitura e o uso dos mesmos (Hoffer et al., 2017).

Segundo Bedeley e Iyer (2014), Cai e Zhu (2020) e Alshboul et al. (2015), o *Big Data* tem características únicas, que são conhecidas como os 4 V's: volume – porque existe uma elevada quantidade de dados; velocidade - os dados estão sempre a ser gerados e devem ser tratados em *real-time*, para evitar que fiquem desatualizados; variedade – existem diferentes formatos de dados que são recolhidos das mais variadas fontes; valor – capacidade de gerar *insights* e conhecimento para o processo de tomada de decisão.

Devido às características do *Big Data*, é difícil quantificar a qualidade dos dados (Cai & Zhu, 2020). Segundo a International Organization for Standardization (2005), a qualidade é definida como o nível em que um conjunto de características cumpre os requisitos propostos. Neste sentido, para mensurar a qualidade dos dados, foram definidas cinco dimensões universais do *Data Quality* (DQ) – disponibilidade, usabilidade, confiabilidade, relevância e qualidade de apresentação. A disponibilidade caracteriza-se pela acessibilidade dos dados a qualquer hora. A usabilidade define a utilidade dos dados, se estes respondem, ou não, às necessidades dos *users*. A confiabilidade refere-se à exatidão e consistência dos dados. A relevância descreve a relação entre a expetativa do utilizador e o conteúdo da *data*. Por último, mas não menos importante, a qualidade de apresentação permite a compreensão e legibilidade dos dados (Cai & Zhu, 2020).

Hoje em dia, as empresas recolhem e armazenam cada vez mais dados, aumentado o risco da má qualidade dos dados e realçando a importância da gestão eficiente de dados. A má qualidade dos dados afeta a organização a diferentes níveis, como por exemplo, no processo de tomada de decisão num nível estratégico (Moges et al., 2013). As ações de melhoria da DQ carecem de uma observação holística dos processos, de modo a identificar os erros e as suas causas (Moges et al., 2013). No entanto, a maioria das instituições financeiras desconhece os seus problemas de qualidade de dados. Por este motivo, estas ações são motivadas principalmente por ação dos reguladores (Moges et al., 2013).

2.1.3. Desafios Atuais

A qualidade dos dados depende diretamente do ambiente de quem os manuseia, incluindo processos e utilizadores (Cai & Zhu, 2020; Moges et al., 2013).

Infelizmente, os regulamentos e as regras publicadas não são concretas, nem fáceis de interpretar, originando lacunas entre a intenção dos requisitos regulatórios e a interpretação dos mesmos pelas organizações (Butler, 2017).

Outro motivo que tem impacto na qualidade dos dados são os recursos inadequados para a integração e exploração dos dados, sejam eles financeiros, humanos ou tecnológicos (Stein & Morrison, 2014). Os processos são repetidos ao longo dos vários departamentos das empresas, podendo originar informações diferentes para os mesmos dados (Butler, 2017; Moges et al., 2013). Além disso, os erros têm de ser identificados para serem corrigidos posteriormente (Moges et al., 2013).

Adicionalmente, a indústria bancária contém dados em várias fontes, provocando erros e inconsistências nos dados (Cai & Zhu, 2020; Chaudhuri & Dayal, 1997; Kadadi et al., 2014). Por um lado, existem dados idênticos em vários ficheiros, o que leva a diferentes significados para o mesmo conceito e, por vezes, diferentes correções para o mesmo dado (Butler, 2017; Moges et al., 2013). Em particular, apesar da automação dos processos, existem muitos dados que são inseridos manualmente nos sistemas de informação do banco, aumentando, assim, o risco operacional. Os inputs manuais são considerados como uma das principais causas para a baixa qualidade dos dados (Moges et al., 2013).

Do mesmo modo, a falta de documentação é um desafio muito presente, porque a maioria do conhecimento é tácito, perdendo-se quando a pessoa em causa sai da organização (Moges et al., 2013). Em acréscimo, também não existem padrões, nem dicionários de dados (Butler, 2017).

Por último, o acesso à informação também é um desafio identificado. Não só pelo facto da recolha de dados ser feita por vários departamentos, como também pelo facto de quem recolhe aos dados, quando existe uma direção de sistemas de informação que os disponibiliza, poder descartar dados que são relevantes (Moges et al., 2013). Desta forma, é necessário otimizar as consultas de dados, bem como a sua integração (Cai & Zhu, 2020).

Deste modo, o presente estudo procura compreender a situação atual do *reporting* na organização em estudo, bem como os desafios a que está exposta, considerando-se como uma dimensão em estudo.

2.2. Data Lake

Com o aparecimento do *Big Data*, a entrega de informação e conhecimento tornou-se um desafio para as empresas. A necessidade de armazenar grandes volumes de dados potencializou o aparecimento do *Data Lake* (Madera & Laurent, 2016). Segundo Laurent e Madera (2016), o *Data Lake* é definido como uma metodologia para lidar com dados brutos, sem qualquer tratamento prévio, aumentando a flexibilidade e *performance* dos mesmos.

2.2.1. Objetivos

Neste sentido, o *Data Lake* tem como principal objetivo o armazenamento dos dados num só repositório, ao invés de repositórios heterogêneos (Mehmood et al., 2019), mitigando a criação de silos de informação, ou seja, os dados deixam de ser geridos departamentalmente (Dama International, 2017; Stein & Morrison, 2014).

Em acréscimo, o *Data Lake* também proporciona a acessibilidade e integração dos dados, ao reunir elevados volumes, resultando em *insights* valiosos para a organização e descoberta de informações (Stein & Morrison, 2014).

No entanto, cada setor tem um potencial uso do *Data Lake* e o foco deve ser sempre um problema de negócio, o qual deve ser respondido através da implementação do *Data Lake* (Moges et al., 2013; Stein & Morrison, 2014). No setor bancário, o forte grau de regulamentação a que este está exposto pode ser o problema de negócio a ser resolvido (Stein & Morrison, 2014).

Assim, para depreender quais são os objetivos da implementação do *Data Lake* na organização em estudo, identificaram-se os objetivos como uma dimensão de estudo na análise de dados.

2.2.2. Estrutura

O *Data Lake* armazena todos os dados no seu formato original sem integração ou modificação dos mesmos. Deste modo, mesmo que os dados não acrescentem valor naquele exato momento, podem ser explorados no futuro, potencializando a identificação de padrões, *insights*, tendências e/ou incoerências. Isto só é possível graças ao *schema on read*, já mencionado, que garante que os dados estão sempre acessíveis (Singh & Ahmad, 2019). Além disso, o projeto do *Data Lake* deve ser flexível e especifico para cada

organização, onde a arquitetura do mesmo deve estar assente em 3 camadas: *Data Governance* e Segurança, Metadados e Gestão do Ciclo de Vida da Informação (Singh & Ahmad, 2019).

O *Data Governance* define como é que devem ser tomadas as decisões sobre os dados, bem como os respetivos comportamentos das pessoas e dos processos. Por outras palavras, elucida quem é o responsável por cada dado, ou seja, o *data owner*. O principal objetivo do *Data Governance* é garantir a melhor gestão dos dados, de acordo com as melhores práticas e políticas, visto que todas as organizações são obrigadas a tomar decisões sobre os seus dados (Dama International, 2017; Khatri & V. Brown, 2010; Madera & Laurent, 2016). Além disso, são definidas as regras necessárias para a análise de dados, por exemplo, através da definição de príncipios de *data governance*, onde se clarifica o objetivo de cada dado como um ativo, elaborando um dicionário de dados (Khatri & V. Brown, 2010; Madera & Laurent, 2016). Assim, o *Data Governance* é crucial na orientação de todas as demais funções de gestão de dados (Dama International, 2017).

Neste sentido, a segurança dos dados está fortemente relacionada com o *data governance*, pois existem decisões cruciais perante os requisitos de segurança, privacidade e confidencialidade. Além disso, para evitar a perda do controlo dos dados, devem ser identificadas as ameaças ao longo do ciclo de vida dos dados, com o objetivo de desenvolver mecanismos de segurança contra as mesmas (Alshboul et al., 2015; Fang, 2015; Moges et al., 2013).

Quanto ao ciclo de vida do dado, e tendo em conta a elevada quantidade de dados no repositório, é necessário garantir que todos os dados têm um caminho a seguir, ou seja, que existem processos definidos para criar, obter, mover, transformar, armazenar, partilhar, aplicar, eliminar, aprimorar e/ou agregar os dados. Os dados não são estáticos e o ciclo de vida pode ter várias iterações, dificultando a compreensão e especificidade do ciclo de vida do dado por parte da organização. As fases abrangidas pelo ciclo de vida dependem da abordagem seguida pela organização. El Arass e Souissi (2018)

compararam 6 ciclos de vida[1] e, em todos eles, as fases mencionadas são diferentes, variando entre 5 e 14 fases. No entanto, segundo o livro Dama International (2017), a criação e o uso dos dados são os pontos mais críticos num ciclo de vida, visto que a produção de dados tem custos e estes só acrescentam valor quando são consumidos. Em acréscimo, existem fases que devem acompanhar todo o ciclo de vida, por exemplo, a gestão da qualidade de dados, a gestão da qualidade de metadados e a segurança dos dados. Além do mais, devem existir controlos adequados a todas as fases do ciclo de vida dos dados, desde a sua correta inserção no *front office* até à sua atualização (Basel Committee on Banking Supervision, 2013; El Arass & Souissi, 2018; Khatri & V. Brown, 2010).

De outra maneira, os metadados são as características de um dado, por exemplo, um conjunto de *tags* que ajudam a compreender o conteúdo dos dados, tornando a pesquisa mais eficiente (Basel Committee on Banking Supervision, 2013; Nogueira et al., 2018). Sem quaisquer metadados ou sem a correta gestão destes, o *Data Lake* pode transformar-se num *Data Swamp*, onde a estrutura e a semântica dos dados não são conhecidas (Hai et al., 2016). Por sua vez, a gestão dos metadados é crucial para consultas, qualidade e integração dos dados e, por este motivo, foram desenvolvidos diversos sistemas para facilitar a gestão inteligente dos metadados, entre eles, o *Constance* (Mehmood et al., 2019; Nogueira et al., 2018). Assim, o processo de integração dos dados pode ser bastante desafiador para a organização, impondo requisitos à gestão de metadados (Cai & Zhu, 2020; Mehmood et al., 2019; Moges et al., 2013).

Em suma, o projeto do *Data Lake* deve ser capaz de desenvolver um modelo de dados unificado que resolve problemas específicos do negócio, ao mesmo tempo que fornece *insights* para uma tomada de decisão mais informada. Este projeto deve estar alinhado com a estratégia da organização e com as suas necessidades de negócio, de processos e de sistemas (Singh & Ahmad, 2019).

[1] Os diversos ciclos de vida são: *Smart Data Lifecycle, Hindawi Lifecycle, Information Lifecycle, USGS Lifecycle, Big Data Lifecycle, IBM Lifecycle*, sendo que o primeiro foi apresentado por El Arass & Souissi (2018) e os restantes foram os primeiros classificados no artigo (Arass et al., 2017).

2.2.3. Vantagens

Após a implementação do *Data Lake*, muitas são as vantagens das quais a organização pode beneficiar. De uma forma resumida e complementar, enumeram-se algumas das mesmas:

- O armazenamento do *Data Lake* é escalável e tem baixo custo, potencializando o crescimento contínuo dos dados, sem comprometer a qualidade e disponibilidade dos mesmos (Gupta & Giri, 2018; Mehmood et al., 2019; Moges et al., 2013; Singh & Ahmad, 2019).
- O *Data Lake* melhora a eficiência operacional ao diminuir o tempo na preparação e análise pormenorizada dos dados, enquanto fornece *insights* e padrões de tendência nos dados (Gupta & Giri, 2018).
- O facto de ser um repositório centralizado com vários tipos de dados (Fang, 2015), reduz a dependência às restantes fontes de informação, facilita a integração dos dados e reduz os silos de informação (Mehmood et al., 2019; Moges et al., 2013; Stein & Morrison, 2014).
- O *Data Lake* é controlado por regras, ferramentas e processos, garantindo o governo de dados, bem como a sua qualidade. A documentação é acessível e pode ser refinada colaborativamente (Madera & Laurent, 2016; Stein & Morrison, 2014).

Desta forma, realça-se a importância de analisar quais as vantagens inerentes à implementação do *Data Lake*, por parte da organização em estudo, identificando-se como uma dimensão da análise qualitativa em estudo.

2.2.4. Desafios

O principal desafio do *Data Lake* é aproveitar todas as oportunidades que este proporciona (Singh & Ahmad, 2019). Por este motivo, apresentam-se, resumidamente, vários desafios na implementação e/ou manutenção do mesmo:

- O processo de implementação do *Data Lake* requer muito esforço técnico (Fang, 2015). Além disso, as ferramentas e interfaces de dados podem não funcionar tão bem num repositório único de dados (Mehmood et al., 2019).
- Encontrar os dados no *Data Lake* pode tornar-se complicado, caso este se transforme num *Data Swamp* (Nargesian et al., 2019). Por esse motivo, a falta de

metadados torna o processo de integração dos dados mais desafiante (Mehmood et al., 2019). Ademais, os termos ambíguos devem ser definidos no dicionário de dados, através do entendimento entre os interessados, convergindo para uma uniformização de conceitos de negócio (Stein & Morrison, 2014). Exemplo disso, é a ambiguidade existente no *ownership* dos dados, que dificulta a determinação do *owner* de um dado (Dama International, 2017). Só assim é que é possível desenvolver interfaces com fontes de dados externas (Nargesian et al., 2019). Em acréscimo, as fontes de dados podem ficar obsoletas ou podem surgir outras novas, imputando dificuldade na manutenção da documentação atualizada (Mehmood et al., 2019).

- Tendo em conta o *schema* do *Data Lake*, a maioria dos utilizadores deste repositório carece de habilidades mais técnicas, para transformar e analisar os dados com as novas ferramentas. Muitas vezes, o desenvolvimento destas novas capacidades tem um grande custo, não só financeiro, como temporal (Fang, 2015).

Neste sentido, o presente estudo pretende estudar quais os desafios sentidos pela organização, identificando-se como uma dimensão de estudo.

3. METODOLOGIA

3.1. Case Study

O *case study* foi a metodologia escolhida para desenvolver a presente investigação. Além de ser uma das metodologias mais utilizadas em investigação de sistemas de informação (Martins & Belfo, 2010), um dos objetivos do estudo é compreender o impacto gerado pela implementação do *Data Lake*. Assim, pretende-se obter uma visão holística dos processos, sendo um motivo para a escolha da metodologia (Gerring, 2004). O estudo do impacto da adoção de uma nova ferramenta é bastante complexo, envolvendo o uso de tecnologia no contexto da organização, sendo também um argumento a favor da metodologia escolhida (Martins & Belfo, 2010). Por outras palavras, o *case study* é uma investigação no seu ambiente real e permite identificar o que está a acontecer, bem como a sua razão, compreendendo os seus efeitos e implicações (Mark Saunders et al., 2016).

