Modelo de Data Quality nas Estruturas da MC

Joana Irene Alves Coelho
Mestrado em Engenharia Matemática
Departamento de Matemática
2023

Orientadora
Maria João Pinto Sampaio Rodrigues, Professora Auxiliar, Faculdade de Ciências da Universidade do Porto


U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Agradecimentos

Foram diversos os intervenientes que contribuíram para a realização deste trabalho, que, de forma direta ou indireta, permitiram a concretização desta etapa final do mestrado em Engenharia Matemática na Faculdade de Ciências da Universidade do Porto (FCUP).

Início por agradecer à minha orientadora, Doutora Maria João Rodrigues, pelas contribuições significativas, apoio incansável e disponibilidade constante para orientação.

Às supervisoras, Daniela Rocha e Ana Sofia Duarte, expresso o meu agradecimento pelo tempo generosamente dedicado, disponibilidade constante e partilha de conhecimento. Saliento também a contribuição relevante que deram ao longo do tempo para o meu desenvolvimento, tanto a nível profissional como pessoal.

Ao meu namorado Jorge, quero manifestar a minha sincera gratidão pelo apoio inabalável e incessante, pela motivação constante e pela infinita paciência. Agradeço por teres desempenhado um papel fundamental na realização deste sonho.

A minha mãe e irmã merecem o meu agradecimento pelo apoio constante, pelos conselhos e pelo incentivo que sempre me encorajaram a ir além dos meus limites. Ao meu avô Manuel, que é a minha referência e fonte de inspiração, expresso a minha sincera gratidão.

Ao meu amigo Paulo Moreira, que acompanhou a minha jornada académica, expresso o meu sincero agradecimento pela orientação e apoio que ofereceu ao longo de todo o percurso. A tua presença foi extremamente gratificante nesta etapa.

Aos meus amigos, que estiveram sempre presentes e me deram força nos momentos cruciais, expresso o meu reconhecimento pelo apoio inestimável.

À minha equipa na Sonae, Supply Chain em BI Delivery, quero expressar a minha gratidão pela dedicação incansável, partilha de conhecimento e pelo apoio durante o meu estágio. A vossa colaboração tornou a minha integração na empresa muito mais fácil.

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto
Departamento de Matemática

Master's Degree in Mathematical Engineering

Modelo de Data Quality nas Estruturas da MC

by Joana COELHO

Companies have large volumes of data, usually referred to as Big Data, coming from different sources and in multiple formats, and if properly analyzed, this can be a great advantage for the company. To be more competitive, companies need to exploit all the richness that is present in their data. Good data quality allows organizations to take advantage of more market opportunities and be more competitive.

The inexistence of information on data quality is a significant obstacle to detecting and rectifying potential flaws that could compromise the integrity of the data in question. At MC Sonae, there was a pressing need to develop a data model that would provide detailed information on data quality through a dashboard. This approach gives users access to KPIs that quantify the quality of the data, while also allowing them to assess its reliability. When poor data quality is detected, it is an indicator of possible errors in the databases and enables timely corrective interventions.

During the internship I developed a single centralized data model to generate a final dashboard, as well as building a predictive model to forecast future data quality.

Keywords: **Palavras-chave**: Data Quality, Data Governance, Business Intelligence, Data Warehouse, Dashboard, Time Series.

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto
Departamento de Matemática

Mestrado em Engenharia Matemática

Modelo de Data Quality nas Estruturas da MC

por Joana COELHO

As empresas possuem grandes volumes de dados, comumente referidos como *Big Data*, provenientes de diversas fontes e em múltiplos formatos, e estes se forem devidamente analisados podem constituir uma grande vantagem para a empresa. Para serem mais competitivas, as empresas necessitam de explorar toda a riqueza que está presente nos seus dados. Uma boa qualidade de dados permite que as organizações possam aproveitar mais oportunidades de mercado e serem mais competitivas.

A ausência de informações relativas à qualidade dos dados constitui um obstáculo significativo na deteção e retificação de potenciais falhas que possam prejudicar a integridade dos dados em questão. Na MC Sonae, surgiu a necessidade premente de desenvolver um modelo de dados que disponibilizasse informações detalhadas sobre a qualidade destes através de um *dashboard*. Tal abordagem possibilita os utilizadores acederem aos KPIs que quantificam a qualidade dos dados, permitindo-lhes também avaliar a sua fiabilidade. Quando se deteta uma baixa qualidade dos dados, este facto é um indicador de eventuais erros nas bases de dados e possibilita intervenções corretivas oportunas.

Durante a realização do estágio procedeu-se ao desenvolvimento de um modelo único centralizado de dados para gerar um *dashboard* final, bem como construir um modelo preditivo para prever a qualidade futura dos dados.

Palavras-chave: Data Quality, Data Governance, Business Intelligence, Data Warehouse, Dashboard, Séries Temporais.

Conteúdo

Lista de Abreviaturas

BI Inteligência de negócios (do inglês *Business Intelligence*)

DQ Qualidade de dados (do inglês *Data Quality*)

DG Governo de dados (do inglês *Data Governance*)

DW Armazéns de dados (do inglês *Data Warehouse*)

BD Base de dados

SQL Linguagem de consulta estruturada (do inglês *Structured Query Language*)

KPI Indicador chave de desempenho (do inglês *Key Performance Indicator*)

ETL Extrair, transformar e carregar (do inglês *Extraction, Trnasformation and Loading*)

CQ Controlo de qualidade

RAID (do inglês *RAID - Revenue Assurance Integration Driller*)

DS (do inglês *IBM Infosphere DataStage*)

BIT (do inglês *Business Information Technology*)

ACF Função autocorrelação (do inglês *Autocorrelation Function*)

PACF Função autocorrelação parcial (do inglês *Partial Autocorrelation Function*)

ADF (do inglês *Augmented Dickey Fuller*)

KPSS (do inglês *Kwiatkowski-Phillips-Schmidt-Shin*)

IQR Intervalo interquartil (do inglês *Interquartile Range*)

Q1 Primeiro quartil

Q3 Terceiro quartil

AR AutoRegressiva (do inglês *AutoRegressive*)

MA Médias móveis (do inglês *Moving Average*)

ARIMA (do inglês *AutoRegressive Integrated Moving Average*)

RMSE Erro quadrático médio (do inglês *Root Mean Squared Error*)

MAPE Erro percentual absoluto médio (do inglês *Mean Absolute Percentage Error*)

AIC Critério Akaike Information (do inglês *Akaike Information Criterion*)

BIC Critério Bayesian information (do inglês *Bayesian Information Criterion*)

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Neste capítulo inicial é introduzido o tema deste projeto de relatório de estágio, efetuando uma contextualização e articulando com as motivações envolventes. Numa segunda parte é descrito o perfil corporativo da organização onde realizei o meu estágo.

1.1 Contextualização e motivação do relatório de estágio

De acordo com Suer (2021), a qualidade dos dados, *Data Quality* (DQ), é definida como o grau em que estes satisfazem as expectativas de uma empresa em termos de exatidão, validade, completude e consistência. [1]

Quando analisa os seus dados, uma empresa pode identificar potenciais problemas que prejudiquem a qualidade dos mesmos, e assegurar que os dados partilhados estejam aptos a ser utilizados para um determinado fim. Quando os dados recolhidos não satisfazem as expectativas da empresa quanto à sua exatidão, validade, completude ou consistência, é de inferir a ocorrência de impactos negativos de dimensão considerável no serviço ao cliente, bem como na produtividade dos funcionários ou ainda na execução das estratégias chave. Os dados são classificados de alta qualidade quando representam de forma fiel, completa e consistente a realidade que se propõem a descrever.

Sem uma boa qualidade de dados, as empresas terão dificuldade em fazer uso de novas oportunidades de mercado. [1] Consequentemente, a baixa qualidade dos dados pode proporcionar uma diminuição acentuada das margens de lucro da organização, bem como dificultar a sua trajetória de crescimento. Por sua vez, as empresas com dados de baixa

qualidade geralmente não são capazes de introduzir medidas de redução de custos, dado que a falta de dados de boa qualidade exige ainda a necessidade de se realizarem mais inspeções e correções manuais. Sem a presença de dados completos e consistentes, é difícil concretizar a automatização dos processos associados à empresa.

Os dados devem atender a um conjunto de requisitos que os tornam mais apropriados. Esses requisitos incidem fundamentalmente na sua conformidade, na segurança, na proteção ou na privacidade. Para que uma organização possa realizar análises preditivas sobre os seus ativos, é essencial que os dados utilizados cumpram o número máximo de requisitos apresentados. Caso algum requisito não seja cumprido, é frequente que surjam dificuldades em definir decisões ótimas ou viáveis. Desta forma, as decisões organizacionais ficam comprometidas, resultando num progresso empresarial mais árduo e menos eficiente. Os principais desafios que atualmente comprometem a qualidade dos dados de uma empresa passam pela sua duplicação, incompletude, inconsistência ou ainda na imprecisão preditiva.

Outro dos desafios que as organizações enfrentam é o desenvolvimento de mecanismos de gestão que estabeleçam um equilíbrio entre o risco e os benefícios, face à quantidade de dados que cresce gradualmente e às inovações tecnológicas que oferecem um armazenamento melhor, mais rápido e mais barato. Os programas e iniciativas de *Data Governance* (DG) são empreendidos por organizações com o objetivo de aumentar a receita e rentabilidade, aumentando assim o valor dos seus produtos e serviços, bem como uma melhor tomada de decisões. [2]

Por fim, relativamente à segurança e privacidade dos dados, as organizações enfrentam desafios cada vez mais complexos na proteção desses dados contra roubos, na utilização sem permissão ou ainda na divulgação não autorizada. Assim, é vantajoso a implementação de um programa de DG com características de segurança e privacidade, com o objetivo de identificar as ameaças e abordar os riscos residuais de maneira eficiente e eficaz. [3]

Dado todo o contexto apresentado, a motivação deste relatório de estágio decorre da

necessidade de desenvolver um *dashboard* à qual a ingestão de dados ocorra automaticamente com informações relacionadas à qualidade dos dados dentro da organização. Isso inclui tanto os dados em si quanto as métricas, levando em consideração as necessidades dos colaboradores internos da empresa. Esses colaboradores compreendem os analistas (equipa de negócios), *developers* e a equipa de suporte (BIT). O objetivo principal é fornecer informações sobre a qualidade dos dados que estão a utilizar nas análises e tomadas de decisão.

A solução envolve o planeamento, desenvolvimento e implementação de um modelo centralizado em *Hadoop* responsável pela ingestão diária de novos dados no *dashboard* de controlo. Os processos de extração, transformação e carregamento (ETL) serão realizados utilizando o ambiente *IBM Infosphere DataStage* (DS).

O primeiro passo consiste em colaborar com as equipas envolvidas para identificar os tipos de controlo de qualidade existentes e determinar em quais bases de dados esses controlos estão registados.

O segundo passo é desenvolver um *dashboard* que represente os dados e os indicadores de qualidade de dados. Isso permitirá que os usuários finais visualizem de modo intuitivo a qualidade dos dados em uso.

Para finalizar, serão realizadas previsões associadas à qualidade de dados, utilizando dados de um controlo de qualidade.

1.2 Contexto empresarial

A **MC Sonae** é uma empresa portuguesa de retalho e distribuição fundada em 1985, que dirige quinze das principais marcas nacionais, constituindo assim uma das principais atividades do grupo *Sonae* (visualmente ilustrado na Figura 1.1). É uma das maiores empresas de retalho em Portugal, sendo líder no retalho alimentar, e a sua atividade teve origem com a abertura do primeiro hipermercado em Portugal .

É composta pelos Continente e Continente Online (hipermercados), Continente Modelo e Continente Bom dia (supermercados de conveniência), Meu Super (lojas de proximidade

em formato franchising), Bagga (cafetaria e restaurante), Note! (livraria/papelaria), Wells, Dr. Wells, Go Natural e Arenal (saúde, bem estar e ótica), ZU (produtos e serviços para animais de estimação) e outros negócios em crescimento como Washy (lavandarias self-service) e Home Story (decoração e artigos para o lar).

A **BIT** (*Business Information Technology*) é a área de sistemas de informação da *MC Sonae* responsável por garantir que a empresa tem as melhores tecnologias e ferramentas para gerir e analisar dados, e ainda capacidades de *Business Intelligence (BI)*[*] para ajudar a empresa a tomar decisões informadas e alcançar os seus objetivos. Detém também a responsabilidade em fornecer suporte aos negócios da empresa, desde a recolha de dados até à tomada de decisões. Neste domínio estão incluídos a gestão de dados, a construção de sistemas de informação, a análise de dados e a criação de relatórios e *dashboards*. Deve-se ainda ressalvar que a BIT responsabiliza-se por garantir a segurança e a privacidade dos dados da empresa, assegurando a sua proteção contra ameaças externas ou internas. Procura ainda garantir que a empresa se encontra em conformidade com as regulamentações de proteção de dados aplicáveis. A BIT é composta por aproximadamente 600 colaboradores (internos e parceiros) com sede localizada no Lugar do Espido, Via Norte – Maia. A Figura 1.1 ilustra o perfil corporativo do grupo *Sonae* e as lojas da *MC Sonae*.

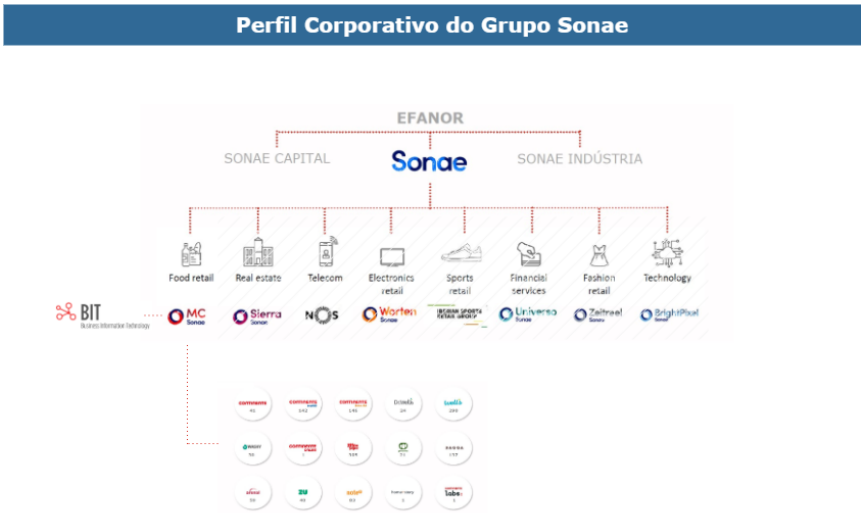

FIGURA 1.1: Perfil corporativo do grupo *Sonae*

[*]Business Intelligence (BI) é um conjunto de tecnologias, processos e ferramentas que ajudam as empresas a reunir, integrar, analisar e apresentar informações úteis para apoiar a tomada de decisões estratégicas.