Neste seguimento, a investigação tem um objetivo exploratório, visto que é o mais indicado para compreender o que está a acontecer e gerar *insights* relevantes sobre várias temáticas (Mark Saunders et al., 2016). No que diz respeito à abordagem seguida, esta será preferencialmente dedutiva, dado que a literatura existente é o ponto de partida para a definição das questões de investigação, bem como dos objetivos, e para a organização da análise de dados (Mark Saunders et al., 2016). No entanto, também são utilizados elementos de uma abordagem indutiva na análise dos dados, permitindo o aparecimento de novos códigos, bem como o seu ajustamento (Mark Saunders et al., 2016). Tal acontece quando a revisão de literatura não permite explorar, de forma adequada, as respostas concedidas pelos participantes (Mark Saunders et al., 2016). O *case study* é aplicado a um caso real e específico, sendo este projeto crítico para a conformidade com o regulador do setor bancário, o que leva a que seja um *case study* único (Mark Saunders et al., 2016; Martins & Belfo, 2010). Por se tratar de um estudo num momento pré-estabelecido, em que a recolha de dados será durante um curto período de tempo, o horizonte temporal é *cross-sectional* (Mark Saunders et al., 2016).

Por último, a unidade de análise é a organização, mais concretamente, um banco que atua no mercado português. A organização é uma referência no setor financeiro português, que contribui para a estabilidade do sistema financeiro. O Banco de Portugal (2020) identificou a instituição como outra instituição de importância sistémica (O-SII),

devido à sua dimensão, importância para a economia, complexidade e potencial impacto, no caso de insolvência, para o sistema financeiro e para a economia em geral. É importante mencionar que não foram identificadas instituições de importância sistémica global (G-SII) em Portugal (Banco de Portugal, 2020). A organização, fundada antes do século XXI, apresenta um conjunto de sistemas de informação que foram adquiridos e adaptados ao longo dos anos. Desta forma, a instituição está presente em todas as áreas do setor bancário e distingue-se pela sua cobertura geográfica, através de uma vasta rede de distribuição internacional. Os objetivos da mesma passam por inovar e melhorar continuamente a prestação de serviços e acompanhar os desenvolvimentos tecnológicos no setor, ao promover o uso de novas tecnologias, tanto pelos clientes, como dos colaboradores, reduzindo os custos operacionais e garantindo a competitividade num mercado financeiro dinâmico. Em 2020, o número total de colaboradores situou-se no intervalo entre 4500 e 7000, dos quais cerca de 300 pertencem à área de sistemas de informação (área técnica em estudo) e cerca de 100 efetuam *reporting* (área de negócio em estudo). Toda a informação descrita anteriormente foi retirada de documentação interna e, com o objetivo de manter a privacidade e sigilo dos dados, não será mencionado o nome da organização. Assim, trata-se de um *case study* holístico, que tem como premissa estudar o impacto do *Data Lake* na organização como um todo (Mark Saunders et al., 2016).

3.2. Recolha de Dados

Segundo Martins e Belfo (2010), a recolha de dados deve ser efetuada de diversas fontes, salientando a documentação e as entrevistas como fontes principais. Em particular, as entrevistas semiestruturadas podem ser usadas para compreender motivos das decisões, das atitudes, bem como das opiniões, dos participantes (Mark Saunders et al., 2016). Por esta a razão, a recolha de dados foi efetuada através de entrevistas semiestruturadas e documentação interna. Através da documentação interna, foi possível caracterizar a organização e compreender o projeto em curso, permitindo a triangulação dos dados recolhidos (Mark Saunders et al., 2016).

Nas entrevistas semiestruturadas, o investigador tem uma lista de questões-chave com o objetivo de recolher dados válidos, confiáveis e relevantes para as questões e objetivos da investigação (Mark Saunders et al., 2016). Por conseguinte, as questões das entrevistas

foram formuladas e estruturadas de modo a dar resposta aos objetivos e, consequentemente, às questões de investigação, por meio dos guiões presentes no Anexo A. No entanto, a ordem das questões pode ser alterada consoante o fluxo da conversa, de igual modo que é possível adicionar e/ou omitir questões, dado o contexto específico (Mark Saunders et al., 2016).

As entrevistas focaram-se em duas áreas distintas da organização: a área técnica, ou seja, a área responsável pela implementação do *Data Lake*; e, a área de negócio, isto é, a área de *reporting* dos dados, a qual vai usufruir do novo contexto. Por este motivo, o processo de seleção do painel de peritos teve em conta as duas áreas distintas. Foram escolhidos três participantes da área técnica, por recomendação, e dois participantes da área de negócio. Num total de cinco entrevistas, apresenta-se, abaixo, na Tabela 3.1, a caracterização sociodemográfica dos entrevistados.

Tabela 3.1

Caracterização Sociodemográfica dos Entrevistados

Entrevista	Tipo	Idade	Grau de Habilitação Literária	Área da Habilitação Literária	Função	Tempo na Função	Tempo na Empresa
E1_N_CRE	Negócio	43	Licenciatura	Economia	Coordenador (Área de Reporte)	4 anos	19 anos
E2_N_DRE	Negócio	50	Licenciatura	Economia	Diretor (*Reporting*)	5 anos	27 anos
E3_T_CAO	Técnica	39	MBA	Informação de Gestão	*Chief Analytics Officer*	1 ano	1 ano
E4_T_CDO	Técnica	47	Mestrado	Engenharia	*Chief Data Officer*	2 anos	1 ano
E5_T_CRI	Técnica	38	Licenciatura	Finanças	Coordenador de Risco	2 anos	6 anos

Fonte: Elaborado pelo próprio autor (2021).

As entrevistas foram realizadas na plataforma *ZOOM*, devido à situação pandémica imposta pelo novo coronavírus SARS-COV-2, e tiveram uma duração média de 43 minutos. Todas as conversas decorreram num ambiente tranquilo, sem ruído, sem dificuldades acrescidas, com uma ou nenhuma falha de rede, e, nesses casos, houve repetição da informação perdida. Além disso, todos os participantes demonstraram total disponibilidade para esclarecimento de novas questões.

À exceção da quinta entrevista (E5_T_CRI), todas as restantes foram gravadas, com o consentimento dos participantes, com o único objetivo de serem transcritas posteriormente. Para facilitar a transcrição do mencionado pelos participantes, utilizou-se um *software*[2] com tecnologia *Speech-to-Text*. Assim, ao reproduzir a gravação da entrevista, o que foi referido é convertido para texto, em tempo real (SpeechTexter, 2014). Segundo o *site* SpeechTexter (2014), os níveis de precisão desta conversão são acima de 90%. Neste caso, a tecnologia simplificou notoriamente este processo, mas existiram dificuldades, como a colocação dos sinais de pontuação e o reconhecimento de frases proferidas mais rapidamente.

No seguimento, efetuou-se a limpeza dos dados, onde se certificou de que a transcrição estava correta, corrigindo eventuais erros (Mark Saunders et al., 2016). Para além da transcrição do mencionado pelos participantes, também foram tiradas notas, principalmente às expressões não verbais dos mesmos, sabendo a importância das mesmas (Mark Saunders et al., 2016). No caso da quinta entrevista (E5_T_CRI), as notas foram mais detalhadas e o entrevistado pausava entre respostas para facilitar a escrita de frases e palavras-chave. Posteriormente, os documentos para análise, tanto as transcrições, como as notas do Entrevistado E5_T_CRI, foram enviados para validação pelos próprios participantes, de forma a garantir que todas as ideias foram bem interpretadas. Apenas 3 respostas foram recebidas, mas importa realçar que as notas do entrevistado E5_T_CRI foram uma destas, detalhando mais cada resposta dada.

Por último, é importante realçar que todos os participantes estavam conhecedores do objetivo da entrevista, bem como do presente estudo, do cumprimento da privacidade e proteção dos dados pessoais e da possibilidade de interromper a entrevista a qualquer momento, caso o mesmo o desejasse.

3.3. Análise de Dados

Os dados foram analisados com recurso à abordagem de Análise Temática, da qual o principal objetivo é a identificação de temas e padrões relacionados com as questões de investigação, através da codificação dos dados qualitativos (Mark Saunders et al., 2016). Esta abordagem é flexível e acessível, não estando associada a nenhuma filosofia

[2] https://www.speechtexter.com/

específica, e permite compreender grandes quantidades de dados qualitativos, integrar dados de diferentes transcrições e notas, identificar temas-chave, bem como descrever e desenvolver relações aparentes de temáticas (Mark Saunders et al., 2016). O *software* usado para a análise de dados foi o *MAXQDA*.

Numa primeira fase, é necessário fortalecer o envolvimento com os dados, através de um processo de imersão nos dados, demonstrando a importância da transcrição na análise de dados (Mark Saunders et al., 2016).

Em seguida, os dados com significados semelhantes devem ser categorizados com um código que simboliza o significado daquele segmento (Mark Saunders et al., 2016). Para manter a coerência e consistência ao longo do processo de codificação, foi elaborada uma lista de códigos com a respetiva definição de trabalho (Mark Saunders et al., 2016), a qual se encontra no Anexo B. Com o objetivo de compreender o grau de conformidade da organização, foram utilizados os critérios de avaliação publicados pelo Banco de Portugal (2019), também no Anexo B. Além disso, foi efetuada uma comparação constante das entrevistas, de forma a atestar que o aparecimento de novos códigos, como também a manifestação de novos *insights* dos códigos existentes, foram tidos em conta na codificação e/ou recodificação dos dados, garantindo a consistência na análise dos dados (Mark Saunders et al., 2016).

Por último, procurou-se extrair significados dos dados, agrupando-os em temas relevantes. Deste modo, avaliou-se a importância de cada tema e, posteriormente, aprimoraram-se as codificações, eliminando códigos desnecessários ao estudo e introduzindo outros novos (Mark Saunders et al., 2016).

4. APRESENTAÇÃO DE RESULTADOS

Após a análise dos dados assimilados, são apresentadas as opiniões e as visões dos entrevistados com recurso a tabelas, figuras e *verbatim* dos mesmos. As frequências dos códigos e subcódigos, perfazendo um total de 607 codificações, encontram-se no Anexo C e serão analisadas em seguida.

4.1. Reporting

Com o objetivo de analisar a situação atual do *reporting* e de identificar as principais dificuldades sentidas no processo atual, os entrevistados da área de negócio foram abordados com 7 questões cujo foco é a qualidade dos dados, a flexibilidade e documentação dos procedimentos e o conhecimento dos objetivos e finalidades do reporte pedido pelo regulador. Neste sentido, foram identificados diversos desafios no procedimento atual do *reporting*, codificando um total de 143 segmentos.

Figura 4.1

Nuvem do Código: Reporting - Desafios Atuais

Fonte: Elaborado pelo próprio autor (2021).

Em primeiro lugar, a existência de várias fontes de dados foi o desafio mais referido ao longo das entrevistas por 4 participantes (80%). Em segundo lugar, 3 entrevistados (60%) mencionaram a falta de recursos, mais concretamente, a falta ferramentas apropriadas, recursos humanos e de processos de controlo de qualidade. Em seguida, surge o acesso à informação como um dos desafios mais referidos, reconhecendo a dificuldade em extrair informação dos sistemas operacionais. Além disso, à exceção do código "*inputs* manuais", todos os demais foram identificados pelos dois entrevistados da área de negócio (E1_N_CRE e E2_N_DRE).

Tabela 4.1

Comentários Verbatim: Data Lake - Desafios Atuais

Entrevistado	Citação Exemplo
Conhecimento Geral do Reporte	
E2_N_DRE	"Não sei se isto é problema transversal a várias empresas ou não, admito que sim, o foco no meio e não no objetivo final é recorrente, a maior parte de nós foca-se no que controla, no entregável."

Recursos Inadequados	
E1_N_CRE	"Não é fácil fazer controlos de milhares de registos, não tendo ferramentas apropriadas para o fazer"
Inputs Manuais	
E2_N_DRE	"Muita da informação que nós tratamos, (…) depende de uma coisa que se chama *inputs* manuais, ou seja, carregamentos manuais. E, tudo o que é manual, obviamente, não é só, mas sobretudo, está sujeito a risco operacional."
Múltiplas Fontes de Dados	
E2_N_DRE	"Os sistemas não estão integrados e nós para os diversos reportes nos vários órgãos de estrutura do banco, muitas vezes, inclusivamente, usamos extrações diferentes das mesmas realidades."
Documentação	
E5_T_CRI	"Inexistência de um dicionário de dados comum."
Acesso à Informação	
E2_N_DRE	"Se houver algum tipo de questão/dúvida sobre os dados, temos que andar um bocadinho mais para trás e temos que pedir intervenção da própria IT, para nos ajudar a fazer o *drill down* dessa informação."
Peso Regulatório	
E1_N_CRE	"Até hoje, nós temos conseguido sempre dar resposta a tudo. Obviamente que temos tido algumas dificuldades neste últimos 2 anos / 3 anos, por força do aumento da confrontação do Banco de Portugal."

Fonte: Elaborado pelo próprio autor (2021).

4.2. Data Lake

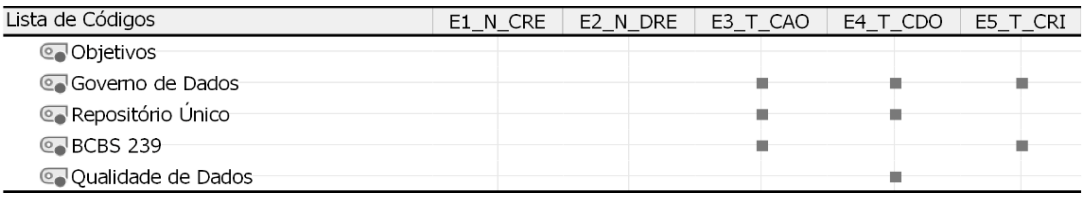
No que toca ao *Data Lake*, procurou-se compreender qual a visão da organização perante a tecnologia. O objetivo do estudo desta dimensão é identificar os objetivos, os *drivers* e as expetativas da implementação do *Data Lake*, como as suas vantagens e os desafios causados pela mesma.

4.2.1. Objetivos e Drivers

Quanto aos objetivos, todos os entrevistados da área técnica (E3_T_CAO, E4_T_CDO e E5_T_CRI) identificaram o melhor conhecimento dos dados como um dos objetivos principais, sendo o "Governo de Dados" o subcódigo com mais frequência de segmentos. Por outro lado, 2/3 dos entrevistados da mesma área (67%) fazem referência ao cumprimento do BCBS 239 e à integração dos dados num repositório único como objetivos da implementação do *Data Lake*.

Figura 4.2

Visualização Binária do Código: Data Lake - Objetivos



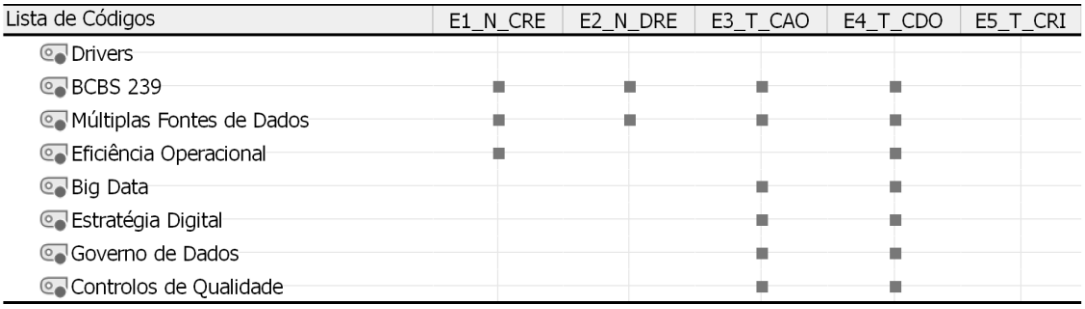
Fonte: Elaborado pelo próprio autor (2021).

No entanto, no decorrer das entrevistas, os participantes foram identificando alguns *drivers* à implementação do *Data Lake*, tendo em conta a sua relação causal com os objetivos, o que levou à criação do código *a posteriori*.

Em concordância com os objetivos identificados, 80% dos entrevistados reconheceram o "BCBS 239" e os "Múltiplos Repositórios" como os principais *drivers* do projeto. Destaca-se, ainda, o código "*Big Data*", que menciona o reforço informático necessário devido ao elevado volume de dados.

Figura 4.3

Visualização Binária do Código: Data Lake - Drivers



Fonte: Elaborado pelo próprio autor (2021).

4.2.2. Vantagens

Quanto às vantagens, todos os participantes contribuíram para a codificação de segmentos. O subcódigo "Eficiência Operacional" foi o mais mencionado ao contrário do subcódigo "Características da Tecnologia", que é alusivo às próprias características da tecnologia, como ser escalável e redundante.

Figura 4.4

Nuvem do Código: Data Lake – Vantagens

Fonte: Elaborado pelo próprio autor (2021).

Apesar de ser o mais referido, 66% das menções do subcódigo "Eficiência Operacional" foram feitas pelo entrevistado E3_T_CAO. O único subcódigo que foi incluído por todos os entrevistados foi o "Controlos de Qualidade", realçando a importância da coerência de informação. Não obstante, à exceção do subcódigo

"Características da Tecnologia", os restantes códigos foram abordados por 4/5 dos entrevistados (80%).

Tabela 4.2

Comentários Verbatim: Data Lake – Vantagens

Entrevistado	Citação Exemplo
Características da Tecnologia	
E3_T_CAO	"Tecnologia alinhada às melhores práticas, escalável e que tenha redundância, tenha *back-ups*, que tenha *disaster recovery*, etc.., é muito importante."
Eficiência Operacional	
E1_N_CRE	"Diminui significativamente o risco operacional de reporte, em termos de análises, também permite ir logo diretamente à operação."
Documentação	
E3_T_CAO	"Teres documentação com qualidade sobre esse elemento de dado, seja de um ponto de vista de definição funcional, seja no ponto de vista de definição técnica, seja no ponto de vista de rastreabilidade."
Repositório Único	
E2_N_DRE	"Passamos a ter a informação das empresas do grupo de uma forma também *standardizada*, também no mesmo tempo, e permite, de alguma forma, também aproveitar algumas sinergias porque, no fundo, não há depois manutenção de extrações dos sistemas fonte para a casa mãe."
Controlos de Qualidade	
E5_T_CRI	"Identificação dos controlos de qualidade por todos, a implementação de *dashboard* é critica para implementar medidas corretivas."

Fonte: Elaborado pelo próprio autor (2021).

4.2.3. Desafios

Os desafios associados à implementação do *Data Lake* foram identificados por 4 dos participantes (80%), no entanto são os entrevistados E4_T_CDO e E2_N_DRE (40%), respetivamente, que mais apoiam esta dimensão.

Figura 4.5

Frequência do Código: Data Lake – Desafios

Lista de Códigos	E1_N_CRE	E2_N_DRE	E3_T_CAO	E4_T_CDO	E5_T_CRI	SOMA
Desafios						0
Data Ownership		2		16		18
Uniformização de Conceitos		8				8
Visão de Grupo				7		7
Processo Complexo		3		2		5
Tamanho da Organização	1	4				5
Novas Skills		2	2			4
∑ SOMA	1	19	2	25	0	47

Fonte: Elaborado pelo próprio autor (2021).

O subcódigo "*Data Ownership*" foi a codificação com mais segmentos, no entanto 89% pertence ao mesmo participante (E4_T_CDO). Em seguida, surge a "Uniformização de Conceitos" e a "Visão de Grupo" como desafios, mas, em ambos os casos, apenas existe 1 entrevistado (20%) a identificar estas dificuldades. Por último, aparecem os

subcódigos com menos frequência e com maior distribuição entre as entrevistas, são eles "Processo Complexo", "Tamanho da Organização" e "*Novas Skills*".

Tabela 4.3

Comentários Verbatim: Data Lake - Desafios

Entrevistado	Citação Exemplo
Processo Complexo	
E4_T_CDO	"Porque obriga não só a criar aquela camada do governo e qualidade de dados, mas também a outra camada tecnológica, que é refazer na realidade todo o *layer* informacional de um banco, são programas que levam anos."
Tamanho da organização	
E1_N_CRE	"Obviamente que isto era muito fácil fazer numa entidade com balanço pequeno, numa entidade como a nossa com balanço grande, obviamente que complica sobremaneira a tarefa."
Visão de Grupo	
E4_T_CDO	"A necessidade de isto ser feito a nível do grupo, a conformidade tem de ser garantida a nível de grupo, então depois há modelos de relação com as entidades do grupo, (…), mas depois a distância geográfica, a autonomia de acionistas, *etc*., levanta aqui uma série de questões e de temas sobre o que é que deve ser delegado ou o que é que não deve ser delegado."
Uniformização de Conceitos	
E2_N_DRE	"No fundo, nós temos que alinhar pela mesma bitola, ou seja, definir os conceitos e perceber efetivamente quais é que são comuns, quais é que não são comuns, os que são comuns, defini-los de uma forma inequívoca, e os que não são comuns, temos que acrescentá-los ao nosso dicionário de dados e perceber como é que vamos orientá-los."
Data Ownership	
E4_T_CDO	"O assumir o *ownership* de um dado, que traz todas estas responsabilidades, não é de ânimo leve, portanto é uma discussão, (…), acho que a discussão que se coloca neste momento é: "porque é que sou eu o *owner* do dado?""
Novas Skills	
E3_T_CAO	"Ok, isso é muito bonito, mas de que serve dares um carro uma pessoa, se a pessoa não sabe conduzir?"

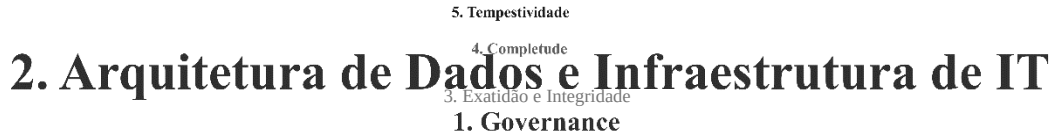
Fonte: Elaborado pelo próprio autor (2021).

4.3. BCBS 239

O intuito do estudo desta dimensão é compreender qual é o estado da organização no que diz respeito à conformidade com os princípios do BCBS 239. Ao longo das entrevistas, foram codificados segmentos que relacionam os príncipios com os desafios atuais e/ou com as vantagens da implementação do *Data Lake*, de modo a existir uma comparação entre cenários: antes *vs* depois.

Figura 4.6

Nuvem do Código: BCBS 239 – Princípios



Fonte: Elaborado pelo próprio autor (2021).

Deste modo, é possível verificar que o princípio 2, "Arquitetura de Dados e Infraestrutura de IT" foi o subcódigo com mais segmentos associados. Seguido do princípio 1, "*Governance*", e do princípio 3, "Exatidão e Integridade", respetivamente.

4.3.1. Princípios

4.3.1.1. Governance e Infraestrutura

A área de *Governance* e Infraestrutura abrange os dois primeiros princípios e, com um total de 157 codificações, é a área dos princípios com mais referências entre os entrevistados (68%). Ambos os princípios foram mencionados por todos os entrevistados (100%), tornando-se os princípios com maior frequência de segmentos, 57 e 100, respetivamente.

Tabela 4.4

Comentários Verbatim: BCBS 239 - Governance e Infraestrutura

Entrevistado	Citação Exemplo
1. Governance	
E3_T_CAO	"Teres os dados governados e com a qualidade que é merecida para uma organização."
2. Arquitetura de Dados e Infraestrutura de TI	
E5_T_CRI	"Criar uma cultura de dados: identificar *ownership*, identificação única dos atributos dos dados, um processo de controlo centralizado no *Data Lake*."

Fonte: Elaborado pelo próprio autor (2021).

Através do subcódigo do princípio "*Governance"*, os participantes salientaram a importância da documentação, enquanto o subcódigo do princípio da "Arquitetura de Dados e Infraestrutura de IT" foca a *golden source* e a uniformização de conceitos em todo o grupo bancário.

4.3.1.2. Integração de Dados

A área de Integração de Dados foi potencializada pelo terceiro princípio, "Exatidão e Integridade", sendo este subcódigo o terceiro maior a nível de número de segmentos. Além disso, foi o único princípio citado por todos os entrevistados (100%) nesta área.

Tabela 4.5

Comentários Verbatim: BCBS 239 - Integração de Dados

Entrevistado	Citação Exemplo
3. Exatidão e Integridade	
E5_T_CRI	"Rastreabilidade dos dados até aos sistemas operacionais."
4. Completude	
E5_T_CRI	"Integração de processos no *Data Lake*."
5. Tempestividade	
E3_T_CAO	"Temos compromissos regulatórios diariamente, semanalmente, mensalmente. Eu acho que isso é o dia-a-dia do banco e nós temos que ter essa visão e, portanto, eu acho que o *Data Lake* vai conseguir responder a isso."
6. Adaptabilidade	
E2_N_DRE	"Tentar reutilizar informação que recebemos o mais possível."

Fonte: Elaborado pelo próprio autor (2021).

4.3.1.3. Práticas de Reporting

A área dos princípios relacionados com as Práticas de *Reporting* foi responsável por codificar 6% dos segmentos alusivos a todos os princípios do BCBS 239, sendo uma área focada na validação de consistência entre reportes. Nenhum dos subcódigos foi referido por todos os participantes, inclusive o princípio da "Abrangência" não foi mencionado ao longo de toda a recolha de dados.

Tabela 4.6

Comentários Verbatim: BCBS 239 - Práticas de Reporting

Entrevistado	Citação Exemplo
7. Exatidão	
E2_N_DRE	"Muitas vezes, gastamos algum tempo, para não dizer bastante, a justificar porque é que o que reportamos nas Estatísticas Monetárias e Financeiras não está, não é o que reportamos para a Central de Responsabilidades de Crédito."
9. Clareza e Utilidade	
E3_T_CAO	"É tu garantires que o dado é relevante, eu quando apresento uma coisa a ti, Mariana, tem que ser algo para ti tenho um sentido, (…), e tu consigas utilizar isso no processo de tomada de decisão."
10. Frequência	
E1_N_CRE	"Até hoje, nós temos conseguido sempre dar resposta a tudo."
11. Distribuição	
E2_N_DRE	"Genericamente tem-se conseguido atingir os objetivos a que nos propomos."

Fonte: Elaborado pelo próprio autor (2021).

Destacam-se os princípios “Exatidão” e “Clareza e Utilidade” como os subcódigos com maior frequência. Por um lado, a “Exatidão” procura garantir a rastreabilidade e coerência entre reportes, por outro, a “Clareza e Utilidade” tem o objetivo de garantir a relevância de cada reporte.

4.3.1.4. Revisão Regulatória

A área da Revisão Regulatória é vocacionada para os reguladores e supervisores, no entanto o quarto entrevistado (E4_T_CDO) abordou o processo de avaliação de conformidade com os princípios, demonstrando a complexidade do processo.

Tabela 4.7

Comentários Verbatim: BCBS 239 - Revisão Regulatória

Entrevistado	Citação Exemplo
12. Revisão	
E4_T_CDO	"Recordo-me de relatórios que ao fim de cinco/seis anos, o número de bancos G-SII na Europa que estavam em conformidade com o BCBS 239 eram um, dois, e, mais do que isso."

Fonte: Elaborado pelo próprio autor (2021).

4.3.2. Desafios

Os entrevistados também mencionaram os principais desafios que enfrentam ao desenvolverem projetos para agirem em conformidade com o BCBS 239, motivando a criação de um novo subcódigo.

Figura 4.7

Nuvem do Código: BCBS 239 – Desafios

Fonte: Elaborado pelo próprio autor (2021).

O subcódigo “Processo Complexo” foi o desafio mais frequente entre os três entrevistados (60%) que o mencionaram (E2_N_DRE, E3_T_CAO, e E4_T_CDO), defendendo que os projetos de conformidade “são programas que levam anos” (E4_T_CDO). Aliado a isso, “os princípios são abstratos” (E3_T_CAO) e é necessário compreender se os atributos estão a ser definidos “em conformidade com aquilo que é pedido” (E1_N_CRE), caracterizando o subcódigo “Correta Interpretação”. Por último, o terceiro entrevistado (E3_T_CAO) reforça que “tem que se garantir que continua a acontecer” a conformidade com o BCBS 239.

4.3.3. Oportunidades Futuras

Quando abordados a respeito do BCBS 239, 4 dos entrevistados (80%) identificaram projetos futuros e consequências do cumprimento do príncipios, fomentando um novo código de análise.

Figura 4.8

Nuvem do Código: BCBS 239 – Oportunidades Futuras

Fonte: Elaborado pelo próprio autor (2021).

Ambos os entrevistados da área de negócio (E1_N_CRE e E2_N_DRE) reportaram que o futuro dos reportes passa pelo detalhe de operação a operação e não pela agregação dos dados. Por outro lado, os entrevistados da área técnica (E3_T_CAO e E4_T_CDO) apontam que, no futuro, a exploração dos dados no *Data Lake* será numa "lógica de *self-service*" (E3_T_CAO), ao mesmo tempo que permite uma "visão 360 de tudo e do cliente" (E3_T_CAO), fomentando o interesse do cliente nos produtos e serviços da organização (E4_T_CDO).

Tabela 4.8

Comentários Verbatim: BCBS 239 - Oportunidades Futuras

Entrevistado	Citação Exemplo
Estruturas dos Futuros Reportes	
E2_N_DRE	"Eu acredito que num futuro, não estou a dizer que seja num ano, nem dois, nem três, mas acredito que num futuro não muito longínquo, cada vez mais os reguladores não nos vão pedir agregações, mas vão-nos pedir os detalhes de todas as operações."
Exploração de dados	
E4_T_CDO	"Nós temos a ambição de tornar isto também muito numa lógica *self-service* dos próprios dados."
Experiência ao Cliente	
E3_T_CAO	"E, nós, com base nos dados que vão estar no *Data Lake*, se conseguirmos desenvolver algoritmia que consiga responder a isto, vai estar num nível de maturidade, que a <organização> ainda não está, nunca esteve e tem a oportunidade de estar, e, aí sim, providenciares um serviço à séria ao teu cliente e teres o cliente no centro de tudo o que nós fazemos."