Capítulo 2

Fundamentos

Neste capítulo, efetua-se uma revisão da literatura que abrange informações concernentes à qualidade de dados, além da exposição das ferramentas utilizadas e da arquitetura organizacional. Adicionalmente, será abordada a análise de séries temporais, incluindo a explicação dos conceitos fundamentais, bem como a apresentação de um modelo matemático aplicado para previsões. Este capítulo desempenhará um papel fundamental como alicerce teórico e metodológico para a formulação do modelo de solução proposto.

2.1 Dados

De um modo geral, considera-se que dados são informações que representam factos sobre algo em relação a um determinado universo. No contexto de tecnologia da informação, o entendimento é mais amplo quando se trata de dados, incluindo qualquer característica, atributo ou facto armazenado eletronicamente. [4]

De acordo com Earley (2017) [4], uma base de dados é um conjunto de dados logicamente significativos, ou seja, um grupo de informações organizadas e relacionadas entre si, que são armazenadas eletronicamente num computador ou em outro dispositivo de armazenamento de dados. Esses dados podem incluir informações sobre pessoas, produtos, transações, eventos, entre outros.

2.1.1 Tipos de dados

Os dados são uma fonte valiosa de percepção das necessidades dos clientes, opiniões e tendências de mercado para as empresas. Eles estão disponíveis em grandes quantidades,

provenientes de diferentes fontes e possuem características distintas. Nesse contexto, de acordo com Halper e Krishnan (2013) [5], é possível distinguir três tipos diferentes de dados:

- **Dados estruturados**: dados com uma estrutura rígida e previamente conhecida, na qual todos os elementos partilham o mesmo formato e tamanho, este é o tipo de dados, tradicionalmente encontrados em aplicações empresariais, que têm sido armazenados em base de dados relacionais;
- **Dados semi-estruturados**: dados com alto grau de heterogeneidade, que não são facilmente representados em estruturas de dados fixas. Normalmente, este tipo de dados tem sido armazenado usando linguagens específicas, tais como dados XML (*Extensible Markup Language*), dados RDF (*Resource Description Framework*), JSON (*JavaScript Object Notation*), entre outros;
- **Dados não estruturados**: dados sem estrutura, tais como texto, vídeos ou conteúdo multimédia. Este tipo de dados cresceu exponencialmente durante os últimos anos, e estima-se que atualmente mais de 80% dos dados gerados são dados não estruturados. Alguns exemplos incluem documentos, imagens, fotos, mensagens, entre outros.

2.1.2 Características dos dados

Conforme definido por *Earley* [4], as características dos dados podem variar entre organizações e até mesmo dentro de uma mesma organização. Por essa razão, os dados têm ciclos de vida diferentes, o que se traduz em requisitos de gestão diferenciados. Os dados podem ser classificados quanto aos seguintes aspetos:

- Tipos de dados: dados transacionais, dados de referência, dados mestre e metadados;
- Conteúdo: domínio de dados e áreas subjetivas;
- Formato;
- Nível de proteção;
- Em que local e de que forma os dados são armazenados, disponibilizados e acedidos.

Ao considerar as diversas funções que os dados desempenham numa organização, é de suma importância reconhecer que a qualidade dos dados terá efeitos variados nos utilizadores desses dados. Como ilustração, numa análise de desempenho de vendas, a falta de completude do nome de um produto (dados mestres) pode não ser tão crucial quanto a ausência de informações completas sobre as vendas do produto ou a presença de valores distorcidos (dados transacionais), os quais acarretarão consequências relevantes na análise. Como resultado, todo sistema de gerenciamento de dados deve levar em consideração que os dados possuem atributos distintos e seguem ciclos de vida diversos. [4]

Na área de retalho, é possível categorizar os dados com base em diferentes critérios, como tipo de informação, conteúdo, formato, grau de segurança e os métodos e locais de armazenamento e acesso. No entanto, o âmbito deste relatório de estágio irá incidir apenas sobre os seguintes tipos de dados:

- Dados transacionais: dados gerados através da atividade operacional do negócio;
- Dados de referência: dados para classificar ou categorizar outros dados;
- Dados mestre: dados acerca das entidades do negócio que fornecem contexto a outros dados.

2.1.3 Metadados

Metadados constituem informações descritivas acerca de outros dados e desempenham um papel fundamental na pesquisa e recuperação de informações, bem como na salvaguarda da qualidade e integridade dos dados subjacentes. Estas informações descritivas podem abranger aspetos como o formato dos dados, a identificação do autor, a data de criação, as dimensões do arquivo e outros dados pertinentes. [6]

A relevância dos metadados está intrinsecamente ligada à sua capacidade de aprimorar a precisão e eficiência do processo de recuperação de informações. Ao oferecer informações descritivas sobre os dados subjacentes, os metadados agilizam a sua localização tornando este processo mais expedito e simples. Adicionalmente, os metadados desempenham um papel crucial na garantia da qualidade e integridade dos mesmos, permitindo aos

utilizadores avaliar a fiabilidade das informações e identificar eventuais problemas de qualidade.

2.1.4 *Data Warehouse*

No contexto de *BI*, *Data Warehouse* (DW) é um tipo de armazém de dados projetado para armazenar e gerenciar grandes quantidades de instâncias provenientes de várias fontes e sistemas de uma organização, a fim de fornecer informações valiosas para análise e tomada de decisão. De acordo com Kimball et al. (2013) [7], o DW é um componente-chave da arquitetura de BI e é projetado para apoiar as necessidades de análise de negócios em larga escala. É projetado para suportar processos de extração, transformação e carregamento (*ETL*) de dados, bem como para fornecer um modelo de dados dimensionais otimizado para consulta e análise eficiente. Em outras palavras, o DW é um repositório centralizado de dados que fornece uma visão única e consistente dos dados da organização para apoiar a tomada de decisão baseada em dados.

Um Data Warehouse inclui, frequentemente, os seguintes elementos: [8]

- Uma base de dados relacional* para armazenar e gerir dados;
- Uma solução de ETL para preparar os dados para análise;
- Análise estatística, relatórios e capacidades de extração de dados;
- Ferramentas de análise para clientes para a visualização e apresentação de dados a utilizadores empresariais;
- Outras aplicações analíticas mais sofisticadas que geram informação acionável através da aplicação de algoritmos de ciência de dados e de inteligência artificial (IA), bem como características gráficas e espaciais que permitem tipologias distintas de análise de dados à escala.

Qualquer desenho de um DW deve incluir o conteúdo específico dos dados, bem como as relações inerentes a um grupo de dados ou entre grupos distintos. Também é possível incluir o ambiente do sistema que dá suporte ao DW, os tipos de transformações

*Trata-se de uma base de dados que armazena e fornece um ponto de acesso aos dados entre si. Desta forma, estes últimos são interligados através de tabelas.

necessárias ou ainda a frequência de atualização de dados.

Um fator primário na conceção são as necessidades e finalidades dos utilizadores finais, uma vez que a sua maioria geralmente está interessada em efetuar análises e visualizar os dados de forma agregada, em vez de os considerar como transações individuais. No entanto, em muitos casos os utilizadores finais não sabem realmente o que querem até surgir uma necessidade específica. Assim, o processo de planeamento deve incluir exploração suficiente para antecipar as necessidades. Finalmente, a conceção do DW deve permitir espaço para que este possa expandir-se, de modo a acompanhar as necessidades e tendências dos utilizadores finais.

Emerge uma distinção essencial entre dois tipos de estruturas: os DW e as bases de dados (BD). Embora ambos desempenhem um papel de primordial relevância na preservação de informações, as suas finalidades, arquiteturas e utilizações diferem de forma significativa. A fim de aprofundar a compreensão das discrepâncias entre um DW e uma BD, é apresentada a Tabela 2.1 que destaca as principais divergências entre estas, fornecendo uma visão clara e estruturada de suas características distintivas. Esta tabela serve como uma referência concisa para avaliar as capacidades e limitações de cada estrutura. [9]

Propriedade	**BD**	**DW**
Utilização	Registo de dados - permitem aos utilizadores introduzir informações em campos separados para referência e utilização numa aplicação.	Análise de dados - permitem que os utilizadores trabalhem com os dados em tempo real.
Método de processamento	Processamento de transações em linha (OLTP) - gere dados agregados para utilização transacional.	Processamento analítico em linha (OLAP) - gere dados armazenados para consultas e revisões analíticas.

Tipo de dados	Os dados em tempo real utilizam um feed em direto para referência do utilizador final na realização de atividades diárias.	Os dados históricos aproveitam as ferramentas que os analistas utilizam para relatórios mais longos e BI.
Armazenamento	Limitado com base nos requisitos de uma única aplicação.	Escalável, dependendo das necessidades de dados entre aplicações e sistemas.
Restrições	Limitado a uma base de dados por relação de aplicação.	Suporte ilimitado entre bases de dados e aplicações.

TABELA 2.1: Resumo das principais propriedades de DB e DW

Conforme ilustrado na Tabela 2.1, as BD e os DW exibem vantagens distintas e aplicações singulares.

2.1.5 *Extraction, Transformation and Loading*

Extraction, Transformation and Loading (ETL) – que significa extrair, transformar e carregar – é um processo de integração de dados que os combina, mediante as múltiplas fontes que podem ter, num único e consistente, que por sua vez é carregado num *Data Warehouse* ou num outro sistema. [10]

O ETL fornece a base para a análise de dados e para a aprendizagem de máquinas. Através de uma série de regras empresariais, o ETL limpa e organiza os dados de uma forma que responde às necessidades específicas de inteligência empresarial, como os relatórios mensais, mas também pode abordar análises mais avançadas, que podem melhorar os processos de *back-end* ou as experiências do utilizador final. Assim, um sistema ETL devidamente concebido extrai dados dos sistemas-fonte, reforça as normas de qualidade e de consistência dos dados, conforma os dados para que fontes separadas possam ser utilizadas em conjunto e, finalmente, entrega os dados num formato para que os utilizadores finais possam tomar decisões.

Especificamente, o sistema ETL remove erros, corrige dados em falta, fornece medidas documentadas de confiança nos dados, captura o fluxo de dados transacionais para manter em segurança e ajusta dados de múltiplas fontes para que posteriormente sejam utilizados em conjunto pelo utilizador final.

2.1.6 Indicadores Chave de Desempenho

Os Indicadores Chave de Desempenho, também referidos como KPIs, desempenham um papel crucial ao auxiliar as organizações na definição e medição do progresso em direção aos seus objetivos organizacionais. Os KPIs têm a função de simplificar a complexidade inerente ao desempenho organizacional, condensando-a num conjunto limitado de indicadores-chave, tornando assim esse desempenho mais compreensível. [11]

Ao selecionar os KPIs, é fundamental limitá-los aos fatores essenciais para que a organização alcance os seus objetivos. Assim, os KPIs transparecem uma imagem clara do que é importante e do objetivo final.

Existem dois tipos de indicadores: **quantitativos** e **qualitativos**. Os indicadores quantitativos são medidas numéricas que podem ser mensuradas e quantificadas (número, média, mediana ou percentagem) enquanto que os indicadores qualitativos são medidas subjetivas que não podem ser facilmente quantificadas, mas podem ser avaliadas através de observação ou feedback (por exemplo, a satisfação do cliente, a qualidade do produto).

Os KPIs SMART constituem uma metodologia para definir indicadores de desempenho que apresentem características de especificidade, mensurabilidade, alcance, relevância e temporalidade. SMART é um acrónimo para: [12]

- **S (Specific)**: O KPI deve ser **específico** e claramente definido, para que todos os envolvidos entendam exatamente o que está a ser medido;
- **M (Measurable)**: O KPI deve ser **mensurável**, de forma que possa ser quantificado e monitorizado ao longo do tempo;
- **A (Attainable)**: O KPI deve ser **alcançável** e realista, levando em consideração os recursos disponíveis e as capacidades da empresa;

- **R (Relevant)**: O KPI deve ser **relevante** e alinhado aos objetivos estratégicos da empresa, para que sua medição seja significativa para o sucesso do negócio;
- **T (Time-bound)**: O KPI deve ser oportuno e ter um **prazo definido** para a sua medição, para que possa ser avaliado regularmente e ajustado, se necessário.

2.2 Data Governance

As organizações estão cada vez mais a confiar nos dados que produzem, o que originou a necessidade organizacional de responsabilidade em relação à qualidade dos dados que utilizam para obter informações. Na atualidade, as organizações tornam-se mais sofisticadas na utilização dos seus dados e estes são inegavelmente um dos seus maiores recursos.

Deste modo, DG é definido como um conjunto de práticas de controlo que ajudam a assegurar a gestão formal dos dados numa organização. [13] DG funciona no fundo como uma autoridade e controlo da gestão de dados. [4] O objetivo de DG é garantir que os dados são geridos de forma adequada, tendo por base boas práticas e políticas que melhor se adequam no contexto da organização, que inclui uma estrutura de decisão para definir de forma apropriada diretrizes, políticas, acordos e prioridades. [13] O foco e particularidades de um programa de DG dependem das necessidades de cada organização.

Em síntese, Data Governance importa-se com os processos, as políticas, os padrões e as tecnologias para gerir e garantir a disponibilidade, acessibilidade, qualidade, consistência, audibilidade e segurança dos dados de uma organização. [14]
Os objetivos subjacentes consistem em:

- Garantir que os dados satisfazem as necessidades da empresa;
- Proteger e gerir os dados como um ativo empresarial valioso;
- Reduzir os custos de gestão de dados.

2.2.1 Dimensões da Qualidade de dados

As dimensões da qualidade dos dados são características que podem ser medidas e utilizadas para quantificar a qualidade dos dados. Embora tenham sido identificadas quinze dimensões para medir a qualidade dos dados, neste trabalho apenas serão consideradas as oito dimensões mais relevantes e utilizadas na empresa: integridade, consistência, unicidade, precisão, completude, conformidade, atualidade e validade. [15]

Deve ser definida uma análise da qualidade dos dados, quando aplicável, para todos os elementos de dados e para cada uma das dimensões da qualidade de dados.

Nesse sentido, na MC duas categorias de regras emergem: as técnicas e as de negócio. As regras técnicas objetivam avaliar a coerência dos dados sob uma perspectiva técnica. As regras de negócio são todas aquelas que foram definidas juntamente com o negócio e que alertam para problemas na coerência e veracidade funcional dos dados.

Torna-se assim relevante fornecer exemplos elucidativos para facilitar a compreensão das distinções entre as duas categorias de regras.

Exemplos de regras técnicas:

- A Tabela TABELA1[*] pode conter exclusivamente os valores I, A, H, PS e U no campo event_type[†] (dimensão: completude);
- Não é admissível a existência de repetições de key_id[‡] no Hadoop na Tabela TABELA2[§] para cada table_name[¶] distinto (dimensão: unicidade);
- A contagem de registos entre a tabela final em FONTE[||] e DESTINO[**] deve manter coerência (dimensão: consistência);
- Entre outros.

[*] Nome original da tabela criptografado
[†] Atributo da TABELA1
[‡] Atributo de TABELA2
[§] Nome original da tabela criptografado
[¶] Atributo de TABELA2
[||] Sistema operacional na fonte dos dados
[**] Sistema operacional no destino dos dados

Exemplos de regras de negógio:

- A idade dos colaboradores deve situar-se acima dos 17 anos e abaixo dos 80 anos;
- No que se refere a horas de ausência e horas teóricas, um alerta é gerado sempre que as horas de ausência superarem as horas teóricas;
- Para a remuneração, a regra estipula a consideração do salário mínimo compatível com o país em questão, assim gera alerta sempre que o salário do colaborador seja inferior ao salário mínimo;
- Entre outros exemplos.

A seguir, é apresentada uma breve descrição das dimensões a considerar no presente relatório:

- **Integridade:** Garante que todos os dados da empresa possam ser rastreados e conectados. Assim, assegura que os atributos sejam mantidos corretamente, mesmo quando os dados são armazenados e utilizados em sistemas diversos.
- **Consistência:** Refere-se à comparação e verificação da exatidão dos dados ao serem copiados ou transferidos entre diferentes fontes. Esta dimensão visa assegurar que os dados sejam replicados de forma precisa e consistente, sem perdas ou alterações indesejadas durante o processo de cópia ou transferência. Não se limita apenas à verificação da igualdade entre os valores dos dados, como também abrange a análise da estrutura, formato e integridade dos dados copiados.
- **Unicidade:** Iindica se existe uma única instância registada no conjunto de dados utilizado. É medida em relação a todos os registos dentro de um conjunto de dados ou entre diferentes conjuntos de dados. Assim, um alto valor de unicidade assegura a minimização de duplicações ou sobreposições, aumentando a confiança nos dados e nas análises realizadas. A garantia de unicidade é essencial para evitar duplicações ou sobreposições.
- **Precisão:** Refere-se à proximidade entre um valor de dados ”X”e um valor de dados ”Y”, considerado como a representação correta do fenómeno do mundo real que o valor de dados ”X”procura representar. Por outras palavras, a precisão mede o quão exato e correto um valor de dados é em relação à realidade que ele representa

desempenhando um papel crucial na qualidade dos dados e na garantia de que as informações sejam confiáveis e úteis para tomadas de decisão.

- **Completude:** Diz respeito à amplitude com que todas as informações relevantes e necessárias estão presentes e devidamente registadas no conjunto de dados em consideração. Em essência, refere-se à garantia de que nenhum dado de relevância seja omisso ou ausente na composição das informações.

- **Conformidade:** Refere-se à adesão dos dados a padrões, normas ou critérios predefinidos. Trata-se da garantia de que os dados estão em consonância com requisitos específicos, regulamentos ou diretrizes estabelecidos.

- **Atualidade:** Consiste no grau em que os dados refletem a situação atual. Por exemplo, se se estiver a avaliar o número de produtos vendidos, os dados do ano passado não seriam considerados dados atualizados. [16]

- **Validade:** Consiste na qualidade intrínseca e na precisão das informações. Dados considerados válidos são aqueles que refletem com exatidão a realidade que representam, estando livres de erros, omissões e inconsistências. A avaliação da validade envolve a verificação da conformidade dos dados com os padrões estabelecidos, bem como a garantia de que as relações e regras de negócio.

A Figura 2.1 representa as dimensões implementadas na organização *MC Sonae*, delineando a descrição individual de cada dimensão (conforme estabelecido pela empresa), juntamente com um exemplo representativo e a regra associada.

A avaliação das dimensões de qualidade dos dados permite identificar oportunidades para aprimorar a sua qualidade. Podem-se utilizar regras, controlos de qualidade (CQ's), dos dados para determinar se estão adequados para uso e identificar áreas que requerem melhorias. Essas regras garantem que os dados representem com precisão, integridade e consistência as entidades do mundo real. A automatização das regras auxilia na identificação rápida de erros nos dados e fornece atualizações contínuas sobre o estado da qualidade dos dados. [17] [18]

Dimensões da Qualidade de Dados

Dimensão	Definição	Exemplo	Regra
Completude	O grau em que os dados necessários são conhecidos. Isto inclui ter os elementos de dados necessários (os factos sobre o objeto ou evento), ter os registos necessários e ter os valores necessários	Estão disponíveis todas as informações necessárias? Ex: Estão disponíveis todos os campos necessários e que deveriam ter sido introduzidos para a classificação do cliente?	Negócio
Conformidade	O grau em que os dados são conhecidos com o nível de pormenor correto (por exemplo, o número correto de casas decimais à direita do ponto decimal)	Existem formatos específicos para determinados tipos de dados? Quais são esses formatos? Estão a ser seguidos? Ex: A regra de formatação do NIF está a ser seguida? Quantos registos não seguem a regra?	Negócio
Consistência	O grau em que os factos redundantes são equivalentes em duas ou mais bases de dados em que os factos são mantidos	Os valores entre estados diferentes (Fases: Ing, Proc;Expl) apresentam resultados diferentes; Os registos de dados em fluxos diferentes geram resultados conflituosos?	Técnica
Atualidade	O grau em que os dados estão disponíveis para apoiar um determinado consumidor ou processo de informação quando necessário	Os dados chegaram? Chegaram a tempo? Ex: Existem SLA's que não estejam a ser cumpridas regularmente?	Técnica
Integridade	A correção com que os dados derivados são calculados a partir dos seus dados de base	Existem relações importantes que não estão a ser preenchidas? Ex: Estou a receber compras fidelizadas sem referência ao cliente?	Negócio
Unicidade	O grau em que não existem ocorrências ou registos redundantes do mesmo objeto ou acontecimento do mundo real	Existem registos duplicados? Quantos registos únicos existem? Ex: Existem clientes ativos com o mesmo número NIF?	Técnica
Validade	O grau em que os dados estão em conformidade com as suas definições, valores de domínio e regras de negócio	Existem registos com valores inválidos? Ex: Existem utilizadores com uma data de nascimento no futuro? Temos lojas fechadas e ativas em simultâneo? A variação entre D1-D2 não deve ser superior a uma determinada %	Negócio
Precisão	O grau em que os dados correspondem à fonte original de dados, como um formulário, aplicação ou outro documento	A informação está correta? Existem erros ortográficos? Ex. Temos Maaaria quando deveria ser Maria?	Negócio

FIGURA 2.1: Dimensões implementadas na MC

2.2.2 Impacto da má qualidade dos dados

As repercussões ocasionadas pela inadequada qualidade dos dados manifestam-se de forma quase quotidiana. Não obstante, frequentemente carece-se de conhecimento quanto à origem subjacente desses impactos. [15]

Os problemas causados pela má qualidade dos dados podem revelar-se em qualquer altura do ciclo de vida dos dados, desde a sua criação até ao seu arquivo. [4] Devido à rápida evolução tecnológica, as organizações têm cada vez mais dependido de dados para fundamentar as suas decisões. Conforme mencionado por Heinrich (2018) [14], num inquérito realizado, 83% dos participantes afirmaram que a má qualidade dos dados

impactou os objetivos dos seus negócios, e 66% relataram que a baixa qualidade dos dados teve efeitos negativos nas suas organizações nos últimos 12 meses. Notavelmente, a Forbes [19] divulgou que, segundo o "2016 Global CEO Outlook"da KPMG, 84% dos CEOs demonstraram preocupação em relação à qualidade dos dados que embasam as suas decisões.

IBM divulgou em 2018 [20] algumas estatísticas relacionadas com os dados, enquadradas nos chamados 4 V's (Volume*, Velocidade †, Veracidade‡ e Variedade§). Observando especificamente os números relativos à veracidade, estima-se que a economia dos Estados Unidos da América tenha incorrido em custos da ordem de 3,1 triliões de dólares devido à má qualidade dos dados. Divulgou ainda que um em cada 3 líderes empresariais não confia nas informações que sustentam as suas decisões.

Os efeitos resultantes da má qualidade dos dados podem incidir em várias dimensões de uma organização, conforme observado por: [21]

- Impactos na confiança ou na satisfação (exemplos incluem a satisfação de clientes, colaboradores ou fornecedores);
- Impactos na produtividade;
- Impactos financeiros.

Este relatório centra-se na problemática da ausência da DQ no setor do retalho. Diversos fatores comprometem a qualidade dos dados, nomeadamente: [21]

- Duplicação de dados;
- Falta de dados;
- Dados incompletos;
- Dados desatualizados;
- Dados imprecisos/errados;

*Volume: escala dos dados(terabytes, petabytes, entre outros)
†Velocidade: refere-se à rapidez com que os dados são gerados e processados
‡Veracidade: fiabilidade dos dados
§Variedade: tipo de dados (estruturados, semi-estruturados ou não estrtuturados)

- Dados que não cumprem as regras de negócio;
- Erros na transformação de dados.

2.2.3 Objetivos da Qualidade de Dados

Os objetivos da DQ visam oferecer a contribuição necessária para delineação da estrutura da DQ, abarcando a formulação de exigências que incluem a definição de requisitos, políticas de inspeção, medidas e monitorizações que refletem a mudança na qualidade e desempenho dos dados. [22]

O ciclo da DQ engloba diversas etapas que devem ser ponderadas, as quais facilitam a administração do processo contínuo de aprimoramento da qualidade dos dados. É imperativo compreender o funcionamento do ciclo da DQ e como deve ser perpetuamente administrado, a fim de alcançar o estado de DQ ideal e satisfazer as necessidades da organização. [23]

O ciclo da DQ é inicializado pela identificação exaustiva de todos os metadados. Tal exercício congregará informações referentes à estruturação da base de dados e às relações existentes na base de dados (exemplo: chaves primárias e estrangeiras). A informação compilada nessa fase é arquivada para futura consulta.

De um modo geral, a etapa de elaboração do perfil de dados é executada simultaneamente à fase de descoberta dos dados. Durante esta última fase, são reunidos todos os metadados, informação que a fase de criação do perfil de dados utiliza para relacionar com os dados reais e comparar definições dos metadados com os dados reais. O resultado é então registado para posterior consulta e análise.

As dimensões da DQ devem encontrar-se acessíveis nas ferramentas empregues na fase de elaboração do perfil de dados, a fim de avaliar as dimensões 2.1, representação e validade dos metadados dos dados.

A definição de regras de DQ devem ter em conta os requisitos de negócio. As fases de descoberta e elaboração de perfil dos dados configuram contribuições essenciais a serem utilizadas na concepção das regras de DQ, dado que fornecem as definições dos

metadados nas quais as regras devem fundamentar-se. Subsequentemente na elaboração de regras de DQ consiste em identificar os elementos de dados cruciais que detêm maior importância em cada domínio de dados do negócio.

A monitorização da DQ, assim como as regras, deve ser delineada com base nas necessidades do negócio e na criticidade dos dados, abarcando aspetos como a localização e frequência da monitorização, o limite de erros, criticidade das exceções e os requisitos de notificações. Esta fase deve fazer uso das regras de DQ desenvolvidas, nas quais os resultados são comparados com os limites definidos que ditam a geração das exceções e notificações a gerar. Os resultados são então armazenados numa base de dados para posteriormente serem apresentados em relatórios.

A fase de *reporting* tem como dependências a geração e armazenamento de todas as exceções, que devem ser armazenadas de forma agregada e disponibilizando a contagem de exceções por dimensão de DQ. Tal etapa deve facultar a visualização das tendências das exceções ao longo do tempo, bem como a capacidade de aprofundar-se numa visão mais detalhada.

A fase de correção de dados deve ser gerida pelos administradores de dados. É necessário executar uma análise da causa do problema para identificar o que está a causar as exceções de DQ. Após a identificação das causas e a implementação do plano de remediação, a monitorização e *reporting* da DQ devem apresentar evidências de melhorias ao longo do tempo.

Os propósitos do ciclo de qualidade de dados têm em vista:

- Avaliar a progressão dos dados relativamente a expectativas de negócio definidas;
- Gerenciar a qualidade dos dados incorporados ao ciclo de vida de desenvolvimento de sistemas, mediante a formulação de requisitos específicos;
- Definir e disponibilizar os processos para monitorizar, medir e reportar a conformidade de níveis aceitáveis da DQ.

2.3 Ferramentas de Qualidade de Dados

Na implementação de controlo de qualidade, a seleção das ferramentas apropriadas é um passo crucial que deve ser realizado durante a fase de planeamento, considerando o contexto específico da organização. [4] Serão apresentadas duas ferramentas atualmente utilizadas pela empresa *MC Sonae*, que permitem a aplicação de CQ's.

2.3.1 IBM InfoSphere DataStage (DS)

O IBM *InfoSphere DataStage* é a componente de integração de dados do Servidor de Informação IBM *InfoSphere*.
Trata-se de uma ferramenta de integração de dados que suporta ETL e oferece recursos avançados de DQ. Com o DS, é possível implementar processos automatizados de validação, transformação, limpeza e monitorização de dados, garantindo que os dados usados pela organização estejam sempre precisos e atualizados, disponibilizando assim funcionalidades que ajudam a: [24]

- Efetuar a compreensão dos dados bem como as suas relações;
- Analisar e monitorizar continuamente a qualidade dos dados;
- Realizar a limpeza, padronizar e combinar dados.

2.3.2 RAID - Revenue Assurance Integration Driller (Mobileum, Inc.)

Revenue Assurance Integration Driller (RAID) é uma ferramenta proprietária da Mobileum, Inc. que agrega um conjunto completo de funcionalidades que lhe permitem integrar dados de todas as fontes e formatos, bem como enriquecer e disponibiliza-los através da implementação e configuração de *dashboards* adaptados a cada contexto. [25]

RAID é uma ferramenta projetada principalmente para garantia de receita, gestão de fraudes e garantia de negócios. No entanto, a natureza versátil dessa ferramenta possibilita a aplicação de Controlo de Qualidade de Dados. Dessa forma, é possível integrar os dados que se pretende controlar, definir regras ou expressões de qualidade a serem aplicadas aos dados e gerar alertas e indicadores resultantes da aplicação dessas regras. A divulgação dos resultados pode ser realizada de diversas maneiras, incluindo através de *dashboards*, *emails* ou mensagens de texto.