Fonte: Elaborado pelo próprio autor (2021).

5. DISCUSSÃO DE RESULTADOS

A discussão de resultados tem como principal objetivo compreender o impacto da implementação do *Data Lake* na concretização dos princípios do BCBS 239. Para isso, relacionaram-se os temas abordados na revisão de literatura, *Reporting* e *Data Lake*, com o(s) princípio(s) que melhor caracteriza(m) cada conceito. Além disso, identificaram-se oportunidades futuras e desafios das ações em prol da conformidade com o BCBS 239, por parte da organização.

5.1. Reporting

Perante a situação atual dos processos de *reporting*, a generalidade dos entrevistados identifica a multiplicidade de fontes como uma dificuldade a enfrentar, tal como defendem Cai e Zhu (2020), Chaudhuri e Dayal (1997), Kadadi et al. (2014). Como consequência, existe "sempre o problema da consistência entre a informação entre as várias áreas" (E4_T_CDO), inclusivamente, são usadas "extrações diferentes das mesmas realidades" (E2_N_DRE), demonstrando que a organização não está em conformidade com o princípio 2, "Arquitetura de Dados e Infraestrutura de TI".

Em concordância com os autores Stein e Morrison (2014), os entrevistados consideram que a falta de recursos é um desafio na qualidade dos dados. Em específico, o entrevistado E2_N_DRE afirma que a estrutura "não assegura a uniformidade dos controlos de qualidade da informação". Neste sentido, realça-se, mais uma vez, que o princípio 2 deve ser um ponto a ser melhorado.

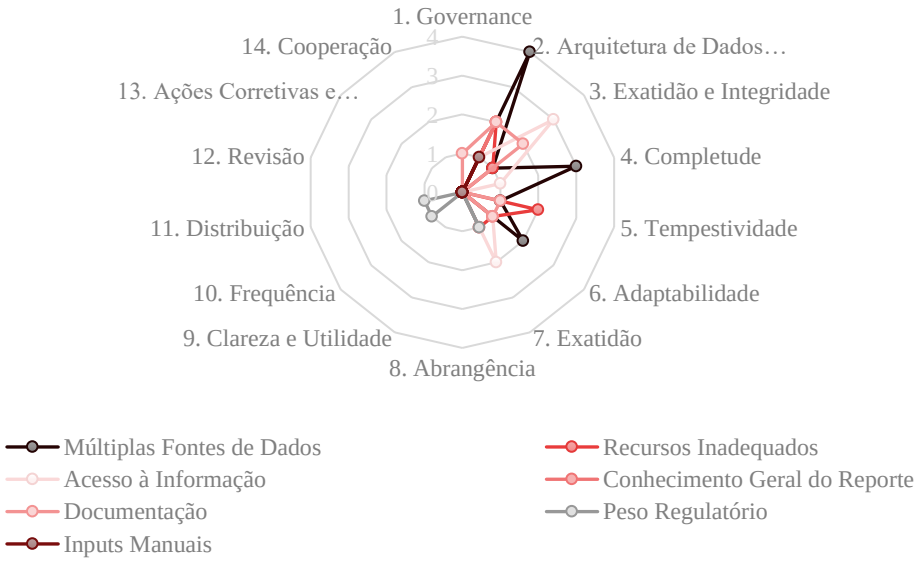
Os peritos identificam que, por falta de capacidade "para fazer o *drill down*" da informação (E2_N_DRE), não é possível compreender o detalhe de uma operação de forma ágil. Tal acontece porque existe uma área de SI responsável pela extração dos dados e, por vezes, pode descuidar o detalhe, o que vai ao encontro do defendido por Moges et al. (2013). Sendo assim, é complicado conseguir rastrear os dados, pelo que o princípio 3, "Exatidão", também não está a ser realizado na sua totalidade.

Por último, o conteúdo das entrevistas apoia que a falta de documentação é um desafio constante, sustentando a tese de Moges et al. (2013). Em concordância com Butler (2017), a inexistência de um dicionário de dados também está patente na organização em estudo (E5_T_CRI), o que reforça o incumprimento do princípio 2.

Em suma, os desafios atuais têm consequências diretas na conformidade dos princípios do BCBS 239, pelo que se apresenta graficamente a relação entre ambos.

Figura 5.1

Relação entre os Desafios Atuais e os Princípios do BCBS 239



Fonte: Elaborado pelo próprio autor (2021).

Assim, é possível concluir que os atuais processos de *reporting* são impactados essencialmente pelas múltiplas fontes de dados, agindo em incoerência com o princípio 2. Além disso, o princípio 3 e 4 também identificam incoerências nos procedimentos.

5.2. Data Lake

5.2.1. Objetivos e Drivers

A maioria dos entrevistados acredita que um dos objetivos do *Data Lake* é "promover maior qualidade, governo e, portanto, disponibilização dos dados" (E3_T_CAO), no entanto, na revisão de literatura, não foi identificado nenhum autor que partilhasse a mesma opinião explicitamente. Segundo Stein e Morrison (2014), a tecnologia facilita a acessibilidade e integração dos dados, realçando a importância da existência de governo de dados. Neste caso, este objetivo contribui positivamente para a conformidade com o princípio 1, "*Governance*".

Além disso, Moges et al.(2013), Stein e Morrison (2014) argumentam que o *Data Lake* deve responder a um problema de negócio e, neste caso, esse problema são os

princípios definidos pelo Basel Committee on Banking Supervision (2013), como sublinha o entrevistado E3_T_CAO, "atacar, no bom sentido, o BCBS 239".

Neste sentido, o BCBS 239 não é apenas um objetivo de concretização, como é um *driver* para o projeto. Em sintonia com o sustentado por Moges et al. (2013), todos os entrevistados apontam que a atividade dos reguladores, neste caso, o BCBS 239, é uma das razões que leva a atividades de melhoria da qualidade dos dados.

Por último, e em complemento ao objetivo de armazenamento dos dados num só repositório, defendido por Mehmood et al. (2019), os participantes identificam que o facto de "tentar ter uma única fonte de verdade" (E2_N_DRE) é uma motivação à implementação da tecnologia já mencionada, contribuindo para o aprimoramento da *compliance* da organização com o princípio 2, "Arquitetura de Dados e Infraestrutura de IT".

5.2.2. Vantagens

A respeito dos benefícios da implementação do *Data Lake*, concluiu-se que estes vão ao encontro da literatura descrita anteriormente e, em acréscimo, despertam um novo proveito: "Controlos de Qualidade".

Todos os entrevistados consideram que os controlos de qualidade da informação são uma mais-valia do projeto, contribuindo positivamente para o princípio 2, visto que assegura a existência de controlos em todas as fases do ciclo de vida do dado. Assim, é possível a garantir que quando se diz que "'"o cliente tem 7 produtos ativos", na realidade tem 7 produtos ativos" (E3_T_CAO) e não tem mais ou menos produtos ativos. Em acréscimo, estes controlos de qualidade são uniformes (E2_N_DRE), pelo facto deste repositório ser centralizado e facilitar a integração dos dados (Fang, 2015). Esta realidade também é reconhecida, pelos intervenientes, como uma vantagem, garantindo que o dado só é produzido uma vez e não está em múltiplos repositórios informacionais (E3_T_CAO). Desta forma, este repositório único apoia a integração dos dados e o uso de identificadores únicos, bem como a agregação dos dados numa única base de dados, agindo em conformidade com o princípio 2.

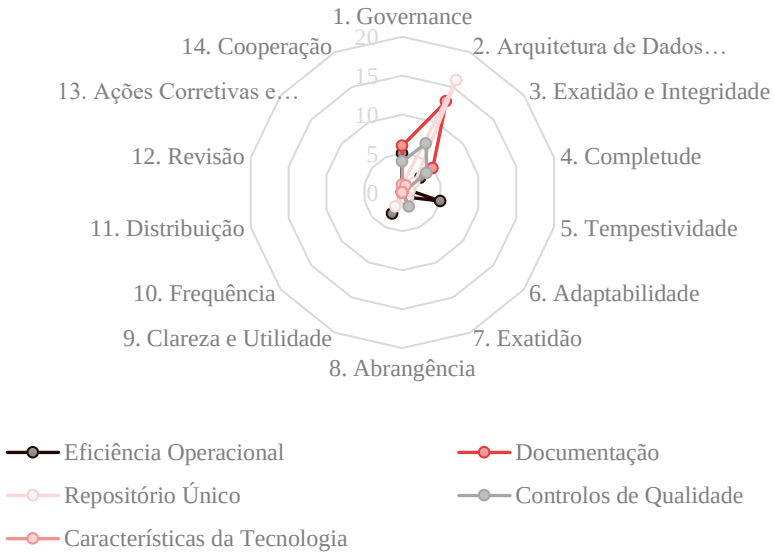
Ademais, a documentação assinalada pelos participantes assegura o governo de dados como é defendido por Madera e Laurent, (2016), Stein e Morrison (2014). Neste sentido, o *Data Lake* promove a conformidade com o princípio 2, ao acautelar que "todos os

mapeamentos de domínios de valores dos vários sistemas fonte são documentados e mapeados ou tratados nesse mesmo *Data Lake*" (E2_N_DRE).

A eficiência organizacional, defendida em grande parte pelo entrevistado E3_T_CAO e também por Gupta e Giri (2018), torna a atividade regulatória mais eficiente e automatizada (E5_T_CRI). Por este motivo, desempenha um papel fundamental na automatização e rapidez do processos, colaborando, assim, no desenvolvimento do princípio 5, "Tempestividade" (Banco de Portugal, 2019).

Figura 5.2

Relação entre as Vantagens e os Princípios do BCBS 239



Fonte: Elaborado pelo próprio autor (2021).

Por último, a Figura 5.2 apresenta, de forma esquemática, as relações entre as vantagens do *Data Lake* e a suas implicações no cumprimento dos princípios abordados anteriormente. Pode concluir-se que a implementação do *Data Lake* irá apoiar essencialmente a "Arquitetura de Dados e Infraestrutura de IT", motivado pelo repositório único.

5.2.3. Desafios

No que toca aos desafios impostos pelo *Data Lake*, o entrevistado E2_N_DRE defende que "não é fácil atribuir o *data owner*" (E2_N_DRE), ilustrando a ambiguidade do *ownership* apresentada pelo Dama International (2017), "(…) Porque é que sou eu? Porque é que é que o outro? Isto não é fácil, porque obviamente isto significa responsabilidade" (E2_N_DRE).

No seguimento da ambiguidade de conceitos, a complexidade da definição dos termos evasivos, como apresentam Stein e Morrison (2014), também foi salientada pelos entrevistados. Em particular, o entrevistado E2_N_DRE admite que "há muito um processo negocial nesta questão de uniformização dos conceitos e isto não é uma coisa fácil".

Ambos os desafios mencionados estão relacionados com o governo de dados, mais especificamente, com a definição de conceitos, funções e responsabilidades, sendo associados ao princípio 2.

Por outro lado, associado à complexidade do processo defendida por Fang (2015), os participantes levantam questões sobre a conformidade ao nível do grupo, impactando o cumprimento do princípio 4, "Completude". De acordo com o entrevistado E4_T_CDO, existem questões "sobre o que é que deve ser delegado ou o que é que não deve ser delegado", ao mesmo tempo, que persistem "fluxos de dados paralelos nas entidades", dificultando a consistência da informação.

Figura 5.3

Relação entre os Desafios e os Princípios do BCBS 239

Fonte: Elaborado pelo próprio autor (2021).

Em suma, os princípios 2 e 4 são os que apresentam mais obstáculos à sua conformidade, como é ilustrado na Figura 5.3.

5.3. BCBS 239

Adicionalmente à literatura existente, os entrevistados apresentam vantagens e desafios que advêm do esforço para agir em conformidade com os princípios determinados pela Basel Committee on Banking Supervision (2013).

Nessa perspetiva, os participantes consideram que o processo é bastante complexo, demorando anos até conseguir atingir a conformidade (E4_T_CDO). Uma vez atingida, é necessário garantir que a conformidade é mantida (E2_N_DRE). Além disso, como em muitos regulamentos publicados pelos reguladores e segundo Butler (2017) e Orgeldinger (2018), os princípios são abstratos, vagos e difíceis de interpretar.

No entanto, também foram consideradas consequências positivas perante esta conformidade. No que toca ao negócio, há uma melhoria na experiência do cliente (E3_T_CAO) e, posteriormente, o número de clientes fidelizados aumenta (E4_T_CDO). A nível técnico, o acesso aos dados será mais acessível por força de uma lógica "*self-service*" (E4_T_CDO), ao mesmo tempo que o nível de adaptabilidade e detalhe requerido nos reportes será facilmente cumprido.

Em suma, na opinião do entrevistado E1_N_CRE, o cumprimento dos princípios do BCBS 239 é "claramente uma mais-valia para os bancos", apesar de ser um "reforço quer a nível informático, quer a nível de reporte, muito grande" (E1_N_CRE). O BCBS 239 é uma "pequena revolução em curso" (E2_N_DRE) que tem de ser cumprida, caso contrário, o mercado obrigará a fazê-la (E2_N_DRE).

6. CONCLUSÕES

6.1. Principais Conclusões

A preocupação com a qualidade dos dados é cada vez maior e, por isso, a Basel Committee on Banking Supervision (2013) publicou 14 princípios. Estes princípios têm como objetivo melhorar a qualidade dos dados reportados por parte dos bancos e, consequentemente, melhorar a eficácia da supervisão bancária. Por esse motivo e aliado ao facto de existirem múltiplos repositórios, a organização em estudo está em fase de implementação de um repositório único, o *Data Lake*.

Atualmente, a existência de várias fontes de dados e de processos não automatizados têm gerado dificuldades na atividade de *reporting*, prejudicando a conformidade com os princípios "2. Arquitetura de Dados e Infraestrutura de TI", "3. Exatidão e Integridade" e "4. Completude". Além disso, não existe um modelo de governo de dados como é definido pelo princípio "1. *Governance*".

A implementação do *Data Lake* na organização apresenta várias vantagens: (1) o facto de ser um repositório único, em consonância com o princípio 2, permite a integração dos dados e controlos da qualidade uniformes para todo o grupo bancário; (2) a documentação potencializa a uniformização de conceitos de negócio e de processos, também em concordância com o princípio 2; (3) por último, em conformidade com o princípio "5. Tempestividade", a melhoria da eficiência organizacional ao automatizar os procedimentos que, consequentemente, conduz a que o foco do colaborador seja o seu objetivo de negócio e não o seu processo.

No entanto, para beneficiar da tecnologia, o modelo de governo de dados tem de estar bem definido, sendo um desafio a ultrapassar. A ambiguidade de conceitos, seja a atribuição de *ownership* ou a elaboração do dicionário de dados, torna o processo de implementação desta tecnologia ainda mais complexo.

Comparativamente ao cenário atual, o princípio 2 é o que mais evolui no seu grau de conformidade, principalmente pela criação de uma *golden source*. A *golden source* também potencializa os demais princípios, com foco na integração, rastreabilidade e validações dos dados. Além disso, a introdução de um modelo de governo de dados é uma novidade para os procedimentos, agindo em concordância com o princípio 1.

Em suma e em resposta à questão de investigação, pode concluir-se que, pela análise de dados, o *Data Lake* é um excelente meio para atingir a conformidade com o regulador, visto que as suas vantagens vão ao encontro das especificidades requeridas, melhorando a qualidade dos dados reportados.

6.2. Implicações

Numa vertente teórica, os estudos existentes focam mais os aspetos tecnológicos do *Data Lake*, pelo que o presente estudo pretende alargar o conhecimento de um ponto de vista mais estratégico. Além disso, a literatura carece de investigações dedicadas a arquiteturas de dados no setor bancário, sendo que a presente investigação permitiu recolher *insights* sobre as dificuldades sentidas, bem como as expetativas das vantagens da implementação de um novo modelo de dados.

Sob o ponto de vista mais prático, e tendo em conta o grau de abstração dos princípios do BCBS 239 (Orgeldinger, 2018), o atual *case study* aprofunda a implementação de um método específico, a tecnologia do *Data Lake*, para fazer cumprir os princípios, tornando-se num exemplo que pode ser seguido.

6.3. Limitações e Investigações Futuras

As principais limitações deste estudo prendem-se com o facto de a implementação da tecnologia ainda estar em desenvolvimento, pelo que o estudo foi feito com base em opiniões e expetativas. Neste sentido, poderá ser interessante, no futuro, avaliar a conformidade com os princípios do BCBS 239 após o projeto estar implementado e comparar os resultados com as expetativas. Além disso, *o case study* recai apenas sob uma organização, de maneira que as futuras investigações podem analisar como é que os restantes bancos respondem a estes princípios. Por último, o estudo da implementação de tecnologias distintas do *Data Lake* também seria notável.

REFERÊNCIAS

Alshboul, Y., Nepali, R. K., & Wang, Y. (2015). Big Data LifeCycle: Threats and Security Model. *AMCIS*. https://www.researchgate.net/publication/281079716

Arass, M. El, Tikito, I., & Souissi, N. (2017). Data lifecycles analysis: Towards intelligent cycle. *Intelligent Systems and Computer Vision, ISCV 2017*.

Banco de Portugal. (2019). *Análise temática sobre qualidade de dados (BCBS 239): conclusões preliminares*.

Banco de Portugal. (2020). *Reserva para outras instituições de importância sistémica | Banco de Portugal*. https://www.bportugal.pt/page/reserva-de-o-sii

Barth, M. E., & Landsman, W. R. (2010). How did financial reporting contribute to the financial crisis? *European Accounting Review*, *19*(3), 399–423.

Basel Committee on Banking Supervision. (2013). *Basel Committee on Banking Supervision Principles for effective risk data aggregation and risk reporting*. www.bis.org

Bedeley, R., & Iyer, L. S. (2014). Big Data Opportunities and Challenges: the Case of Banking Industry. *SAIS 2014 Proceedings*, *1*, 7. http://aisel.aisnet.org/sais2014/2/

Butler, T. (2017). Towards a Standards - Based Technology Architecture for RegTec. *JOURNAL OF FINANCIAL TRANSFORMATION*, *45*(1), 49–59.

Cai, L., & Zhu, Y. (2020). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, *14*, 1–10.

Chaudhuri, S., & Dayal, U. (1997). An Overview of Data Warehousing and OLAP Technology. *ACM SIGMOD Record*, *26*(1), 65–74.

Ćurko, K., Bach, M. P., & Radonić, G. (2007). Business intelligence and business process management in banking operations. *Proceedings of the International Conference on Information Technology Interfaces, ITI*, 57–62.

Dama International. (2017). *DAMA-DMBOK: Data Management Body of Knowledge* (2nd ed.). Technics Publications.

El Arass, M., & Souissi, N. (2018). Data Lifecycle: From Big Data to SmartData. *IEEE 5th International Congress on Information Science and Technology (CiSt)*, 80–87.

Fang, H. (2015). Managing Data Lakes in Big Data. *2015 IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, IEEE-CYBER 2015*, 820–824. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7288049

Gerring, J. (2004). What Is a Case Study and What Is It Good for ? *The American Political Science Review*, *98*(2), 341–354.

Gupta, S., & Giri, V. (2018). Practical enterprise data lake insights: Handle data-driven challenges in an enterprise big data lake. Em *Practical Enterprise Data Lake Insights: Handle Data-Driven Challenges in an Enterprise Big Data Lake* (1.ª ed.). Apress Media LLC.

Hai, R., Geisler, S., & Quix, C. (2016). Constance: An Intelligent Data Lake System. *Proceedings of the 2016 International Conference on Management of Data*, 2097–2100.

Hoffer, J., Venkataraman, R., & Topi, H. (2017). *Modern Database Management* (Pearson (ed.); 13.ª ed.).

International Organization for Standardization. (2005). *Quality Management Systems-Fundamentals and Vocabulary*. ISO Press.

Kadadi, A., Agrawal, R., Nyamful, C., & Atiq, R. (2014). Challenges of Data Integration and Interoperability in Big Data. *2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, 38–40. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7004486

Khatri, V., & V. Brown, C. (2010). Designing Data Governance. *Communications of the ACM*, *53*(1), 148–152.

Madera, C., & Laurent, A. (2016). The Next Information Architecture Evolution : The Data Lake Wave. *8th International Conference on Management of Digital EcoSystems*, 174–180.

Mark Saunders, Philip Lewis, & Adrian Thornhill. (2016). *Research Methods for Business Students* (7th ed.). Pearson Education. https://books.google.co.ma/books/about/Research_Methods_for_Business_Students.html?id=LtiQvwEACAAJ&redir_esc=y

Martins, J. C., & Belfo, F. (2010). Métodos de investigação qualitativa estudos de casos na investigação em sistemas de informação. *Proelium-Revista da Academia Militar*, *14*, 39–72.

Mehmood, H., Gilman, E., Cortes, M., Kostakos, P., Byrne, A., Valta, K., Tekes, S., & Riekki, J. (2019). Implementing Big Data Lake for Heterogeneous Data Sources. *2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW)*, 37–44. https://ieeexplore.ieee.org/document/8750951/

Moges, H. T., Dejaeger, K., Lemahieu, W., & Baesens, B. (2013). A multidimensional analysis of data quality for credit risk management: New insights and challenges. *Information and Management*, *50*(1), 43–58.

Nargesian, F., Zhu, E., Miller, R. J., Pu UOIT, K. Q., & Arocena, P. C. (2019). Data Lake Management: Challenges and Opportunities. *Proceedings of the VLDB Endowment*, *12*(12), 1986–1989.

Nogueira, I. D., Romdhane, M., & Darmont, J. (2018). Modeling Data Lake Metadata with a Data Vault. *IDEAS 2018: 22nd International Database Engineering & Applications Symposium*, 253–261.

Orgeldinger, J. (2018). The Implementation of Basel Committee BCBS 239: Short analysis of the new rules for Data Management. *Journal of Central Banking Theory and Practice*, *7*(3), 57–72.

Singh, A., & Ahmad, S. (2019). Architecture of Data Lake. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *5*(2), 2456–3307.

SpeechTexter. (2014). *SpeechTexter*. https://www.speechtexter.com/

Stein, B., & Morrison, A. (2014). The enterprise data lake: Better integration and deeper analytics. *Technology Forecast: Rethinking integration*, *1*.

ANEXOS

Anexo A - Guiões da Entrevista

Tabela A1 – Guião da Entrevista de Negócio

Pergunta	Objetivo de Investigação	Questão de Investigação
P1. Quais são as principais dificuldades sentidas em reportar dados estatísticos?	Identificar as dificuldades sentidas atualmente;	De que forma está a organização em conformidade, ou não, com os princípios BCBS 239?
P2. Considera que reporta informações cruciais? Na sua opinião, está sempre disponível à hora certa? A informação está correta? Existe uma versão da verdade?	Analisar a situação atual dos processos de *reporting*;	Como é que o *Data Lake* impacta a concretização dos princípios BCBS 239?
P3. Qual a sua opinião sobre a qualidade dos dados? Sente que a arquitetura atual dos dados está preparada e adequada à agregação de dados? Considera que os dados são precisos?	Analisar a situação atual dos processos de *reporting*;	Como é que o *Data Lake* impacta a concretização dos princípios BCBS 239?
P4. Considera que os procedimentos são flexíveis? É possível usar os mesmos dados para diferentes fins?	Analisar a situação atual dos processos de *reporting*;	Como é que o *Data Lake* impacta a concretização dos princípios BCBS 239?
P5. Os procedimentos estão documentados detalhadamente? Existem soluções alternativas para o mesmo procedimento?	Analisar a situação atual dos processos de *reporting*;	Como é que o *Data Lake* impacta a concretização dos princípios BCBS 239?
P6. Todos os reportes elaborados têm um destino? Conhece os seus objetivos? A equipa consegue dar resposta a todos os pedidos em tempo útil?	Analisar a situação atual dos processos de *reporting*;	Como é que o *Data Lake* impacta a concretização dos princípios BCBS 239?
P7. Conhece o *Data Lake*? Qual é a sua perceção sobre o mesmo? Gostaria de trabalhar com este repositório de dados?	Identificar as expetativas para a implementação do *Data Lake*;	Quais são as vantagens e os desafios do *Data Lake* do ponto de vista da organização?
P8. Na sua opinião, como é que uma *golden source* afetaria o *reporting*? Quais as suas vantagens e desvantagens?	Estudar vantagens e desvantagens do *Data Lake*;	Quais são as vantagens e os desafios do *Data Lake* do ponto de vista da organização?
P9. Tendo em conta que o BCBS 239 tem como objetivo fortalecer a capacidade de agregação dos dados e define princípios para serem cumpridos, considera que o facto de fazer cumprir os seus princípios se torna num obstáculo ou numa ajuda para o reporte de dados? Quais são as dificuldades sentidas?	Identificar as dificuldades sentidas atualmente;	De que forma está a organização em conformidade, ou não, com os princípios BCBS 239?

Fonte: Elaborado pelo próprio autor (2021).

Tabela A2 – Guião da Entrevista Técnica

Pergunta	Objetivo de Investigação	Questão de Investigação
P1. O *Data Lake* foi construído de raiz ou tem por base ferramentas/processos já existentes?	Identificar os objetivos da implementação do *Data Lake* na organização;	Quais são as vantagens e os desafios do *Data Lake* do ponto de vista da organização?
P2. Quais as principais razões que motivaram o desenvolvimento do *Data Lake*?	Compreender as motivações para a implementação do *Data Lake*;	Quais são as vantagens e os desafios do *Data Lake* do ponto de vista da organização?
P3. Quais os principais objetivos para o desenvolvimento do *Data Lake*? O que é que se quer atingir com este projeto? Qual a sua maior vantagem?	Identificar os objetivos da implementação do *Data Lake* na organização;	Quais são as vantagens e os desafios do *Data Lake* do ponto de vista da organização?
P4. Tendo em conta que o BCBS 239 tem como objetivo fortalecer a capacidade de agregação dos dados, considera que o facto de cumprir os seus princípios foi preponderante na decisão do desenvolvimento do *Data Lake* ou apenas um extra?	Compreender as motivações para a implementação do *Data Lake*;	Quais são as vantagens e os desafios do *Data Lake* do ponto de vista da organização?
P5. Na sua opinião, como é que uma *golden source* afetaria a empresa? Quais as suas vantagens e desvantagens? Em particular, qual as consequências para a concretização do BCBS 239?	Estudar vantagens e desvantagens do *Data Lake*;	Quais são as vantagens e os desafios do *Data Lake* do ponto de vista da organização?
P6. Qual será a abordagem para o *Data Governance*? Quais são as políticas e regras que considera mais importante no *Data Lake*?	Relacionar as vantagens do *Data Lake* com os princípios do BCBS 239;	Como é que o *Data Lake* impacta a concretização dos princípios BCBS 239?
P7. Como será garantida a qualidade dos dados? Como será abordada a gestão de metadados? Existirão controlos a todos os ciclo de vida dos dados? Como são definidos os atributos de cada dado?	Relacionar as vantagens do *Data Lake* com os princípios do BCBS 239;	Como é que o *Data Lake* impacta a concretização dos princípios BCBS 239?
P8. O desenvolvimento do *Data Lake* tem em consideração o conhecimento e *skills* dos futuros utilizadores? Como e com que frequência será efetuada a extração dos dados? Quais são as principais preocupações?	Estudar vantagens e desvantagens do *Data Lake*;	Quais são as vantagens e os desafios do *Data Lake* do ponto de vista da organização?
P9. Que direções participam no desenvolvimento do *Data Lake*? Qual a sua importância no desenvolvimento deste projeto? Quais as direções que mais beneficiarão?	Identificar as expetativas para a implementação do *Data Lake*;	Quais são as vantagens e os desafios do *Data Lake* do ponto de vista da organização?
P10. Concorda com a frase "um projeto de *Data Lake* é mais um projeto de *Governance* do que análise de dados"?	Identificar as expetativas para a implementação do *Data Lake*;	Quais são as vantagens e os desafios do *Data Lake* do ponto de vista da organização?

Fonte: Elaborado pelo próprio autor (2021).