2.3.3 Comparação das ferramentas

A Tabela 2.2 fornece um resumo das vantagens e desvantagens associadas à *IBM InfoSphere Information Server* (IBM) e à tecnologia *Revenue Assurance integration Driller* (RAID).

Ferramenta	Vantagens	Desvantagens
IBM	Serviço na cloud; Capacidade de processamento de dados; Integração com *Big Data*; Inteligência artificial e *Machine Learning*.	Complexidade na aprendizagem inicial; Instalação complexa.
RAID	Serviço na cloud; *Dashboards* de visualização de dados versáteis; Integração com *Big Data*; Inteligência artificial e *Machine Learning*.	Pequena comunidade de utilizadores para partilha de conhecimento; Não tem módulo dedicado de DQ.

TABELA 2.2: Comparação de ferramentas

2.4 Arquitetura de dados de BI

A Figura 2.2 ilustra a arquitetura de dados em vigor na área de BI na *MC Sonae*. Nessa representação, é possível visualizar como os dados são movimentados (fluxo de dados), desde a sua geração nos sistemas operacionais até estarem disponíveis nos sistemas analíticos. Cumpre ressaltar que a arquitetura é objeto de avaliação contínua e passível de atualizações, as quais são realizadas de acordo com as necessidades emergentes e o desenvolvimento constante da empresa.

.

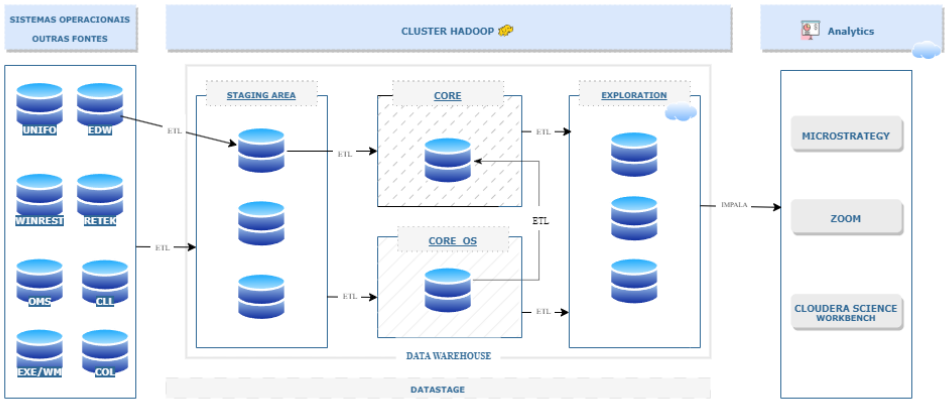

FIGURA 2.2: Arquitetura de dados de BI

2.4.1 Sistemas Operacionais e Outras Fontes

Os sistemas operacionais, *data sources*, desempenham um papel de extrema relevância ao fornecer um suporte direto às atividades empresariais, sendo encarregues de gerar uma parcela substancial dos dados factuais da organização. Adicionalmente, existem outras fontes de dados que desempenham um papel primordial na parametrização, distribuição e armazenamento das informações empregadas pelos sistemas operativos, bem como no armazenamento dos dados gerados por tais sistemas. Conforme se observa na Figura 2.2, os sistemas operativos e outras fontes incluem: *RETEK, EDW, UNIFO, WM, WINREST, CLL, OMS e COL.*

Sistemas Operacionais e Outras Fontes

Sistema	Dados a serem integrados	Tipologia dos dados	Periocidade
RETEK (ERP)	Contém dados sobre: Encomendas, Descontos, Faturação, Entregas, Quebras, Devoluções, Meios de pagamento, stocks, fornecedores, clientes.	Estruturados (tabelas)	Diária
UNIFO	Sistema único de Front Office para transação e faturação de artigos.	Estruturados (tabelas)	Diária
EDW Enterprise Data Warehouse	Contém informação sobre vendas, stocks, entre outros. Base da dados central com toda a informação para apoio à decisão da Sonae MC.	Estruturados (tabelas)	Diária
WM Warehouse Management	Sistema operacional para controlo da cadeia de entrepostos.	Estruturados (tabelas)	Diária
CLL Customer Link Loyalty	Gere programas de fidelização, execução de promoções e dados de clientes fidelizados (cartão continente, universo, worten).	Estruturados (tabelas)	Diária
OMS Order Management System	Sistema de gestão das encomendas de cliente de ecomerce da MC Sonae: Encomendas, Descontos, Faturação, Entregas, Quebras, Devoluções, Meios de pagamento.	Estruturados (tabelas)	Diária
COL Continente Online	Plataforma de eCommerce para Continente e outras insígnias da MC (www.continente.pt).	Estruturados (tabelas)	Diária
WINREST	Sistema opracional para vendas (Front Office e Back Office) das lojas Bagga.	Estruturados (tabelas)	Diária

FIGURA 2.3: Sistemas Operacionais e outras fontes da MC

2.4.2 *Cluster Hadoop*

O *Cluster Hadoop* corresponde à área de trabalho (DW de BI) para as equipas de negócio efetuarem exploração de dados (diretamente na base de dados) e criarem novos processos de transformação de dados (área com acesso a *storage* e processamento reservado para equipas).

As tarefas *batch* referem-se a processos que são executados na ausência total (ou quase total) de interação humana ou são agendadas para executar com base em premissas.
Estas tarefas são caracterizadas por processar grandes volumes de dados de uma só vez, em intervalos predefinidos, como diários, semanais ou mensais. Durante o processamento *batch*, os dados são extraídos das fontes, passam por processos de ETL para se adequarem ao modelo de dados do data *warehouse*. Posteriormente, são carregados no destino final. [26]

2.4.3 Analytics

Analytics, conforme evidenciado na Figura 2.2, constitui a etapa responsável por disponibilizar a informação. Após o processamento dos dados mediante *batch*, a informação é apresentada em *dashboards* ou em ambientes orientados ao *self-service*, acessíveis aos utilizadores que necessitam da informação na empresa. Consoante os objetivos ou requisitos dos utilizadores, os dados podem ser consultados por diferentes aplicações, tais como *MicroStrategy* ou *Cloudera Science Workbench*. A informação é então disponibilizada em modelos organizados que podem ser explorados de forma flexível diretamente na base de dados ou na plataforma de visualização (*Microstrategy*). Existem dois tipos de perfis de utilizadores com propósitos distintos: os consumidores e os analistas. Os consumidores têm como objetivo apenas de analisar a informação, enquanto que os analistas avançados podem explorar e produzir novas análises.

2.4.4 Dashboard

Um *dashboard* é um tipo de interface gráfica do usuário que geralmente fornece visualizações rápidas dos principais KPIs relevantes para um objetivo ou processo de negócios específico.

Para desenvolver *dashboards* , é de suma importância utilizar dados de alta qualidade. Esses dados representam os KPIs, que por sua vez são cruciais para monitorizar o progresso e a eficácia das organizações. Quando os dados utilizados para calcular os KPIs são imprecisos ou incorretos, a confiabilidade desses indicadores fica comprometida, o que pode resultar em tomadas de decisão erradas por parte da empresa. Portanto, a qualidade dos dados é fundamental para garantir a integridade e a eficácia dos *dashboards* de controlo e, por extensão, para assegurar que as decisões empresariais sejam fundamentadas e acertadas. [16]

2.5 Séries temporais

Uma série temporal é definida como um conjunto de observações associadas a um determinado fenómeno aleatório, efetuadas em períodos de tempo sucessivos e estatisticamente relacionadas. A análise de séries temporais é uma abordagem sistemática que permite não apenas compreender como uma variável se altera ao longo desse período de tempo, proporcionando assim uma descrição dos dados, como também permite estabelecer uma relação entre duas ou mais variáveis ao longo de um determinado período de tempo. [27]

Dado este contexto, nas próximas subsecções serão apresentados alguns conceitos fundametais relacionados a séries temporais, juntamente com alguns modelos utilizados para a previsão de séries temporais.

2.5.1 Séries temporais - o conceito

De acordo com Chatfield (2000) [28], uma série temporal é um conjunto de observações registadas em sequência ao longo do tempo. Essas observações podem ser realizadas de forma contínua ou em momentos específicos, geralmente com intervalos regulares, como dias, meses, trimestres ou anos. Independentemente da natureza da variável medida - discreta ou contínua - esses dois tipos de séries são denominados séries temporais contínuas e séries temporais discretas, respetivamente.

Para representar cada obervação na série temporal, utiliza-se geralmente a notação Y_t. No caso em que os dados tenham sido observados em intervalos uniformes de tempo,

a indexação é realizada por: $t = \ldots, -1, 0, 1, 2, \ldots$, refletindo o momento em que cada observação foi efetuada. No entanto, quando os dados não são recolhidos em intervalos igualmente espaçados: apesar de se adotar a notação $t = 1, 2, \ldots$, os dados recolhidos apenas podem ser referentes a instantes particulares, num subconjunto de 1,2,3, ... Consequentemente, $t_i - t_{i-1}$ não é necessariamente igual a um, conforme abordado por Silva (2021). [29]

As séries temporais podem ser classificadas como univariadas, constituídas apenas pela variável de interesse associada a um certo espaço temporal, ou multivariadas, constituídas por mais do que uma variável de interesse dependente do tempo.

O estudo de uma série temporal, de um modo geral, tem como principais objetivos:

- Previsão: prever o comportamento futuro da série, ou seja, estimar os valores futuros da série para qualquer horizonte temporal;
- Análise descritiva dos dados: descrever os dados de modo a identificar tendências ao longo do tempo, como padrões sazonais e identificar o motivo que justifique esse comportamento;
- Modelação: encontrar um modelo matemático adequado para a evolução de série temporal;
- Controlo: Uma previsão precisa possibilita ao analista intervir de maneira eficaz para gerenciar um processo específico.

2.5.2 Componentes de uma série temporal

A variação de uma série temporal pode ser decomposta em quatro componentes: tendência (T), sazonalidade (S), componente residual/irregular (E) e cíclica (C). Estas são caracterizadas como:

- Tendência: refere-se ao padrão e à inclinação que a série temporal apresenta ao longo do tempo, podendo ser linear (ou não), crescente ou decrescente. Caso uma série não exiba um comportamento crescente ou decrescente, a série diz-se estacionária em média.

- Sazonalidade: corresponde ás posições típicas que ocorrem ao longo do tempo, ou seja, a um padrão, seja de aumento ou de diminuição, que ocorre regularmente em períodos específicos. A variação sazonal é muitas vezes removida através do ajustamento sazonal, uma vez que pode ocultar ou atenuar outras componentes relevantes, como a tendência. Um exemplo de sazonalidade no setor do retalho é a sazonalidade positiva associada às festas de final de ano, como o Natal e o Ano Novo. Durante a temporada de festas, as vendas em muitas lojas tendem a aumentar significativamente devido ao aumento das compras de presentes, decorações e alimentos. No entanto, após as festas, as vendas podem diminuir à medida que os consumidores reduzem os seus gastos e as promoções de pós-festas diminuem.

- Componente residual: é uma componente irregular de natureza aleatória também designada por ruído branco (do inglês *white noise*). Esta componente contém qualquer variação não explicada pelas restantes componentes.

- Componente cíclica: É a componente que representa movimentos oscilatórios de longo prazo em torno da tendência e com periocidade não regular. Esta componente é desafiadora de prever, uma vez que é recursiva e não segue um padrão regular, ou seja, os intervalos entre os picos não tem regularidade definida. Portanto, é comum ignorá-la em séries ”curtas”.

Admitida a existência destas quatro componentes, a maioria dos métodos convencionais para a análise de séries temporais fundamenta-se na decomposição da variação da série nas suas componentes individuais.

Considere-se Y_t o valor da série temporal no instante t, T_t a componente de tendência no instante t, S_t a componente sazonal no instante t e E_t a componente irregular no instante t. As duas hipótese mais correntes são as seguintes:

- Hipótese aditiva:

$$Y_t = T_t + S_t + E_t \tag{2.1}$$

- Hipótese multiplicativa:

$$Y_t = T_t \times S_t \times E_t \tag{2.2}$$

Um modelo aditivo é preferível quando a magnitude das oscilações sazonais permanece constante, independentemente do nível da série, ao passo que o modelo multiplicativo é mais apropriado quando a magnitude das flutuações sazonais varia de acordo com a tendência da série.

De seguida descrevem-se os conceitos que serão usados posteriormente.

2.5.3 Estacionaridade

Uma série temporal é considerada estacionária quando as suas propriedades estatísticas, como a média, a variância e a covariância, permanecem inalteradas ao longo do tempo. Então, para uma série estacionária temporal Y_t, a sua distribuição $(Y_t, ..., Y_{t+s})$, para todo o s, não depende de t, ou seja, a série não pode ter tendência crescente ou decrescente nem apresentar sazonalidade. [29–31]

Para testar a estacionaridade das sérias existem vários testes com as respetivas hipóteses nula e alternativa, tais como:

- Augmented Dickey Fuller (ADF):

  Hipótese nula: A série tem uma raiz unitária, ou seja, a série é não estacionária
  Hipótese alternativa: A série não tem uma raiz unitária, ou seja, a série é estacionária

- Kwiatkowski-Phillips-Schmidt-Shin (KPSS):

  Hipótese nula: A série é estacionária (*trend stationary*)
  Hipótese alternativa: A série tem uma raiz unitária, ou seja, a série é não estacionária

A raiz unitária é uma caraterística das séries temporais não estacionárias, ou seja, esta está presente para $\alpha = 1$ na seguinte equação:

$$Y_t = \alpha Y_{t-1} + \beta X_e + \epsilon \tag{2.3}$$

onde Y_t corresponde ao valor da série temporal no instante t, X_e representa uma variável exógena (variável externa) que é explicativa e pertence à série temporal e ϵ é um processo

estocástico de média zero, não correlacionado em série, com variância constante σ^2. [29, 32, 33]

No caso em que os testes indicam que a série não é estacionária, torna-se necessário aplicar transformações a fim de converter numa associada estacionária. Dentre as transformações possíveis, destacam-se: [31]

- Diferenciação sazonal. O número de raízes unitárias presentes na série temporal indica a ordem da diferenciação (sendo desaconselhável a utilização de uma ordem superior a 2);
- Transformação Box-Cox (aproxima os dados a uma distribuição normal);
- Transformação logarítmica;
- Aplicação da raiz quadrada ou raiz cúbica aos dados.