Anexo B – Definição dos Códigos MAXQDA

Tabela B1 – Definição Conceptual dos Códigos

Subcódigo	Tipo	Definição Operacional ou Conceptual
Reporting - Desafios		
Múltiplas Fontes de Dados	Prévio	A indústria bancária contém dados em várias fontes, provocando erros e inconsistências nos dados (Cai & Zhu, 2020; Chaudhuri & Dayal, 1997; Kadadi et al., 2014).
Recursos Inadequados	Prévio	Os recursos inadequados para a integração e exploração dos dados, sejam eles financeiros, humanos ou tecnológicos (Stein & Morrison, 2014).
Acesso à Informação	Prévio	Não só pelo facto de a recolha de dados ser feita por vários departamentos, como também pelo facto de quem acede aos dados, (...) , pode descartar os dados que não são relevantes (Moges et al., 2013).
Conhecimento Geral do Reporte	Prévio	Os regulamentos e as regras publicadas não são concretas, nem fáceis de interpretar, originando lacunas entre a intenção dos requisitos regulatórios e a interpretação dos mesmos pelas organizações (Butler, 2017).
Documentação	Prévio	A maioria do conhecimento é tácito (Moges et al., 2013). (...) Há, também, carência de padrões e de dicionários de dados (Butler, 2017).
Inputs Manuais	Prévio	Existem muitos dados que são inseridos manualmente nos sistemas de informação do banco (Moges et al., 2013).
Peso Regulatório	Novo	
Data Lake - Objetivos		
Governo de Dados	Novo	
Repositório Único	Prévio	O armazenamento dos dados num só repositório, ao invés de repositórios heterogêneos (Mehmood et al., 2019).
BCBS 239	Prévio	O foco deve ser sempre um problema de negócio, o qual deve ser respondido através da implementação do *Data Lake* (Moges et al., 2013; Stein & Morrison, 2014).
Qualidade de Dados	Novo	
Data Lake - Drivers		
BCBS 239	Novo	
Múltiplas Fontes de Dados	Novo	
Eficiência Operacional	Novo	
Big Data	Novo	
Estratégia Digital	Novo	
Governo de Dados	Novo	
Controlos de Qualidade	Novo	

Data Lake - Vantagens		
Eficiência Operacional	Prévio	O *Data Lake* melhora a eficiência operacional ao diminuir o tempo na preparação e análise pormenorizada dos dados (Gupta & Giri, 2018).
Documentação	Prévio	O *Data Lake* é controlado por regras, ferramentas e processos, garantindo o governo de dados, bem como a sua qualidade. A documentação é acessível e pode ser refinada colaborativamente (Madera & Laurent, 2016; Stein & Morrison, 2014).
Repositório Único	Prévio	O facto de ser um repositório centralizado de dados, ou seja, com vários tipos de dados (Fang, 2015).
Controlos de Qualidade	Novo	
Características da Tecnologia	Prévio	O armazenamento do *Data Lake* é escalável e tem baixo custo, (...) sem comprometer a qualidade e disponibilidade dos dados (Gupta & Giri, 2018; Mehmood et al., 2019; Moges et al., 2013; Singh & Ahmad, 2019).
Data Lake - Desafios		
Data Ownership	Prévio	A ambiguidade existente no *ownership* dos dados (…) dificulta a determinação do *owner* de um dado (Dama International, 2017)
Uniformização de Conceitos	Prévio	Os termos ambíguos devem ser definidos no dicionário de dados, através do entendimento entre os interessados, convergindo para uma uniformização de conceitos de negócio (Stein & Morrison, 2014).
Visão de Grupo	Novo	
Processo Complexo	Prévio	O processo de implementação do *Data Lake* requer muito esforço técnico (Fang, 2015).
Tamanho da organização	Novo	
Novas *Skills*	Prévio	A maioria dos utilizadores deste repositório carecem de habilidades mais técnicas, para transformarem e analisar os dados com as novas ferramentas (Fang, 2015).
BCBS 239 - Data Governance		
1. *Governance*	Prévio	A estrutura de dados deve estar definida em políticas da empresa sobre a confidencialidade, integridade e disponibilidade dos dados (Basel Committee on Banking Supervision, 2013).
2. Arquitetura de Dados e Infraestrutura de TI	Prévio	O banco deve estabelecer taxonomias, caracterizar os dados, usar identificadores únicos, integrar os dados de todo o grupo bancário e garantir que existem controlos adequados para todo o ciclo de vida de um dado (Basel Committee on Banking Supervision, 2013).
BCBS 239 - Integração de Dados		
3. Exatidão e Integridade	Prévio	Um banco deve ser capaz de gerar dados precisos e confiáveis. Para minimizar a probabilidade de erros, os dados devem ser agregados numa base de dados automatizada e devem ser comparados com os dados contabilísticos (Basel Committee on Banking Supervision, 2013).
4. Completude	Prévio	Um banco deve ser capaz de capturar e agregar todos os dados de risco do seu grupo, com o nível de detalhe requerido pelo reporte em questão (Basel Committee on Banking Supervision, 2013).
5. Tempestividade	Prévio	Um banco deve ser capaz de agregar dados atualizados em tempo útil, sem descuidar dos princípios relativos à exatidão, integridade e adaptabilidade (Basel Committee on Banking Supervision, 2013).

6. Adaptabilidade	Prévio	Um banco deve ser capaz de agregar dados de risco de maneira a responder às diversas solicitações de reporte (Basel Committee on Banking Supervision, 2013).
BCBS 239 - Práticas de Reporting		
7. Exatidão	Prévio	Os relatórios devem transmitir os dados agregados de maneira exata (Basel Committee on Banking Supervision, 2013).
8. Abrangência	Prévio	A profundidade e o âmbito dos relatórios devem ser coerentes com o tamanho e complexidade das operações e do perfil de risco do banco, bem como com os requisitos dos destinatários (Basel Committee on Banking Supervision, 2013).
9. Clareza e Utilidade	Prévio	Os relatórios devem conseguir transmitir informação de uma forma clara e concisa, facilitando a sua leitura (Basel Committee on Banking Supervision, 2013).
10. Frequência	Prévio	Os requisitos da frequência dependem de fatores como as necessidades da receção dos dados, a natureza do risco e a volatilidade do risco (Basel Committee on Banking Supervision, 2013).
11. Distribuição	Prévio	Os relatórios devem ser distribuídos a todas as partes relevantes, assegurando a confidencialidade dos dados (Basel Committee on Banking Supervision, 2013).
BCBS 239 - Revisão Regulatória		
12. Revisão	Prévio	Os supervisores devem rever e avaliar a conformidade dos bancos com os 11 princípios já definidos (Basel Committee on Banking Supervision, 2013).
13. Ações Corretivas e Medidas de Supervisão	Prévio	Os supervisores devem usar ferramentas e recursos apropriados para exigirem ações corretivas, eficazes e oportunas, de modo a resolver deficiências nos recursos de agregação de dados de risco do banco (Basel Committee on Banking Supervision, 2013).
14. Cooperação	Prévio	Os supervisores devem cooperar com os supervisores relevantes de outras jurisdições no que toca à supervisão e à revisão dos princípios definidos, bem como a implementação de qualquer ação corretiva (Basel Committee on Banking Supervision, 2013).
BCBS 239 - Oportunidades Futuras		
Estruturas dos Futuros Reportes	Novo	
Experiência ao Cliente	Novo	
Exploração de dados	Novo	
BCBS 239 - Desafios		
Processo Complexo	Novo	
Correta Interpretação	Novo	
Inputs Manuais	Novo	
Manutenção	Novo	

Fonte: Elaborado pelo próprio autor (2021).

Tabela B2 – Critérios de Avaliação dos Princípios BCBS 239

Princípio 1 - Governance

Modelo de Governo;

Políticas;

Framework de Risco e Estratégia;

Validação Independente;

Princípio 2 - Arquitetura de Dados e Infraestrutura de TI

Golden Source;

Dicionário de Conceitos e Repositório de Metadados;

Funções e Responsabilidades;

Controlos de Qualidade;

Princípio 3 - Exatidão e Integridade

Reconciliações;

Processos Manuais e *End-User Applications* (EUA);

Rastreabilidade;

Validações Automáticas;

Princípio 4 - Completude

Perímetro de Consolidação;

Homogeneização de Sistemas, Processos e Conceitos;

Princípio 5 - Tempestividade

Automatização e Rapidez dos Processos;

Procedimentos em Situações de Crise;

Princípio 6 - Adaptabilidade

Flexibilidade dos Processos e Sistemas;

Capacidade de Interpretação e Implementação de Novos Requisitos;

Princípios 7 a 11

Validações de Consistência entre Reportes;

Requisitos para os Reportes Internos (Abrangência, Periodicidade, Destinatários e Aprovação, Conteúdos);

Fonte: Banco de Portugal (2019, p.9-15)

Anexo C – Frequências dos Códigos e Subcódigos MAXQDA

Tabela C1 – Frequências dos Códigos

Códigos	**Segmentos** [i]		**Entrevistas** [ii]	
	Frequência	**Percentagem**	**Frequência**	**Percentagem**
Reporting	15	2,47%	4	80%
Reporting\Desafios	128	21,09%	4	80%
Data Lake\Objetivos	16	2,64%	3	60%
Data Lake\Drivers	53	8,73%	4	80%
Data Lake\Vantagens	90	14,83%	5	100%
Data Lake\Desafios	47	7,74%	2	40%
BCBS 239\Princípios	232	38,22%	5	100%
BCBS 239\Desafios	11	1,81%	3	60%
BCBS 239\Oportunidades Futuras	15	2,47%	2	40%
Total	**607**	**1**		

[i] Número de vezes que cada código foi abordado na totalidade das entrevistas.

[ii] Número de entrevistados que mencionou cada um dos códigos.

Fonte: Elaborado pelo próprio autor (2021).

Tabela C2 – Frequências dos Subcódigos

Código	**Subcódigo**	**Segmentos** [i]		**Entrevistados** [ii]	
		Frequência	**Percentagem**	**Frequência**	**Percentagem**
Reporting		15	2,47%	4	80%
Reporting\Desafios	Múltiplas Fontes de Dados	34	5,60%	4	80%
Reporting\Desafios	Recursos Inadequados	29	4,78%	3	60%
Reporting\Desafios	Acesso à Informação	25	4,12%	4	80%
Reporting\Desafios	Conhecimento Geral do Reporte	16	2,64%	2	40%
Reporting\Desafios	Documentação	13	2,14%	4	80%
Reporting\Desafios	Peso Regulatório	9	1,48%	2	40%
Reporting\Desafios	Inputs Manuais	2	0,33%	1	20%
Data Lake\ Objetivos	Governo de Dados	7	1,15%	3	60%
Data Lake\ Objetivos	Repositório Único	4	0,66%	2	40%
Data Lake\ Objetivos	BCBS 239	3	0,49%	2	40%
Data Lake\ Objetivos	Qualidade de Dados	2	0,33%	1	20%
Data Lake\Drivers	BCBS 239	13	2,14%	4	80%
Data Lake\Drivers	Múltiplas Fontes de Dados	12	1,98%	4	80%
Data Lake\Drivers	Eficiência Operacional	8	1,32%	2	40%

Data Lake\Drivers	*Big data*	6	0,99%	2	40%
Data Lake\Drivers	Estratégia Digital	6	0,99%	2	40%
Data Lake\Drivers	Governo de Dados	4	0,66%	2	40%
Data Lake\Drivers	Controlos de Qualidade	4	0,66%	2	40%
Data Lake\ Vantagens	Eficiência Operacional	30	4,94%	4	80%
Data Lake\ Vantagens	Documentação	21	3,46%	4	80%
Data Lake\ Vantagens	Repositório Único	20	3,29%	4	80%
Data Lake\ Vantagens	Controlos de Qualidade	13	2,14%	5	100%
Data Lake\ Vantagens	Características da Tecnologia	6	0,99%	3	60%
Data Lake\ Desafios	Data *Ownership*	18	2,97%	2	40%
Data Lake\ Desafios	Uniformização de Conceitos	8	1,32%	1	20%
Data Lake\ Desafios	Visão de Grupo	7	1,15%	1	20%
Data Lake\ Desafios	Processo Complexo	5	0,82%	2	40%
Data Lake\ Desafios	Tamanho da Organização	5	0,82%	2	40%
Data Lake\ Desafios	Novas *Skills*	4	0,66%	2	40%
BCBS 239\ Princípios	Data *Governance*\1. *Governance*	57	9,39%	5	100%
BCBS 239\ Princípios	Data *Governance*\2. Arquitetura de Dados e Infraestrutura de IT	100	16,47%	5	100%
BCBS 239\ Princípios	Integração de Dados\3. Exatidão e Integridade	26	4,28%	5	100%
BCBS 239\ Princípios	Integração de Dados\4. Completude	14	2,31%	4	80%
BCBS 239\ Princípios	Integração de Dados\ 5. Tempestividade	10	1,65%	4	80%
BCBS 239\ Princípios	Integração de Dados\ 6. Adaptabilidade	10	1,65%	3	60%
BCBS 239\ Princípios	Práticas de *Reporting*\7. Exatidão	4	0,66%	2	40%
BCBS 239\ Princípios	Práticas de *Reporting*\ 8. Abrangência	0	0,00%	0	0%
BCBS 239\ Princípios	Práticas de *Reporting*\	6	0,99%	2	40%

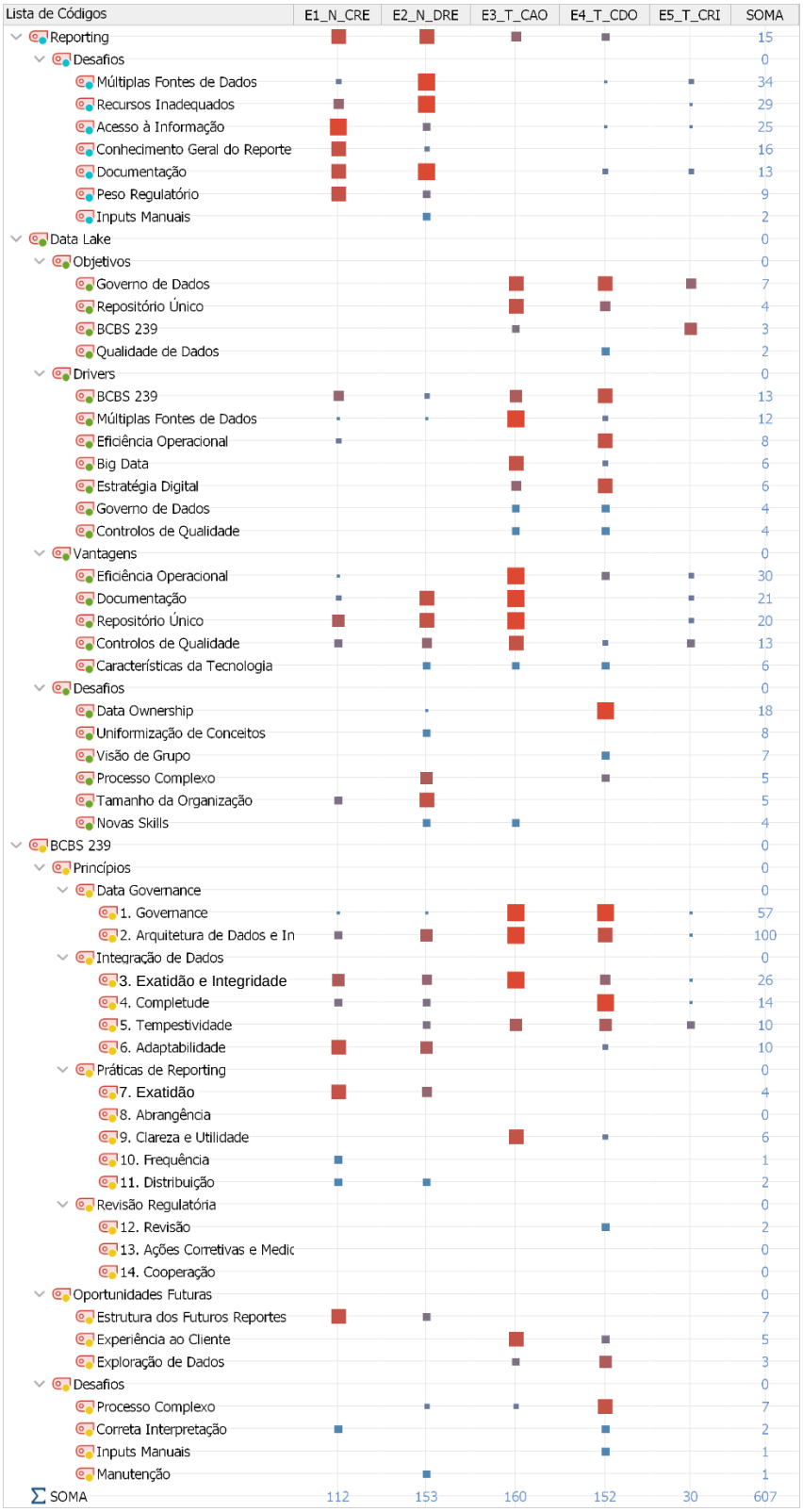
BCBS 239\ Princípios	9. Clareza e Utilidade Práticas de *Reporting*\ 10. Frequência	1	0,16%	1	20%
BCBS 239\ Princípios	Práticas de *Reporting*\ 11. Distribuição	2	0,33%	2	40%
BCBS 239\ Princípios	Revisão Regulatória\12. Revisão	2	0,33%	1	20%
BCBS 239\ Princípios	Revisão Regulatória\ 13. Ações Corretivas e Medidas de Supervisão	0	0,00%	0	0%
BCBS 239\ Princípios	Revisão Regulatória\ 14. Cooperação	0	0,00%	0	0%
BCBS 239\ Oportunidades Futuras	Estrutura dos Futuros Reportes	7	1,15%	2	40%
BCBS 239\ Oportunidades Futuras	Experiência ao Cliente	5	0,82%	2	40%
BCBS 239\ Oportunidades Futuras	Exploração de Dados	3	0,49%	2	40%
BCBS 239\ Desafios	Processo Complexo	7	1,15%	3	60%
BCBS 239 \Desafios	Correta Interpretação	2	0,33%	2	40%
BCBS 239\ Desafios	Inputs Manuais	1	0,16%	1	20%
BCBS 239\ Desafios	Manutenção	1	0,16%	1	20%
Total		**607**	**100%**		

[i] Número de vezes que cada código foi abordado na totalidade das entrevistas.