2.5.4 Função Autocorrelação (ACF)

A função de autocorrelação (ACF) é utilizada para identificar a correlação que uma determinada série temporal apresenta com ela própria, desfasada em k períodos e é utilizada como indicador estatístico das características dos dados. Por outras palavras, uma série de *lag* k * exprime a correlação entre Y_t e Y_{t-k}, o que significa que a função avalia a correlação entre o valor da série temporal num dado instante t e o seu próprio valor deslocado k momentos temporais. [34]

A ACF é definida como:

$$\rho_k = \frac{cov(Y_t, Y_{t-k})}{V(Y_t)} \quad , k = 1, \ldots, n-1 \tag{2.4}$$

E pode ser estimada através de:

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \quad , k = 1, \ldots, n-1 \tag{2.5}$$

Onde:

- $\text{cov}(Y_t, Y_{t-k})$ representa a covariância entre Y_t e Y_{t-k},

*o termo "lag"refere-se a um desfasamento temporal, indicando o número de unidades de tempo pelas quais um ponto de dados está atrasado em relação a outro ponto de dados na mesma série.

- $\text{Var}(Y_t)$ é a variância de Y_t. Esta representa a dispersão dos valores em torno da média;
- $\bar{Y}$ representa a média de Y.

É comum antecipar que o primeiro valor do gráfico da ACF seja sempre igual a 1, uma vez que a correlação entre qualquer variável e ela própria é invariavelmente igual a 1.

Além de contribuir para a descrição dos dados, a ACF também desempenha um papel fundamental na verificação da estacionariedade, na seleção de modelos e na identificação da presença de sazonalidade. [35]

2.5.5 Função Autocorrelação Parcial (PACF)

Se o valor do primeiro elemento da série temporal, denotado por Y_1, estiver significativamente correlacionado com o valor do segundo elemento, Y_2, e o valor do segundo elemento estiver por sua vez correlacionado com o terceiro elemento, Y_3, então existe uma relação de alguma forma entre o primeiro e o terceiro elementos, ou seja, entre Y_1 e Y_3. Para analisar as autocorrelações na série temporal de forma a eliminar essa influência, é introduzido um conceito adicional: a função de autocorrelação parcial (PACF). [34]

A PACF de *lag k* corresponde à autocorrelação entre Y_t e Y_{t-k} que não é explicada pelos atrasos entre 1 e k:

$$\varphi_{kk} = corr(Y_t, Y_{t-k}|Y_{t-1}, \ldots, Y_{t-k+1}) \tag{2.6}$$

A PACF identifica a correlação * entre a variável no instante t e um dos seus desfasamentos, removendo os efeitos de outros desfasamentos. Por outras palavras, define a correlação entre as observações Y_{t-k} e Y_t. Esta remoção é realizada determinando Y_t e Y_{t-k} como combinação linear das observações $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-k+1}$.

2.5.6 Divisão dos dados

Antes de serem utilizados na aplicação dos modelos de previsão, os dados necessitam de ser preparados para garantir o seu processamento pelos algoritmos selecionados. A

*A correlação tem o objetivo de identificar e quantificar a relação entre duas variáveis quantitativas.

preparação dos dados varia com o tipo de algoritmo a utilizar.

A prática de dividir um conjunto de dados num conjunto de treino e noutro conjunto de teste visa avaliar o desempenho de um algoritmo. Poder-se-ia utilizar a totalidade do histórico de dados, criando assim um modelo de *Machine Learning* (ML) pronto para receber novas instâncias e realizar previsões. Contudo, esta técnica não garante que se consiga determinar o real desempenho do algoritmo associado. Podem ainda ocorrer casos de *overfitting* - situações em que os dados levam à criação de um modelo que se ajusta muito bem aos mesmos, mas é incapaz de prever novos resultados.

Idealmente, o conjunto de dados é dividido em duas partes: o conjunto de treino e o conjunto de teste. O conjunto de treino é usado para instruir o modelo, permitindo que ele aprenda e faça previsões futuras. O conjunto de teste é usado para avaliar imparcialmente o desempenho do modelo, comparando as suas previsões com os dados "não observados".
A escolha da proporção entre os conjuntos pode variar, geralmente a divisão é feita em proporções como 80% - 20%, 85% - 15% ou 90% - 10%.

Em projetos e algoritmos envolvendo séries temporais, nos quais as observações estão intrinsecamente ligadas ao componente temporal, a escolha dos conjuntos de dados não pode ser arbitrária. Não se justifica treinar o modelo com dados futuros a fim de prever o passado. Portanto, o conjunto de treino corresponde ao primeiro conjunto selecionado, enquanto o conjunto de teste corresponde ao conjunto subsequente (levando em consideração que a série temporal encontra-se organizada cronologicamente).

2.5.7 Processamento dos dados

O processamento de dados é o conjunto de técnicas e métodos utilizados para transformar dados brutos num formato mais adequado para análise. O pré-processamento de dados é uma etapa crucial na análise de dados e pode incluir técnicas como limpeza de dados, normalização, seleção de características e redução de dimensionalidade. [36]
As várias técnicas de processamento que existem podem enquadrar-se em quatro grupos principais: [37]

Limpeza dos dados: tratamento dos valores em falta através da sua remoção ou preenchimento (através de média, regressão ou preencher com zeros).

- Identificação de ruído nos dados: utilizando métodos como regressão;
- Remoção de *outliers*: é um procedimento estatístico no qual valores atípicos num conjunto de dados são identificados e excluídos. Esses outliers podem ser detetados utilizando intervalos interquartis (IQR). O IQR é uma medida da dispersão dos dados e é calculado como a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1) do conjunto de dados. Portanto, qualquer valor que inferior a $Q1 - 1,5 * IQR$ ou superior a $Q3 + 1,5 * IQR$ é considerado um possível outlier e pode ser removido do conjunto de dados, pois esses valores são considerados incomuns ou extremos em relação à distribuição geral dos dados. *

Integração dos dados: refere-se ao processo de combinação de diferentes ficheiros ou fontes de dados num único conjunto de dados unificado, muitas vezes representado sob a forma de um *dataframe* ou tabela.

Transformação dos dados: engloba a alteração do tipo de dados, como por exemplo, a conversão de uma variável do tipo texto para uma variável do tipo data.

- Normalizações e estandardizações
- Agregações dos dados: envolve a consolidação dos mesmos em níveis superiores, como, por exemplo, a agregação de dados diários para uma frequência semanal ou mensal.

2.6 Modelação

2.6.1 *AutoRegressive Integrated Moving Average* (ARIMA)

O modelo ARIMA, também conhecido como modelo Box-Jenkins, foi proposto por George Box e Gwilym Jenkins em 1976. Este modelo é uma abordagem matemática que utiliza processos autorregressivos, processos de médias móveis e um elemento de integração para realizar previsões futuras de séries temporais. [27]

*Neste procedimento basta que o outlier seja moderado (inferior a $Q1 - 1,5 * IQR$ ou superior a $Q3 + 1,5 * IQR$) contudo ainda existe a categoria de severo (inferior a $Q1 - 3 * IQR$ ou superior a $Q3 + 3 * IQR$).

ARIMA(p, d, q) é uma combinação do modelo AR(p) (*Auto Regressive*) com a componente de integração (I) e com o modelo MA(q) (*Moving Average*), onde p representa a ordem da parte do modelo autorregressivo, d o grau da primeira diferenciação envolvida e q indica a ordem da parte do modelo correspondente à média móvel. [31]

O modelo AR(p) pode ser escrito da seguinte forma :

$$y_t = c + \sum_{n=1}^{p} \alpha_n y_{t-n} + \epsilon_t \tag{2.7}$$

onde ϵ_t é o ruído branco.

O termo ruído branco refere-se a uma série temporal que é caracterizada por ser independente e identicamente distribuída (i.i.d) (variáveis aleatórias independentes e identicamente distribuídas) com , com média igual a zero. Isso implica que a variância, σ^2, é constante ao longo da série e que cada valor não possui correlação com os restantes. Os erros (ruído branco) na série temporal são distribuídos de forma independente e seguem uma distribuição normal, ou seja, ϵ_t segue uma distribuição normal com média zero, $\epsilon_t \sim N(0, \sigma^2)$.

o modelo MA(q), é definido como uma combinação linear de erros passados num modelo que se assemelha a uma regressão:

$$y_t = c + \sum_{n=1}^{q} \theta_n \epsilon_{t-n} + \epsilon_t \tag{2.8}$$

onde ϵ_t é o ruído branco.

No entanto, o modelo MA(q) não é considerado um modelo de regressão no seu sentido convencional, uma vez que os valores de ϵ_t não são efetivamente observados.

Assim, ARIMA(p, d, q) pode ser escrito em função dos valores desfasados de y_t e dos valores desfasados dos erros: [31]

$$y_t = c + \sum_{n=1}^{p} \alpha_n y_{t-n} + \sum_{n=1}^{q} \theta_n \epsilon_{t-n} + \epsilon_t \tag{2.9}$$

onde ϵ_t é o ruído branco.

É imperativo salientar que a aplicação de qualquer modelo ARIMA é condicional à estacionaridade das séries temporais. Desta forma, caso as séries temporais em questão não apresentem tal característica, torna-se necessário proceder a transformações conforme descritas em 2.5.3. O recurso à variável correspondente à diferenciação da série temporal y_t, no grau indicado pelo parâmetro d, é por vezes adotado como variável de interesse. A representação desta versão do modelo ARIMA é descrita da seguinte forma: [31]

$$d_t = c + \sum_{n=1}^{p} \alpha_n d_{t-n} + \sum_{n=1}^{q} \theta_n \epsilon_{t-n} + \epsilon_t, \tag{2.10}$$

onde d_t é a série temporal diferenciada, sendo $d \neq 0$.

Em suma, para aplicar um modelo ARIMA, será necessário: [38]

- Transformar a série temporal numa série estacionária, sendo necessário diferenciar a série d vezes; este processo pode ser realizado visualmente ou aplicando algum teste da raiz unitária que permita estimar o valor de d;
- Estimar os parâmetros p e q, através da análise do gráfico da ACF. Como nem sempre é fácil estimar estes valores dada a complexidade dos cálculos é vantajoso recorrer a algumas ferramentas disponíveis, como o *pyhton* * ou $R^{\dagger}$, pois conseguem de forma automática estimar e devolver os parâmetros ótimos a considerar.

De modo a comparar quais os parâmetros que originam o melhor modelo (seleção do modelo) existem os critérios de informação, dos quais se destacam o critério *Akaike Information* e o critério *Bayesian Information*.

O critério de Informação de Akaike (AIC) fundamenta-se na premissa de que o melhor modelo nem sempre é aquele que possui o maior número de coeficientes, sendo calculado como: [39]

$$AIC = -2log\text{-}like + 2p,$$

*Python é uma linguagem de programação também utilizada para desenvover algoritmos de ML
†R é uma linguagem de programação para computação estatística e gráficos.

sendo p o número de coeficientes e *log-like* o logaritmo da função de verosimilhança (do inglês *likelihood*), usando os coeficientes obtidos por estimação.

O critério Bayesian Information (BIC) impõe uma penalização mais rigorosa aos modelos com um maior número de coeficientes, com essa penalização a ser mais pronunciada em amostras de maior dimensão, com base na seguinte fórmula:

$$BIC = -2log\text{-}like + p * ln(n),$$

onde n é o número de observações.

2.6.2 Erros de previsão

Todas as previsões incluem uma componente de erro ou resíduo no resultado apresentado.

Com o objetivo de comparar diversos métodos e modelos de previsão, serão utilizadas duas medidas de erro distintas, tais como: Erro Percentual Absoluto Médio (MAPE) e a Raiz do Erro Quadrático Médio (RMSE). Dado um conjunto de dados com um total de T observações dividido em conjuntos de treino (com N observações) e teste (com $T - N$ observações) representados como $(y_1, ..., y_N)$ e $(y_{N+1}, ..., y_T)$ respetivamente. O modelo é treinado utilizando o primeiro conjunto em causa, e em seguida são feitas previsões para o conjunto de teste. Raramente um modelo consegue prever com exatidão os valores reais, motivo pelo qual calculamos os erros de previsão, que consistem na diferença entre os valores reais y_t do conjunto de teste e as previsões $\hat{y}_t$ correspondentes para o mesmo intervalo de tempo: [40] [41]

$$e_t = y_t - \hat{y}_t, \quad t \in \{N+1, \ldots, T\}. \tag{2.11}$$

Duas medidas de erro com uma utilização mais regular são: [41]

$$MAPE = \frac{1}{T-N} \sum_{t=N+1}^{T} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \times 100 \tag{2.12}$$

$$RMSE = \sqrt{\frac{1}{T-N} \sum_{t=N+1}^{T} (y_t - \hat{y}_t)^2} \tag{2.13}$$

Para analisar os resídos é possível realizar duas abordagens distintas: [42]

1. Distribuição dos erros: refere-se à forma como essas discrepâncias estão distribuídas em torno de zero. Idealmente, os resíduos devem seguir uma distribuição normal, o que significa que eles devem estar distribuídos simetricamente em torno de zero, formando uma curva em forma de sino. No entanto, em muitos casos, os resíduos podem não seguir uma distribuição normal, indicando desvios do pressuposto de normalidade;

2. Teste de normalidade: teste de *Shapiro-Wilk*, é utilizado para verificar se os resíduos seguem ou não uma distribuição normal. O teste avalia a hipótese nula de que os resíduos são normalmente distribuídos. Se o valor p associado ao teste for menor que um nível de significância predefinido (geralmente 0,05), a hipótese nula é rejeitada, indicando que os resíduos não seguem uma distribuição normal.

Capítulo 3

Modelo de *Data Quality*

O objetivo principal deste modelo consiste em consolidar dados de várias fontes, realizar transformações específicas e disponibilizar os dados consolidados para um *dashboard* de controlo de qualidade em *Microstrategy*. Portanto, com o intuito de explorar o domínio da DQ foi elaborado um projeto, *Data Quality Model*, em *Hadoop*. Neste projeto estiveram presentes as seguintes fases:

- Efetuar a compreensão das necessidades do negócio;
- Realizar a análise e compreensão dos dados;
- Preparar os dados e desenvolver a solução.

O processo inicial envolveu a identificação das tabelas que continham as informações relacionadas com o controlo de qualidade nas estruturas da MC. Após uma pesquisa extensa, foram identificadas as bases de dados que abrigam as informações provenientes desses controlos. Na secção 3.1.1, encontra-se uma exposição minuciosa dessas tabelas, acompanhada dos seus atributos e métricas, bem como das fontes de informação correspondentes.

A etapa subsequente consistiu em identificar e delinear a solução de ETL desejada, a qual concretizou a elaboração de um modelo centralizado e automatizado de ingestão de dados destinados a um *dashboard* final.

Para aprimorar as análises, foram incorporadas previsões baseadas em ML. Essas previsões têm como finalidade oferecer suporte às equipas de suporte, uma vez que, sempre

que a qualidade do CQ estiver abaixo de 98%, um registo é aberto. A equipa de suporte é então designada para investigar e compreender as razões pelas quais o CQ não teve nível de qualidade máximo. Se houver valores ausentes ou incoerentes, ou seja, se não tiver ocorrido a ingestão de todos os dados ou se tiver ocorrido a ingestão de dados, mas estes forem incorretos (incoerentes), a equipa é responsável por acionar os mecanismos de reprocessamento dos dados. Portanto, a previsão associada a cada control fornece informações futuras sobre a qualidade deste, permitindo assim que as equipas resolvam os problemas de forma atempada.

Por fim, culminou-se na criação de um *dashboard* em *MicroStrategy*, o qual é gerado com os dados provenientes do modelo único centralizado, contendo todas as informações dos controlos aplicados na MC. Este *dashboard* será atualizado diariamente com os dados mais recentes, juntamente com métricas relativas à qualidade dos dados. Além disso, estará acessível a toda a equipa de BI com foco na equipa de suporte e a utilizadores chave (analistas avançados), proporcionando-lhes acesso às métricas da DQ.

3.1 Diagrama da solução

Após concluir a pesquisa e identificar os elementos anteriormente mencionados, emergiu a estrutura da solução, *Data Quality Model*, como representado na Figura 3.1. A figura descreve sucintamente o processo que concretizou o diagrama de solução. Esta estrutura é composta por quatro componentes distintas, que se encontram identificadas e denominadas na figura como: *Oracle/Hadoop*, *DataStage*, *Azure Databricks e Microstrategy*. Inicialmente, na camada *Oracle/Hadoop* encontram-se localizadas as BD com os dados resultantes dos CQ's. Em seguida, ocorre a recolha desses dados contendo as informações relevantes para o estudo. Estes são ingeridos em *Hadoop* através do modelo desenvolvido em DS, "*Data Quality Model*". Posteriormente, é realizada a modelação dos dados (do inglês *Data Analysis*) utilizando algoritmos de ML através da utilização do serviço *Azure Databricks* da *Microsoft*. Por fim, é elaborado um *dashboard* em *Microstrategy* com os KPIs e métricas relevantes para a organização.

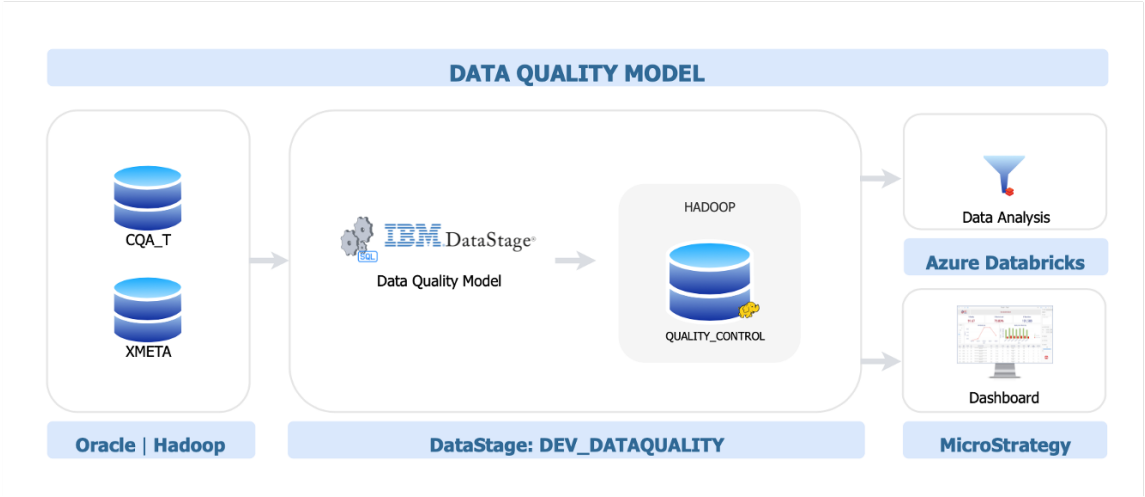

FIGURA 3.1: Diagrama da solução

As secções subsequentes providenciarão uma descrição detalhada de cada fase do modelo.

3.1.1 Tipos de dados e fontes de informação

Os dados que constituem a base deste relatório de estágio compreendem as informações resultantes dos processos automatizados de controlo de qualidade aplicado aos sistemas operacionais. Esses controlos são implementados nas plataformas RAID (RAID CQA) e em DS (IBM DQ). As informações são armazenadas em locais distintos: os dados de DS são armazenados na base de dados **XMETA** do *Hadoop*, enquanto os dados do *RAID* são armazenados na base de dados **CQA_T** em *Oracle*. Esses dados encontram-se ilustrados graficamente na Figura 3.1, localizados em 'Oracle | Hadoop'.

A base de dados XMETA é composta por várias tabelas, no entanto, foram selecionadas apenas as duas tabelas relevantes: `XMETA_EMEXCEPTIONSET` (Tabela A.1) , e `EMSETPROPERTY` (Tabela A.2). A base de dados CQA_T é composta por várias tabelas, das quais apenas seis são relevantes para o estudo: `CQA_T_CASES` (Tabela A.3), `CQA_T_CONTROLOS` (TabelaA.4), `CQA_T_AREAS_OWNER` (TabelaA.5), `CQA_T_UTILIZADORES` (TabelaA.6), `CQA_T_SERVICOS` (TabelaA.7), `CQA_T_DOMINIOS` (Tabela A.8).

Cada tabela é composta por um conjunto de atributos e métricas que sintetizam cada execução referente a um CQ específico. Este conjunto abarca informações como a identificação do CQ, a descrição do controlo, o tipo de regra, a dimensão (conforme ilustrado na Figura 2.1), a data da execução, o nome da tabela onde o CQ está a ser realizado, o domínio dos dados, o projeto associado, a dimensão da tabela (ou seja,

o número total de linhas), o número de dados inconsistentes e, quando aplicável, a informação acerca de entre quais sistemas operacionais a comparação está a ser efetuada. Além disso, é registado se o CQ foi executado com sucesso ou não, informações sobre o responsável pela implementação do CQ, entre outros.

A descrição pormenorizada de cada atributo e métrica encontra-se no Apêndice A. Adicionalmente, os tipos de dados e as relações entre as tabelas são apresentados visualmente nas figuras subsequentes.

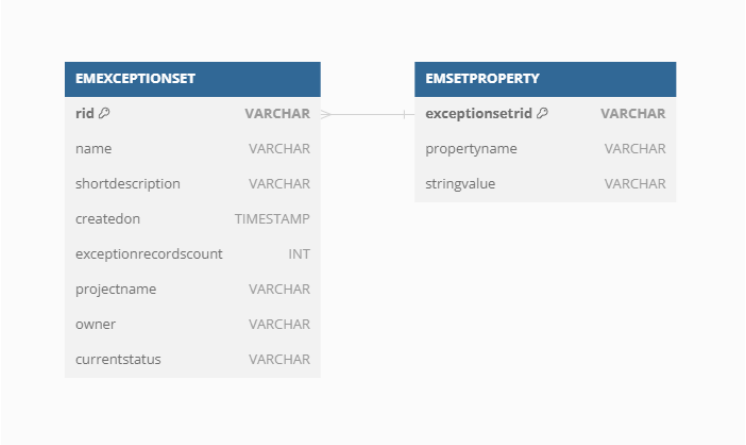

FIGURA 3.2: XMETA com as tabelas constituintes

FIGURA 3.3: CQA_T com as tabelas constituintes

Após uma análise minuciosa para compreender as informações contidas nas tabelas, bem como para entender as necessidades das equipas de suporte e negócio, resultou a estrutura da tabela final "*QUALITY_CONTROL*", representada na Figura 3.5. Esta tabela é o produto da integração de todas as informações provenientes tanto das tabelas XMETA como RAID_CQA, além da incorporação de novos atributos.

Conforme é observável, as tabelas na BD XMETA, em comparação com as de RAID_CQA, parecem conter menos informações. No entanto, é importante destacar que a informação nas tabelas XMETA não segue o formato tradicional de disposição. Neste contexto, os demais atributos são encontrados na coluna denominada `propertyname`, enquanto que as informações correspondentes estão contidas na coluna `stringvalue`. Esta formatação dificulta a análise dos dados então, para superar essa limitação, foi desenvolvida uma *query* que visa consolidar todas as informações numa única. Além disso, foi possível ainda criar novos atributos que não estavam originalmente presentes na coluna `propertyname`. Dentre esses novos atributos, destacam-se exemplos como `cq_rule`, `source1` e `source2`, os quais foram derivados da análise da coluna `shortdescription`. Esses atributos fornecem informações sobre o tipo de regra, as fontes de comparação e suas origens correspondentes.

É importante ressaltar que algumas informações precisaram de ser ajustadas, o que envolveu a manipulação dos dados (informação *Hard-Code* [*]), visto que essas informações não estavam disponíveis inicialmente e eram comuns a todos os registos da tabela. A título de exemplo, o campo `origem` foi padronizado como 'IBM DQ', o `status` foi definido como 'ativo', o domínio (`domain`) foi categorizado como 'Information Quality' e o tipo de alerta (`alert_type`) foi estabelecido como 'Email', entre outros ajustes necessários para uniformizar os dados.

As informações contidas nas tabelas RAID_CQA também foram agregadas de forma a se adequarem à estrutura da tabela final. Para alcançar esse resultado, também foi necessário efetuar *Hard-Code* aos dados, a titulo de exemplo, a criação da coluna `origem` como 'RAID' bem como `cq_rule` que foi definido como 'consistency'.

[*]Trata-se de uma seção do código em que um resultado, uma variável ou uma constante são inseridos diretamente no código fonte, sem passar por qualquer processo de transformação ou cálculo intermediário.

As tabelas presentes em *Hadoop* devem obedecer a determinadas regras específicas. Estas regras incluem a presença de colunas com a data de execução no formato de data e decimal, bem como a data da última atualização de cada linha e uma coluna de particionamento *. Para cumprir essas exigências, foram adicionados os seguintes atributos à tabela "*QUALITY_CONTROL*":

- `create_date`: Esta coluna armazena a data de execução no formato de data (timestamp).
- `time_key`: Trata-se de um número decimal derivado da data de execução. Inicialmente, a data é formatada como 'yyyy-MM-dd' e, em seguida, transformada no formato 'yyyyMMdd'.
- `last_updt_date`: Contém a data da última atualização de cada dado (ou linha) na tabela.
- `year_key`: Esta coluna é utilizada para fins de particionamento. Na tabela em questão o particionamento é realizado ao ano.

Essas adições permitem cumprir as regras estabelecidas pela organização e garantir a estrutura adequada da tabela. Nas figuras abaixo, encontra-se a estrutura da tabela, juntamente com a descrição dos atributos e métricas.

```
CREATE EXTERNAL TABLE QUALITY_CONTROL(

time_key DECIMAL(8,0) COMMENT 'Date of the execution: yyyyMMdd',
origem VARCHAR(128) COMMENT 'Data source: IBM DQ or RAID',
cq_id VARCHAR(128) COMMENT 'Id of quality control',
cq_name VARCHAR(128) COMMENT 'name of quality control ',
status VARCHAR(128) COMMENT 'IF CQ is: A: Ativo, E: Em Estabilização or I: Inativo',
cq_type VARCHAR(128) COMMENT 'Type of control: CQ, MON or DQ',
data_rule VARCHAR(128) COMMENT 'Data rule definition of quality control: Business OR Technical',
cq_rule VARCHAR(128) COMMENT 'Type: Consistency,  Completeness, Conformity, Integrity, Uniqueness, Validity or Coverage',
short_description VARCHAR(128) COMMENT 'Description of quality control',
source1 VARCHAR(128) COMMENT 'Source 1 (origem)',
source2 VARCHAR(128) COMMENT 'Source 2 (destino)',
domains VARCHAR(128) COMMENT 'Data Domain of Quality Control: Pós-venda, Consumidor e Mercado, Contabilidade, Gestão' ,
area VARCHAR(128) COMMENT 'Area: Business Intelligence, Supply Chain, Digital, Merchandising,...'
table_name VARCHAR(128) COMMENT 'Name of table'
project_name VARCHAR(128) COMMENT 'Project name'
owner VARCHAR(128) COMMENT 'Name of responsible Team',
respons_name VARCHAR(128) COMMENT 'Name of responsible Person',
service_name VARCHAR(128) COMMENT 'Service name: Descontos, Preço Venda, Preço Custo, Gama...',
fecho_contas VARCHAR(128) COMMENT 'Fecho contas: S (yes) or N (no) ',
alert_type VARCHAR(128) COMMENT 'Type of alert:  E: Email or R: Registo de ITSM',
SDTicket VARCHAR(128) COMMENT 'Ticket: Sim/Não'
thershold VARCHAR(128) COMMENT 'Threshold of CQ'
stage VARCHAR(128) COMMENT 'Stage of CQ: Created, Ativo, Em Estabilização, ...'
cq_result VARCHAR(128) COMMENT 'Status: Ok/Not ok',
success_cq_nrs DECIMAL(8,0) COMMENT 'Number of successful cqs: OK ',
unsuccess_cq_nrs DECIMAL(8,0) COMMENT 'Number of unsuccessful cqs: NOT OK',
alarms_nrs DECIMAL(8,0) COMMENT  'Total number of inconsistent values (nr total de linhas com diferença de valores)',
total_records DECIMAL(8,0) COMMENT 'Total number of inconsistent records (#linhas Inconsistentes)',
create_date TIMESTAMP COMMENT 'Execution Date: yy.MM.dd 00:00:00,000000000',
last_updt_date TIMESTAMP COMMENT 'Record Last Update Date',
year_key DECIMAL(8,0) COMMENT 'Partition column: yyyy'
)
```

FIGURA 3.4: Tabela final QUALITY_CONTROL com descrição dos atributos e métricas

*Uma tabela particionada é dividida em segmentos, chamados partições, que facilitam o gerenciamento e a consulta dos dados.

FIGURA 3.5: QUALITY_CONTROL

Após a definição dos atributos e métricas cruciais e estratégicas para as equipas envolvidas, bem como as tabelas estruturadas, avança-se para a próxima fase do modelo de solução. Neste contexto, a próxima etapa compreende a implementação, em DS, do modelo responsável pelos processos ETL e por automatizar o fluxo de dados desde a sua origem nos sistemas operacionais até ao seu destino final, que se localiza no ambiente *CORE* do *Hadoop*. Esta implemetação encontra-se detalhada na secção subsequente.