[ii] Número de entrevistados que mencionou cada um dos códigos.

Fonte: Elaborado pelo próprio autor (2021).

Tabela C3 – Visualizador da Matriz de Códigos do MAXQDA



Lista de Códigos	E1_N_CRE	E2_N_DRE	E3_T_CAO	E4_T_CDO	E5_T_CRI	SOMA
Reporting						15
Desafios						0
Múltiplas Fontes de Dados						34
Recursos Inadequados						29
Acesso à Informação						25
Conhecimento Geral do Reporte						16
Documentação						13
Peso Regulatório						9
Inputs Manuais						2
Data Lake						0
Objetivos						0
Governo de Dados						7
Repositório Único						4
BCBS 239						3
Qualidade de Dados						2
Drivers						0
BCBS 239						13
Múltiplas Fontes de Dados						12
Eficiência Operacional						8
Big Data						6
Estratégia Digital						6
Governo de Dados						4
Controlos de Qualidade						4
Vantagens						0
Eficiência Operacional						30
Documentação						21
Repositório Único						20
Controlos de Qualidade						13
Características da Tecnologia						6
Desafios						0
Data Ownership						18
Uniformização de Conceitos						8
Visão de Grupo						7
Processo Complexo						5
Tamanho da Organização						5
Novas Skills						4
BCBS 239						0
Princípios						0
Data Governance						0
1. Governance						57
2. Arquitetura de Dados e In						100
Integração de Dados						0
3. Exatidão e Integridade						26
4. Completude						14
5. Tempestividade						10
6. Adaptabilidade						10
Práticas de Reporting						0
7. Exatidão						4
8. Abrangência						0
9. Clareza e Utilidade						6
10. Frequência						1
11. Distribuição						2
Revisão Regulatória						0
12. Revisão						2
13. Ações Corretivas e Medic						0
14. Cooperação						0
Oportunidades Futuras						0
Estrutura dos Futuros Reportes						7
Experiência ao Cliente						5
Exploração de Dados						3
Desafios						0
Processo Complexo						7
Correta Interpretação						2
Inputs Manuais						1
Manutenção						1
SOMA	112	153	160	152	30	607

Nota: O cálculo do tamanho do símbolo refere-se à linha em causa.

Fonte: Elaborado pelo próprio autor (2021).

Anexo D – Verbatim dos Subcódigos MAXQDA

Tabela D1 – Comentários Verbatim: Data Lake - Desafios Atuais

Entrevistado	Citação Exemplo
Conhecimento Geral do Reporte	
E1_N_CRE	"Ambiguidade dos reportes, porque há reportes que são muito ambíguos e tu ficas na dúvida: então, mas este campo, eu posso reportar este dado assim, mas olha, o dado B também fica bem aqui assim."
E2_N_DRE	"Não sei se isto é problema transversal a várias empresas ou não, admito que sim, o foco no meio e não no objetivo final é recorrente, a maior parte de nós foca-se no que controla, no entregável."
Recursos Inadequados	
E1_N_CRE	"Não é fácil fazer controlos de milhares de registos, não tendo ferramentas apropriadas para o fazer"
E2_N_DRE	"Termos processos que por vezes não são o mais eficiente possível, e, no fundo, estamos a fazer processos, a tratar dados várias vezes com objetivos diferentes, não é? Torna, muitas vezes, a capacidade de fazer outras coisas que não essa tarefa em si muito difícil."
E5_T_CRI	"Inexistência de uma base de dados única e corporativa da Instituição."
Inputs Manuais	
E2_N_DRE	"Muita da informação que nós tratamos, (…) depende de uma coisa que se chama *inputs* manuais, ou seja, carregamentos manuais. E, tudo o que é manual, obviamente, não é só, mas sobretudo, está sujeito a risco operacional."
Múltiplas Fontes de Dados	
E1_N_CRE	"Neste momento, temos várias direções, a fazer reportes semelhantes, mas cada um tem a sua fonte. E, isso acarreta problemas de coerência, problemas de critério, não quer dizer que os dados estejam errados, simplesmente o seu manuseamento com critérios diferentes, provoca resultados diferentes."
E2_N_DRE	"Os sistemas não estão integrados e nós para os diversos reportes nos vários órgãos de estrutura do banco, muitas vezes, inclusivamente, usamos extrações diferentes das mesmas realidades."
E4_T_CDO	"Depois há sempre o problema da consistência entre a informação entre as várias áreas que depois reportam para cima ou para fora."
E5_T_CRI	"Multiplicidade de processos fragmentados."
Documentação	
E1_N_CRE	"Nós não temos uma documentação formal dos procedimentos, o que inviabiliza, em certa parte, a perspetiva de uma alternativa, não é?"
E2_N_DRE	"E, sendo que o objetivo principal é entregar, o foco é entregar, obviamente que se eu tiver que documentar e se tiver que priorizar, eu vou priorizar, claramente, o entregar em relação ao documentar."
E4_T_CDO	"Quem precisa de os utilizar, perde muito tempo à procura deles, a desenvolver processos de extração e, depois, a fazer algum tipo de tratamento para os poder consumir."
E5_T_CRI	"Inexistência de um dicionário de dados comum."
Acesso à Informação	
E1_N_CRE	"Nós temos a capacidade de rapidamente se nos questionarem o valor que está (..), colocado numa célula, rapidamente nós conseguimos ter detalhe com a maior granularidade possível, e isso, neste momento, em alguns casos, ainda não está com possibilidade de fazer isso de forma ágil, ou de forma com relativa destreza."

E2_N_DRE	"Se houver algum tipo de questão/dúvida sobre os dados, temos que andar um bocadinho mais para trás e temos que pedir intervenção da própria IT, para nos ajudar a fazer o *drill down* dessa informação."
E4_T_CDO	"Evita-nos todo este trabalho, que é bastante repetitivo entre os vários departamentos do banco, as várias áreas do banco, que para além de ser um trabalho repetitivo e que é feito em paralelo por várias áreas."
E5_T_CRI	"Não existe rastreabilidade dos dados até aos sistemas operacionais."
Peso Regulatório	
E1_N_CRE	"Até hoje, nós temos conseguido sempre dar resposta a tudo. Obviamente que temos tido algumas dificuldades neste últimos 2 anos / 3 anos, por força do aumento da confrontação do Banco de Portugal."
E2_N_DRE	"A própria evolução, o próprio ritmo da exigência dos vários reportes que temos significa que a dinâmica é difícil manter a documentação atualizada."

Fonte: Elaborado pelo próprio autor (2021).

Tabela D2 – Comentários Verbatim: Data Lake - Objetivos

Entrevistado	Citação Exemplo
BCBS 239	
E3_T_CAO	"Criação desta *Golden Source* tem como principal objetivo, de facto, atacar os princípios base, atacar, no bom sentido, o BCBS 239, são 14 princípios que existem no BCBS 239, que se resumem em cinco grandes dimensões."
E5_T_CRI	"Responder aos princípios BCBS 239."
Repositório Único	
E3_T_CAO	"Termos uma *golden source* informacional para a organização."
E4_T_CDO	"É preciso que seja um repositório gerido com um patrocínio de padrões de *Golden Source*, só que promovam o *Golden Source* e, desde logo, com processos de desenvolvimento que promovam, utilizando aqueles artefacto de governo que eu referi há pouco, tendo técnicas de modelação de dados, que permitam garantir que não se está a criar o dado mais do que uma vez, dados bem arrumados para ver o seu consumo de utilização."
Qualidade de Dados	
E4_T_CDO	"Garantir que os dados têm a qualidade necessária para a sua utilização."
Governo de Dados	
E3_T_CAO	"O principal objetivo é tu teres os dados governados e com a qualidade que é merecida para uma organização e para a responsabilidade que uma organização, como <organização>, tem."
E4_T_CDO	"Os principais objetivos é conhecermos os dados, então, para isso, há dois temas, há um conjunto de artefactos que têm de ser geridos."
E5_T_CRI	"Criar uma cultura de dados: identificação *ownership*, identificação única dos atributos dos dados, um processo de controlo centralizado no *Data Lake*."

Fonte: Elaborado pelo próprio autor (2021).

Tabela D3 – Comentários Verbatim: Data Lake - Drivers

Entrevistado	Citação Exemplo
Big Data	
E3_T_CAO	"Tudo o que nós fazemos, nós geramos *terabytes* de informação a cada minuto."
E4_T_CDO	"Tem de garantir que (…), é uma tecnologia de *Data Lake* ou de *Big Data*, *NoSQL*."
BCBS 239	
E1_N_CRE	"Havendo o BCBS 239, temos o normativo que nos obriga a trabalhar se calhar um bocadinho mais rápido e com etapas pré-estabelecidas, o que também não é mau."
E2_N_DRE	"Eu não vejo claramente como um obstáculo, eu vejo com uma condição necessária."
E3_T_CAO	"A necessidade existia antes, mas claramente o BCBS foi o veículo que nos permitiu fazer isso"
E4_T_CDO	"Um *driver* regulatório, porque o setor financeiro é obrigatório em conformidade com o BCBS 239."
Múltiplas Fontes de Dados	
E1_N_CRE	"Deveríamos caminhar no sentido do que estamos a caminhar, que é do repositório único com os atributos todos reconciliados com os balancetes todos das entidades que compõem o grupo."
E2_N_DRE	"É termos informação, tentar ter uma única fonte de verdade, essa única fonte verdade esteja residente numa estrutura tecnológica única, documentada, *standardizada* e isso é fundamental."
E3_T_CAO	"Fruto também do seu crescimento nas últimas (…), criou múltiplos repositórios de informação, informação departamental, informação específica para certas áreas, informação para certos produtos e para certos canais e o que acontece é que tu criaste uma inconsistência informacional na organização."
E4_T_CDO	"Depois há sempre o problema da consistência entre a informação entre as várias áreas com que depois reportam para cima ou para fora."
Eficiência Operacional	
E1_N_CRE	"Tem essa necessidade para conseguir fazer o trabalho."
E4_T_CDO	"O segundo aspeto é um tema de eficiência operacional, os dados por não terem qualidade, (…), os problemas são os dois, não sabemos dos dados e depois que os encontramos, eles não têm qualidade."
Estratégia Digital	
E3_T_CAO	"Que independentemente da indústria e do país, que todas as empresas passam por este processo, umas mais maduras que outras."
E4_T_CDO	"Cada vez mais, há estratégias suportadas em cima do digital e, para termos certezas do digital, precisamos de ter dados e saber o que possamos encontrar, que estejam governados, que estejamos sempre à procura ou a criá-lo de raiz e que tenha uma qualidade necessária."
Governo de Dados	
E3_T_CAO	"A não atribuição de *ownership* a todos os elementos de dados, portanto tudo o que tu tens ao nível do elemento do dado, tem que ter um *owner*, (…), também não existia isso."
E4_T_CDO	"É preciso um modelo que diga o que é que significa a qualidade, como é que podemos medir a qualidade."
Controlos de Qualidade	
E3_T_CAO	"Ter controlos de qualidade para garantir que quando eu digo que "o cliente tem 7 produtos ativos", na realidade tem 7 produtos ativos e tu tens controlos de qualidade que validam isto, tudo isto não existe."
E4_T_CDO	"É preciso (…) ir medindo essa qualidade com distinção de *thresholds* do que é que significa ter qualidade ou não ter qualidade."

Fonte: Elaborado pelo próprio autor (2021).

Tabela D4 – Comentários Verbatim: Data Lake - Vantagens

Entrevistado	Citação Exemplo
Características da Tecnologia	
E2_N_DRE	"Admito que o DL, daquilo que me dizem, tem uma capacidade de armazenamento, tratamento, ferramentas de exploração de informação bastante mais versáteis do que outras estruturas de dados, estruturas de suporte à informação."
E3_T_CAO	"Tecnologia alinhada às melhores práticas, escalável e que tenha redundância, tenha *back-ups*, que tenha *disaster recovery*, etc.., é muito importante."
E4_T_CDO	"Este tipo de tecnologia permite depois ter outros tipos de dados, de informação e tratar os dados de outra maneira que a tecnologia clássica o obriga, mas é mais uma questão de, eu diria quase de, *performance*."
Eficiência Operacional	
E1_N_CRE	"Diminui significativamente o risco operacional de reporte, em termos de análises também permite ir logo diretamente à operação."
E3_T_CAO	"Risco regulatório tem muito a ver com nós termos uma visão mais, eu diria, robusta da nossa exposição ao risco, rácios de capital, solvabilidade, o nosso apetite o risco também na concessão de crédito, (…), temos compromissos regulatórios diariamente, semanalmente, mensalmente. Eu acho que isso é o dia-a-dia do banco."
E4_T_CDO	"Vai ser muito mais fácil quando a instituição precisar de tomar decisão, decisões, encontrar os seus dados, ter confiança nos seus dados, ou seja, o processo entre conseguir tomar uma decisão, produzir um relatório externo, identificar uma necessidade de um cliente, responder a essa necessidade do cliente, vai ser muito mais rápida e vai ser muito mais fiável."
E5_T_CRI	"Tornar toda a componente regulatória mais eficiente e automatizada."
Documentação	
E1_N_CRE	"Temos uma fonte, se tens só uma fonte, a característica dos atributos é única e está toda documentada."
E2_N_DRE	"Quando eu estou a falar de um código de finalidade, por exemplo, e este é apenas um exemplo, eu sei que eu estou a falar daquilo, como a direção de risco estará a falar daquilo, como a direção de planeamento estará a falar do mesmo. Porquê? Porque estamos a falar do mesmo conceito, o que é uma mais-valia gigante."
E3_T_CAO	"Teres documentação com qualidade sobre esse elemento de dado, seja de um ponto de vista de definição funcional, seja no ponto de vista de definição técnica, seja no ponto de vista de rastreabilidade."
E5_T_CRI	"A documentação sobre os dados é um dos pontos mais importantes."
Repositório Único	
E1_N_CRE	"Termos um local onde esteja tudo armazenado, devidamente catalogado e reconciliado a nível corporativo, não só a nível da sede, mas também a nível das entidades do grupo."
E2_N_DRE	"Passamos a ter a informação das empresas do grupo de uma forma também *standardizada*, também no mesmo tempo, e permite, de alguma forma, também aproveitar também algumas sinergias porque, no fundo, não há depois manutenção de extrações dos sistemas fontes para a casa mãe."
E3_T_CAO	"Tu só produzes o dado uma única vez, isto é um ponto muito importante, portanto tu tens que identificar de onde é que vem o dado, mas só produzes esta informação uma única vez. A data de nascimento dos clientes só pode existir uma vez, não pode estar espartilhada em não sei quantos repositórios informacionais."
E5_T_CRI	"BD única corporativa."

Controlos de Qualidade	
E1_N_CRE	"E, em que haja uma reconciliação contabilística, com os balanços das entidades e da própria sede também, e em que haja um controlo, vários controlos de qualidade da informação que reside nesse repositório."
E2_N_DRE	"O mais importante acho que é aquilo que referi, (...) e, obviamente, também controlos de qualidade uniformes, processos de correção uniformes."
E3_T_CAO	"Controlos de qualidade, que é garantir que o que tu estás a desenvolver bem, a qualidade que foi definida e que tu consegues garantir, portanto que, de facto, foi implementado aquilo que tinha sido decidido."
E4_T_CDO	"Temos aqui a qualidade monitorizada, *thresholds* definidos, vamos a trabalhar no dia-a-dia para que o nível esteja acima do *thresholds*, e se estiver, nós aceitamos e continuamos a trabalhar. Quando baixa daquele *thresholds*, temos de fazer alguma coisa."
E5_T_CRI	"Identificação dos controlos de qualidade por todos, a implementação de *dashboard* é critica para implementar medidas corretivas."

Fonte: Elaborado pelo próprio autor (2021).

Tabela D5 – Comentários Verbatim: Data Lake - Desafios

Entrevistado	Citação Exemplo
Processo Complexo	
E2_N_DRE	"O que significa que também é difícil, é um dos desafios que por acaso não tinha referido ainda, que é conseguir manter o foco, alargando cada vez mais o âmbito."
E3_T_CAO	"Os dados que foram ingeridos, (..), que seguem as normas do BCBS 239, é muito mais importante isso do que teres a maior robustez possível num algoritmo."
E4_T_CDO	"Porque obriga não só a criar aquela camada do governo e qualidade de dados, mas também a outra camada tecnológica, que é refazer na realidade todo o *layer* informacional de um banco, são programas que levam anos."
Tamanho da organização	
E1_N_CRE	"Obviamente que isto era muito fácil fazer numa entidade com balanço pequeno, numa entidade como a nossa com balanço grande, obviamente que complica sobremaneira a tarefa."
E2_N_DRE	"Na maior parte dos casos, estamos a falar de volumetrias bastante elevadas, o que significa também ter competência e capacidade de desenvolver trabalho com ferramentas tecnologicamente evoluídas."
Visão de Grupo	
E4_T_CDO	"A necessidade de isto ser feito a nível do grupo, a conformidade tem de ser garantida a nível de grupo, então depois há modelos de relação com as entidades do grupo, (…), mas depois a distância geográfica, a autonomia de acionistas, etc., levanta aqui uma série de questões e de temas sobre o que é que deve ser delegado ou o que é que não deve ser delegado."
Uniformização de Conceitos	
E2_N_DRE	"No fundo, nós temos que alinhar pela mesma bitola, ou seja, definir os conceitos e perceber efetivamente quais é que são comuns, quais é que não são comuns, os que são comuns, defini-los de uma forma inequívoca, e os que não são comuns, temos que acrescentá-los ao nosso dicionário de dados e perceber como é que vamos orientá-los."
Data Ownership	
E2_N_DRE	"Isto por vezes não é fácil, atribuir o *data owner*, dizer quem é que é, se sou eu, é o colega do lado, é o outro... Porque é que sou eu? Porque é que é que o outro? Isto não é fácil, porque obviamente isto significa responsabilidade."

E4_T_CDO	"O assumir o *ownership* de um dado que traz todas estas responsabilidades, não é de ânimo leve, portanto é uma discussão, (…), acho que a discussão que se coloca neste momento é: "porque é que sou eu o *owner* do dado?""
Novas Skills	
E2_N_DRE	"Isto também obriga a uma outra coisa que é a ter competências que se calhar não tínhamos que ter até agora, ou seja, quem trata os dados (…). As empresas também têm que ter pessoas, recursos, colaboradores que tenham competências que permitam tratar volumetrias, explorar a informação, se calhar de diversos ângulos, utilizar ferramentas que se calhar nós não usamos hoje em dia, capacidade de construir *dashboards* e informação para outros consumirem de forma mais imediata e também que não tenham essas competências."
E3_T_CAO	"Ok, isso é muito bonito, mas de que serve dares um carro uma pessoa, se a pessoa não sabe conduzir?"

Fonte: Elaborado pelo próprio autor (2021).

Tabela D6 – Comentários Verbatim: BCBS 239 - Princípios

Entrevistado	Citação Exemplo
1. Governance	
E1_N_CRE	"Cada um sabe exatamente aquilo que pode extrair ou não pode extrair da fonte "
E2_N_DRE	"O mais importante acho que é aquilo que referi, que é centralização, *standardização*, documentação ... e, obviamente, também controlos de qualidade uniformes, processos de correção uniformes."
E3_T_CAO	"Teres os dados governados e com a qualidade que é merecida para uma organização."
E4_T_CDO	"É conhecer os dados e saber a qualidade que os dados têm."
E5_T_CRI	"Inexistência de um dicionário de dados comum."
2. Arquitetura de Dados e Infraestrutura de TI	
E1_N_CRE	"O desafio está em interpretares corretamente o atributo."
E2_N_DRE	"A uniformidade de conceitos, que é saber que aquele conceito é o único para toda a organização."
E3_T_CAO	"Toda a gestão, seja ao nível de metadados, ao nível do valor do dado, ao nível da definição, ao nível da utilização, da finalidade, dos requisitos, *etc.*, tem que ser definido ao nível do *owner* do dado é que tem de definir isso, porque ele é que tem essa responsabilidade no final do dia."
E4_T_CDO	"A linhagem permite também, por exemplo, nós queremos garantir que, e isto é uma das questões básicas do *golden source*, é que o dado é produzido uma vez e usado para todos os fins para se que destina."
E5_T_CRI	"Criar uma cultura de dados: identifica *ownership*, identificação única dos atributos dos dados, um processo de controlo centralizado no *Data Lake*."
3. Exatidão e Integridade	
E1_N_CRE	"Haja uma reconciliação contabilística, com os balanços das entidades e da própria sede também, e em que haja um controlo, vários controlos de qualidade da informação que reside nesse repositório."
E2_N_DRE	"O que está previsto é termos um único ponto de reconciliação contabilística."
E3_T_CAO	"Rastreabilidade é tu saberes que a data de nascimento vem do sistema x, portanto tu puxas esta informação do sistema operacional x e pões no *Data Lake*."
E4_T_CDO	"É relevante também termos a linhagem, saber como é que os dados são construídos a partir de outros dados, não só para despistar ou depurar problemas que possam ocorrer, mas também é para perceber como é que os dados são construídos."
E5_T_CRI	"Rastreabilidade dos dados até aos sistemas operacionais."

4. Completude	
E1_N_CRE	"Nós caminhamos claramente para esse nível de granularidade, ou seja, dar a informação ou reportar a informação contrato a contrato, operação a operação."
E2_N_DRE	"Se eu tiver nível mais baixo de detalhe, depois, no fundo, posso agregar de acordo com as necessidades que vou precisar para cada um dos reportes."
E4_T_CDO	"A necessidade de isto ser feito a nível do grupo, a conformidade tem de ser garantida a nível de grupo, então depois há modelos de relação com as entidades do grupo, em que o CDO pode ter uma função mais presente ou menos presente, mas depois a distância geográfica, a autonomia de acionistas, *etc.*, levanta aqui uma série de questões e de temas sobre o que é que deve ser delegado ou o que é que não deve ser delegado."
E5_T_CRI	"Integração de processos no *Data Lake*."
5. Tempestividade	
E2_N_DRE	"Conseguiríamos fazer o mesmo de uma forma mais eficiente? Com menos esforço? Com melhor qualidade? Também diria que sim, mas em termos de, se estamos a entregar no tempo, acho que sim, e com qualidade, penso eu."
E3_T_CAO	"Temos compromissos regulatórios diariamente, semanalmente, mensalmente. Eu acho que isso é o dia-a-dia do banco e nós temos que ter essa visão e, portanto, eu acho que o *Data Lake* vai conseguir responder a isso."
E4_T_CDO	"Qualidade é ser tempestivo."
E5_T_CRI	"Tornar toda a componente regulatória mais eficiente e automatizada."
6. Adaptabilidade	
E1_N_CRE	"As extrações serem efetivamente assertivas naquilo que o Banco, o Banco Central Europeu, espera receber."
E2_N_DRE	"Tentar reutilizar informação que recebemos o mais possível."
E4_T_CDO	"Muito mais fácil quando a instituição precisar (...) de produzir um relatório externo."
7. Exatidão	
E1_N_CRE	"Conseguir justificar que os dados estão bem desde o *front office* até ao *back office*."
E2_N_DRE	"Muitas vezes, gastamos algum tempo, para não dizer bastante, a justificar porque é que o que reportamos nas Estatísticas Monetárias e Financeiras não está, não é o que reportamos para a Central de Responsabilidades de Crédito."
9. Clareza e Utilidade	
E3_T_CAO	"É tu garantires que o dado é relevante, eu quando apresento uma coisa a ti, Mariana, tem que ser algo para ti tenho um sentido, (…), e tu consigas utilizar isso no processo de tomada de decisão."
E4_T_CDO	"Muito mais fácil quando a instituição precisar de tomar decisão."
10. Frequência	
E1_N_CRE	"Até hoje, até hoje, nós temos conseguido sempre dar resposta a tudo."
11. Distribuição	
E2_N_DRE	"Genericamente tem-se conseguido atingir os objetivos a que nos propomos."
12. Revisão	
E4_T_CDO	"Recordo-me de relatórios que ao fim de cinco/seis anos, o número de bancos G-SIBS na Europa que estavam em conformidade com o BCBS 239 eram um, dois, e, mais do que isso."

Fonte: Elaborado pelo próprio autor (2021).

Tabela D7 – Comentários Verbatim: BCBS 239 - Oportunidades Futuras

Entrevistado	Citação Exemplo
Estruturas dos Futuros Reportes	
E1_N_CRE	"Vai começar a acontecer em 2022, a conglomeração de reportes, ou seja, nós vamos ter reportes aumentados, alguns reportes a desaparecer, mas o nível de informação não desaparece, o que desaparece é o reporte porque a informação que é exigida ou o conteúdo da informação, ele passa de um reporte para o outro, portanto nós continuaremos sempre a precisar da informação, o que não existe é que tanto reportes."
E2_N_DRE	"Eu acredito que num futuro, não estou a dizer que seja num ano, nem dois, nem três, mas acredito que num futuro não muito longínquo, cada vez mais os reguladores não nos vão pedir agregações, mas vão-nos pedir os detalhes de todas as operações e vão ser eles a agregar aquilo da forma como acharem que devem agregar."
Exploração de dados	
E3_T_CAO	"Uma das coisas que eu também gostava de implementar, era ter pessoas que fizessem uma espécie de um estágio na minha equipa, e que aprendessem a utilizar essas práticas e depois conseguissem ser o *Steward*, o *Data Steward*."
E4_T_CDO	"Nós temos a ambição de tornar isto também muito numa lógica *self-service* dos próprios dados."
Experiência ao Cliente	
E3_T_CAO	"E, nós, com base nos dados que vão estar no *Data Lake*, se conseguirmos desenvolver algoritmia que consiga responder a isto, vai estar num nível de maturidade, que a <organização> ainda não está, nunca esteve e tem a oportunidade de estar, e, aí sim, providenciares um serviço à séria ao teu cliente e teres o cliente no centro de tudo o que nós fazemos."
E4_T_CDO	"Neste momento, se calhar, com base em termos mais dados, termos os dados que temos, cada vez mais podemos ter um modelo provavelmente suportado em Inteligência Artificial, que vai identificar melhor quais são os clientes que podem ser alvo de uma determinada oferta. E aí, as *leads*, em vez de termos listas e listas, não, temos uma lista muito mais reduzida, que provavelmente até vai estar integrado numa aplicação e que podem interagir diretamente com o cliente numa *app* ou então suportar um colega nosso que esteja numa agência, ou que esteja num *contact center* a falar com o cliente e sabe que aquele cliente existe uma propensão muito grande para ele ter interesse naquele produto. Com isto, os nossos clientes vão ficar muito mais fidelizados, vão estar muito mais interessados em nós, em ouvir-nos, em ouvir as nossas ofertas e, provavelmente a adquirir os nossos produtos e serviços."

Fonte: Elaborado pelo próprio autor (2021).

Tabela D8 – Comentários Verbatim: BCBS 239 - Desafios

Entrevistado	Citação Exemplo
Processo Complexo	
E2_N_DRE	"Claro que se eu tiver tudo ajustado, toda esta documentação, toda esta informação num sistema tecnológico que depois não responda, também não funciona."
E3_T_CAO	"Muitas vezes, tu tens é de conseguir, às vezes tu sacrificas a qualidade do teu output analítico, para que, de facto, consiga ser utilizado."
E4_T_CDO	"O tema é bastante complexo, os princípios são abstratos, não são concretos e a sua interpretação também requer algum conhecimento."
Correta Interpretação	
E1_N_CRE	"Se estamos a preenchê-los bem (os atributos), em conformidade com aquilo que é pedido, para depois (…) as extrações serem efetivamente assertivas naquilo que o Banco, o Banco Central Europeu, espera receber."

E4_T_CDO	"O tema é bastante complexo, os princípios são abstratos, não são concretos e a sua interpretação também requer algum conhecimento."
Inputs Manuais	
E4_T_CDO	"As *end-user-computing* são aplicações desenvolvidas por pessoas não programadoras ou fora do IT, muitas vezes, excel, access ou até cálculo manual ou em papel, em que os dados depois são inseridos nos processos de agregação de dados, para produzir o relatório, *etc*. Isto precisa de estar identificado."
Manutenção	
E2_N_DRE	"Agora, que nós temos de ter os príncipios do BCBS 239, bem implementados, implementados, em primeiro lugar. Segundo, bem implementados e bem mantidos, depois não basta fazer uma vez, tem que se garantir que continua a acontecer, também é muito difícil, embora seja mais fácil o que fazer a primeira vez."

Fonte: Elaborado pelo próprio autor (2021).