3.1.2 Modelo implementado em DataStage

Após a delineação de todas as informações relevantes a serem avaliadas, foi desenvolvido um processo automatizado em DS, no ambiente DEV_DATAQUALITY, que possibilita a obtenção da informação final desejada, mantendo-a atualizada e disponível no ambiente *Hadoop*. Este processo consiste na criação de um processo de ETL para a consolidação de

dados provenientes de diferentes fontes.

Antes de implementar o processo (*job*) no DS, foi elaborado um diagrama de fluxo com os principais passos a serem executados. Estes passos incluíram:

- Cálculo de granularidades: Determinação dos períodos de tempo a serem processados;
- Preparação e tratamento de dados: Utilização de consultas SQL para inserir dados em duas tabelas temporárias denominadas: ”working_xmeta”e ”working_raid”bem como aplicação de transformações e regras específicas para tratar e enriquecer os dados;
- Criação da Tabela Final: Combinação dos dados das tabelas temporárias com os dados históricos para formar a tabela final;
- Exploração: Explorar e disponibilizar os dados da tabela final para o negócio.
- *View*: Disponibilização dos dados da exploração em *Microstrategy*, a fim de serem utilizados no *dashboard*.

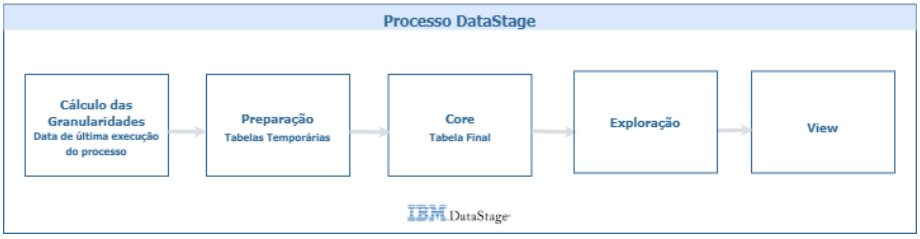

FIGURA 3.6: Diagrama com os cinco principais passos da solução em DS

O modelo em análise será agora abordado, com a ressalva de que, por motivos de segurança e proteção de dados, a explicação seguirá uma abordagem mais geral. Este modelo de solução pode ser dividido em três componentes distintas. Para uma melhor compreensão do modelo final, o mesmo pode ser visualizado na Figura 3.7.

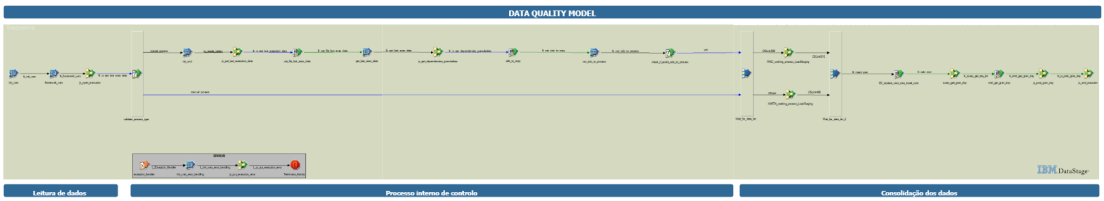

FIGURA 3.7: DEV_DATAQUALITY

Primeira componente - Leitura de Dados a Processar

A primeira parte do projeto envolve a preparação inicial para a execução do processo em DS. Esta inclui a abertura de uma execução na framework do DS, onde é calculada e registada numa variável a data da última execução bem-sucedida, para referência futura. Neste contexto, a execução identifica o período a ser processado, o qual é através da comparação da data atual na *framework* com a data da última execução bem-sucedida. Caso a data da *framework* seja menor que a data da última execução, o processo é desencadeado; caso contrário, o processo é terminado, indicando a ausência de novos dados a serem processados.
Adicionalmente, são estabelecidas as conexões necessárias para possibilitar a posterior interligação do DS com os sistemas operacionais, viabilizando assim a leitura dos dados.

Segunda componente - Processo Interno de Controlo

No segundo momento, ocorre um processamento interno de controlo. Sempre que o processo automático é executado com sucesso é gerado um ficheiro que contém o intervalo de tempo correspondente aos dados ingeridos. Então, o processamento interno envolve a leitura do ficheiro anteriormente referido, um arquivo estruturado no formato *JSON*, que contém a data da última execução bem sucedida bem como as datas que foram processadas nessa mesma execução. A análise desses dados é fundamental para a determinação do período temporal a ser processado, ao mesmo tempo que impede a inclusão de informações duplicadas na tabela final e minimiza a repetição excessiva dos mesmos dados.

Terceira componente - Consolidação

A terceira fase do projeto é dedicada à consolidação dos dados, preparando-os para a etapa subsequente de transformação e carregamento. Inicialmente, são realizadas leituras das tabelas para construir os parâmetros necessários para uso posterior no projeto. A etapa do *OBDC*, no DS, é responsável pela ingestão em *Hadoop* dos novos dados (dados sem histórico) resultantes das consultas (ver ApêndiceB) que serão executadas nas fontes de dados. Para efetuar essa consolidação, são criados dois fluxos de processamento paralelos, cada um deles conectado a um OBDC específico e munido da respetiva consulta correspondente. Os resultados obtidos através desses paralelos são, então, escritos nas tabelas temporárias denominadas working_raid e working_xmeta. A tabela working_raid contém os dados provenientes da tabela RAID_CQA que ainda não estão presentes na tabela final em *Hadoop*, enquanto a tabela working_xmeta aloca os dados da tabela XMETA com o mesmo propósito. Não obstante, importa salientar que estas tabelas ainda não assume a sua forma final, dado que se encontram na "*Staging Area*"da arquitetura de BI, conforme ilustrado na Figura 2.2. Consequentemente, após a conclusão desta fase de consolidação, é executada uma validação para determinar o número de registos resultante dessa operação. Essa validação é crucial para decidir se o processo deve prosseguir normalmente ou ser encerrado devido à ausência de dados para importação. Caso o número de registos seja distinto de zero, os dados consolidados são então integrados na tabela final. Como resultado, a tabela final é composta pelos dados históricos acumulados ao longo do tempo, acrescidos dos dados provenientes das tabelas 'working'. Após a ingestão dos dados em *Hadoop* é realizada a etapa de exploração. Esta etapa consiste em disponibilizar os dados na vista (*view*) para posteriores análises no *microstrategy*, os dados estão estruturados de acordo com a CORE. Por fim, a *view* anterior mencionada possibilita a utilização da informação no *dashboard* de controlo.

Este procedimento é automatizado e ocorre diariamente às 9h. Para definir este horário, foi necessário apresentar uma ficha de requisitos (ficha *cawa*) à equipa encarregue da implementação do processo automatizado. Nesta ficha, foram registados o nome do projeto (*job*) e o horário a ser automatizado.

Por outro lado, existe também a possibilidade de executar o processo manualmente. Este

procedimento assemelha-se ao processo automatizado, exceto na execução da *framework* (na Figura 3.6 corresponde processo interno de controlo), que é desconsiderada. O intervalo de tempo a ser processado deve ser introduzido manualmente nos parâmetros antes da execução do *job*.

3.1.3 *Dashboard*

Data Quality Report foi o *dashboard* desenvolvido com a finalidade de proporcionar uma análise abrangente da qualidade dos dados aos utilizadores. Este centraliza, de forma concisa, todos os dados relativos aos procedimentos de controlo de qualidade, os quais são provenientes do modelo *"Data Quality Model"* 3.1.2 que faz a ingestão dos dados em *Hadoop*.

Na figura subsequente é apresentada uma representação gráfica do protótipo deste *dashboard*:

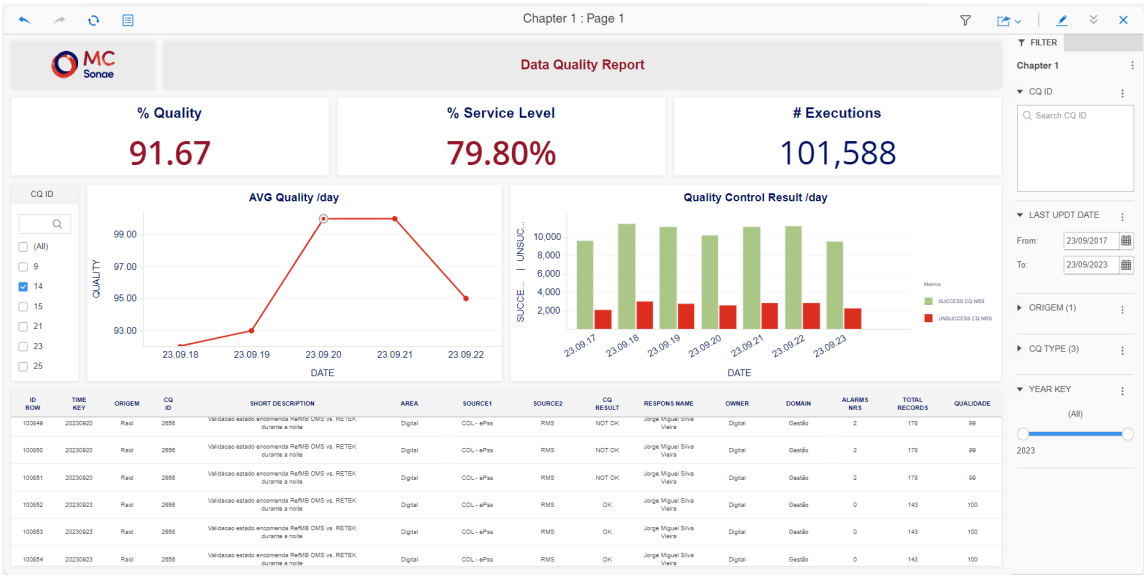

FIGURA 3.8: Protótipo do *Dashboard: Data Quality Report*

De seguida, procederei à análise pormenorizada de cada componente deste, destacando as suas funcionalidades e contribuições para a compreensão da qualidade dos dados.

Na parte superior, encontram-se os principais KPIs, estabelecidos pela empresa para a avaliação da qualidade dos dados. Estes compreendem a "Percentagem de Qualidade"(%*Quality*) e a "Percentagem de Nível de Serviço"(%*ServiceLevel*), cujas

respetivas fórmulas serão disponibilizadas visualmente na Figura 3.9 (figura retirada de relatórios internos da organização). Adicionalmente, é apresentado o número total de execuções (♯ Executions) dos CQ's no intervalo temporal definido.

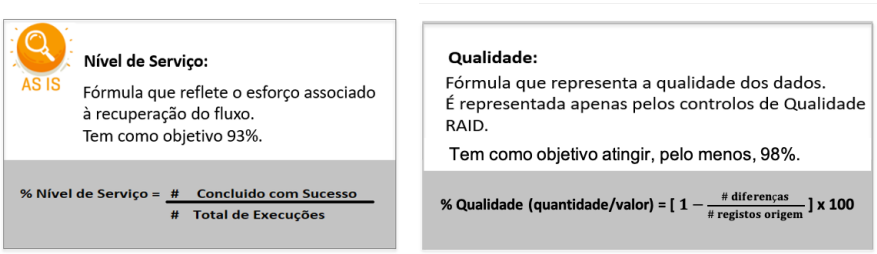

FIGURA 3.9: Principais KPIs

O *dashboard* integra ainda dois gráficos essenciais. O primeiro deles, um gráfico de linhas, representativo da média diária da qualidade de todos os CQ's. Um filtro está implementado, permitindo a seleção de um CQ específico e a adaptação do gráfico de acordo com a escolha efetuada. No caso de serem selecionados dois ou mais CQ's, também são exibidos dois ou mais gráficos respetivos. O segundo gráfico, de barras, apresenta o número de resultados dos CQ's diariamente, isto é, quantos controlos foram categorizados como "OK"e "NOT OK"em cada dia.

No espaço situado abaixo dos gráficos, encontra-se uma tabela que reúne informações essenciais provenientes da tabela final 'Quality_Control' , Tabela 3.5. Esta tabela detalha as execuções diárias de cada CQ, fornecendo uma visão mais aprofundada sobre as verificações realizadas.

Para proporcionar maior flexibilidade aos utilizadores, disponibilizam-se opções de filtros que permitem selecionar intervalos de tempo específicos ("LAST UPDT DATE"), identificadores de CQ (CQ ID), origem dos dados (ORIGEM), tipos de controlo (CQ TYPE) e também o ano (YEAR). Estas opções de filtro estão localizadas na barra lateral direita, permitindo aos utilizadores uma navegação intuitiva e personalizada dos dados, de acordo com as suas necessidades.

O público-alvo deste *dashboard* de controlo abrange analistas avançados da equipa de negócios e *developers* da BIT, com o objetivo de obter uma visão da qualidade dos dados em tempo real. Este conhecimento permitirá a tomada de decisões fundamentadas e a

implementação de correções, quando necessário, com vista à melhoria da qualidade dos dados.

Finalmente é relevante destacar que o *dashboard* de controlo é atualizado diariamente, assegurando que os utilizadores tenham acesso a informações atuais. Desta forma, *Data Quality Report* apresenta-se como uma ferramenta que oferece uma representação visual e intuitiva dos CQ's efetuados diariamente e disponibiliza métricas cruciais relacionadas com os KPIs da DQ. Importa ainda mencionar que o desenvolvimento deste *dashboard* foi efetuado na plataforma *MicroStrategy*.

3.2 Previsão associada a um Controlo de Qualidade

Com o intuito de antecipar a percentagem de qualidade de cada CQ, foi desenvolvido um algoritmo de previsão de séries temporais conhecido como ARIMA. Inicialmente, este algoritmo será aplicado de forma exclusiva aos dados de um cq_id específico, selecionado aleatoriamente da tabela final (conforme demonstrado na Figura 3.5). Entretanto, é relevante ressaltar que futuramente a previsão de qualidade para os demais CQ's seguirá o mesmo procedimento e metodologia.

A análise exploratória dos dados, bem como a construção do modelo adaptado aos dados, foram integralmente conduzidas na plataforma *Azure Databricks* da *Microsoft*, utilizando a linguagem de programação *Python*.

3.2.1 Identificação dos dados

Como foi identificado anteriormente, os dados que servirão de análise e posterior previsão são provenientes da tabela 3.5, originada pelo modelo único centraliazado que contém o resultado da execução dos CQ's.

Nesse contexto, procedeu-se à seleção aleatória de um identificador de CQ (`cq_id`), resultando na escolha do CQ com `cq_id=33`. A partir dessa seleção, foi construída o *dataframe* `data_raid`, que contém todas as informações pertinentes a este `cq_id`. A referida tabela abrange um período de 970 observações, desde 1 de janeiro de 2021 até ao dia 25 de setembro de 2023. Levando em consideração que essas datas englobam dados

correspondentes a um CQ com origem em 'RAID', cujo propósito consiste na comparação dos dados de vendas entre dois sistemas operacionais, nomeadamente EDW e *Hadoop*, com a finalidade de validar a coerência da informação relacionada com a margem de venda entre as publicações em EDW e em *Hadoop*.

Inicialmente, a métrica a ser prevista era `alarms_nrs`, que representa o número de linhas inconsistentes. Entretanto, observou-se que essa análise poderia ser subjetiva, uma vez que o valor absoluto de `alarms_nrs` não forneceria uma avaliação adequada, a menos que fosse contextualizado em relação ao volume total da tabela. Por exemplo, `alarms_nrs=50` numa tabela com 100 linhas teria um impacto diferente de `alarms_nrs=1500` numa tabela com 2.000.000 de linhas. Diante disso, a análise foi redirecionada para a previsão da qualidade do CQ, a qual é comparada a uma taxa de exceções em percentagem:

$$\%Qualidade(\text{quantidade/valor}) = [1 - \frac{\#\text{Diferenças}}{\#\text{Registos na Origem}}] \times 100\% \tag{3.1}$$

Na organização *MC Sonae*, foi estabelecido que o objetivo é atingir uma percentagem de qualidade/valor $\geq 98\%$.

Nesse contexto, criou-se um novo *dataframe* df_raid composto pelas colunas : `create_date` e `qualidade` calculada de acordo com a fórmula anterior mencionada. Esta a métrica está então restrita ao intervalo [0,100].

3.2.2 Análise exploratória dos dados

Os dados destinados à previsão foram submetidos a uma pré-seleção aleatória. Então, resultou num *dataframe* com 970 observações como anteriormente mencionado. De toda a informação disponível, apenas foram selecionadas duas variáveis: a métrica de qualidade expressa em percentagem (`qualidade`) e o atributo que representa a data de cada execução (`create_date`). Como resultado, esses dados resultaram num dataframe contendo apenas duas variáveis. Dado que a previsão se concentra exclusivamente numa variável em relação ao tempo, essa abordagem é denominada de univariada. Em consonância com a revisão de literatura mencionada anteriormente, será realizada uma análise com o objetivo de identificar possíveis valores discrepantes (*outliers*) e

compreender a natureza dessas discrepâncias.

É importante salientar que nenhum dos atributos `qualidade` e `create_date` apresentam valores em falta (*NULL ou NA*), portanto, os dados estão prontos para análises subsequentes e eventuais transformações necessárias.

Na Figura 3.10, a evolução da série temporal ao longo do tempo, bem como o IQR, são representados de forma gráfica. O eixo x representa a variável `create_date` (dia da obervação no formato 'yyy-MM-dd'), enquanto o eixo y representa a métrica `qualidade` (em percentagem). Além disso, no gráfico, as bandas que representam o IQR severo também são exibidas.

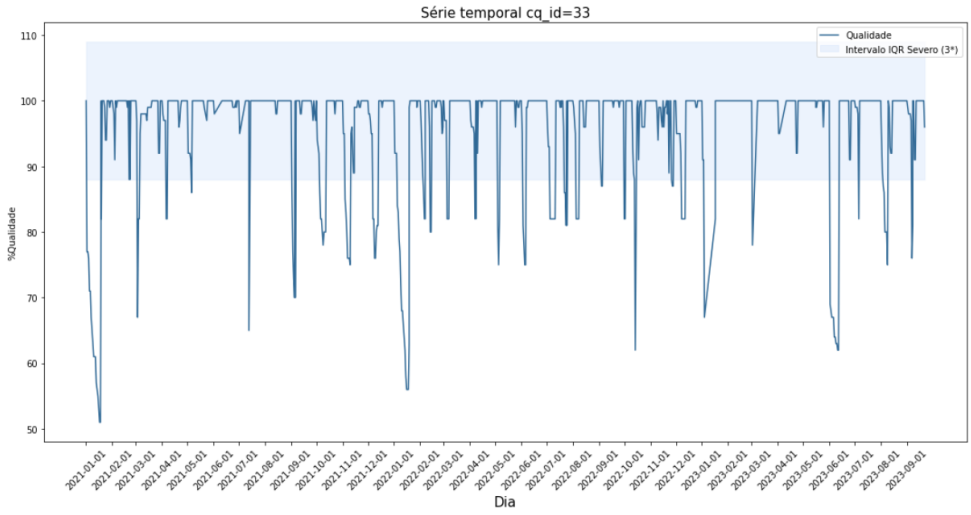

FIGURA 3.10: Gráfico com a evolução da qualidade para cq_id=33 com o respetivo IQR

Uma análise do gráfico revela claramente a presença de *outliers* e estes compreendem aproximadamente 18.04% do conjunto de dados, o que corresponde a um total de 175 obervações. É importante notar que esses *outliers* correspondem a observações genuínas e, portanto, num primeiro modelo não serão removidos do conjunto de dados.

Na análise unidimensional realizada, a Figura 3.11 contém o sumário das medidas centrais de dispersão da variável `qualidade` que caracterizam a distribuição dos dados:

	qualidade
count	970.000000
mean	95.734021
std	9.161287
min	51.000000
25%	97.000000
50%	100.000000
75%	100.000000
max	100.000000

FIGURA 3.11: Medidas centrais de dispersão

A média da variável em estudo foi calculada como sendo aproximadamente 95.73, indicando uma medida de tendência central que sugere que os valores em questão tendem a agrupar-se em torno desse valor médio. A mediana, que é o ponto que divide a distribuição em duas partes iguais, foi determinada como sendo igual a 100, evidenciando que metade dos valores observados está acima (ou igual) desse valor e a outra metade está abaixo. O desvio padrão, que quantifica a dispersão dos dados em relação à média, foi calculado como aproximadamente 9.16, o que indica uma dispersão relativamente baixa dos valores em torno da média. Esse valor sugere que os dados têm uma certa variabilidade, mas não apresentam uma dispersão elevada, o que pode ser interpretado como uma relativa consistência ou estabilidade nas medições. Os quartis, que são medidas estatísticas que dividem a distribuição em quatro partes iguais, revelaram que o Q1 é igual a 97, o Q2 é igual a 100 e o Q3 também é igual a 100.

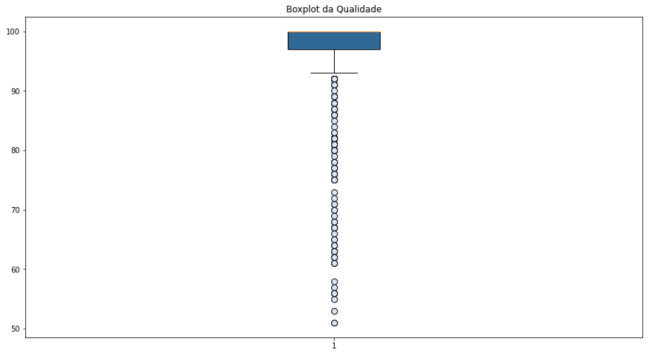

FIGURA 3.12: Gráficos *boxplot* da série temporal

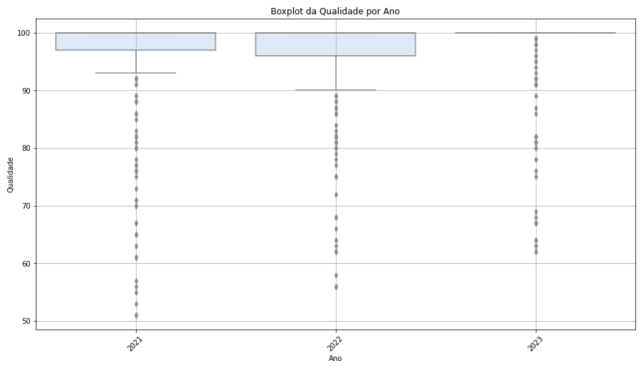

FIGURA 3.13: Gráficos *boxplots* da série temporal anual

Nos dados de qualidade apresentados na Tabela 3.13, podemos observar algumas tendências ao longo dos anos de 2021, 2022 e 2023.
No que diz respeito à média da qualidade, vemos que ela se mantém alta em todos os anos, com valores próximos a 96%. Isso sugere que, em média, o indicador mantém um nível elevado de qualidade ao longo do período analisado. Em relação à dispersão dos dados, o desvio padrão nos anos de 2021 e 2023 é relativamente maior, indicando uma maior variabilidade na qualidade durante esses anos. Por outro lado, o ano de 2022 apresenta um desvio padrão menor, sugerindo uma maior consistência na qualidade.

A análise dos quartis também é reveladora. O primeiro quartil (25%) e o terceiro quartil (75%) em todos os anos são consistentemente altos, indicando que a maior parte dos dados está concentrada em valores próximos a 100%. Em relação aos valores mínimos e máximos, todos os anos têm um valor mínimo acima de 50%, o que indica que a qualidade nunca caiu para níveis críticos. Além disso, o valor máximo em todos os anos é 100%, sugerindo que os dados atingem consistentemente o nível máximo de qualidade. Além disso, os dados correspondentes aos anos de 2021 e 2022 exibem uma assimetria negativa, já que a mediana (Q2) está mais próxima do Q3. Por outro lado, no ano de 2021, a distribuição da qualidade é mais concentrada no valor máximo de 100, com exceção de alguns pontos (*outliers*).

Em suma, os dados revelam que, ao longo dos anos de 2021, 2022 e 2023, a qualidade do indicador permaneceu consistentemente alta, com variações na disponibilidade de dados e algumas diferenças na dispersão. Essa análise fornece uma visão abrangente

da evolução da qualidade ao longo do tempo, com uma ênfase na alta qualidade média mantida durante todo o período.

3.2.3 Estacionaridade

Nesta etapa procedeu-se à avaliação da estacionaridade da série temporal. Para isso, foram realizado os testes de *Dickey-Fuller* Aumentado (ADF) e Phillips-Perron Estacionário em Tendência (KPSS), cujos resultados são apresentados e resumidos na seguinte figura:

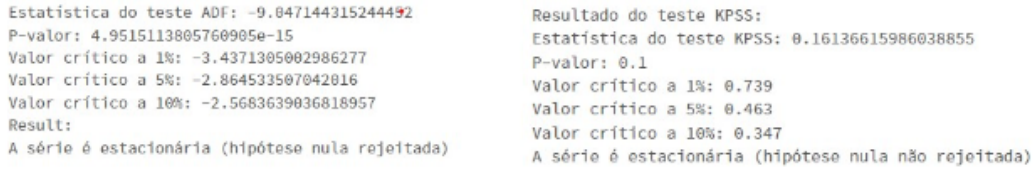

FIGURA 3.14: Testes de estacionaridade: ACF e KPSS respetivamente

Com base nos resultados dos testes estatísticos, é possível derivar conclusões pertinentes para a análise da estacionariedade da série temporal em consideração.

Primeiramente, ao analisar o teste ADF, constatamos que a estatística do teste obteve um valor de -9.047 e um valor-p extremamente baixo, próximo de zero (4.952e-15). Esse P-valor é muito inferior ao nível de significância de 0.05, que foi estabelecido como um limite crítico. Portanto, a estatística do teste ADF fornece evidências sólidas para a rejeição da hipótese nula de não estacionariedade. Adicionalmente, a estatística do teste é menor do que os valores críticos em todos os níveis de significância comuns (1%, 5%, e 10%). Isso fortalece ainda mais a conclusão de que a série temporal é claramente estacionária.

Por outro lado, o teste KPSS produziu uma estatística de teste de 0.161 e um valor-p de 0.1. A estatística do teste KPSS é menor do que os valores críticos para todos os níveis de significância comuns (1%, 5%, 10%). Isso indica que os dados são mais estacionários em comparação com uma série temporal não estacionária em tendência. A hipótese nula, neste caso, não é rejeitada.

Portanto, a decisão de manter a série temporal como estacionária é respaldada pelos resultados dos testes e pelo objetivo de previsão. Não são necessárias transformações na série, e, portanto, prossegue-se com a divisão em conjuntos de treino e teste.

3.2.4 Divisão dos dados

A divisão adequada de séries temporais em conjuntos de treino (*train_set*) e teste (*test_set*) desempenha um papel crucial na modelagem e previsão de dados temporais. Neste contexto, a ordem dos dados é de extrema importância, pois as observações estão dispostas cronologicamente. Este procedimento visa garantir que o modelo seja treinado em dados anteriores no tempo e testado em dados mais recentes, replicando assim as condições do mundo real em que as previsões são necessárias.

Inicialmente, procedeu-se à ordenação dos dados contidos no dataframe df_raid com base na data de registo das execuções dos CQ's, representada. Tal procedimento foi executado com a finalidade de garantir que as observações estivessem dispostas numa sequência cronológica precisa e ordenada.

Em seguida, devido ao considerável volume de dados presente no dataframe, optou-se por adotar uma abordagem convencional de divisão dos dados, alocando 85% das observações ao conjunto de treino e reservando os restantes 15% para o conjunto de teste. Portanto, o conjunto de treino compreende um total de 825 registos, abrangendo o período entre 1 de janeiro de 2021 e 10 de maio de 2023. Por consequência, o conjunto de teste é composto pelas observações remanescentes de df_raid. A Figura 3.15 ilustra a série temporal em questão, destacando a decomposição nos respetivos conjuntos.

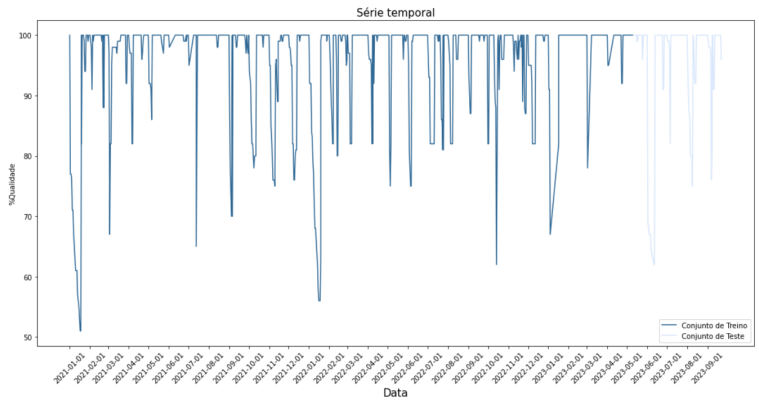

FIGURA 3.15: Série temporal df_raid identificados os conjuntos de treino e teste

3.3 Modelo ARIMA

Conforme previamente delineado nas seções precedentes, o modelo ARIMA é concebido para abordagens de análise e previsão de séries temporais. Nesse contexto, é factível a utilização direta do conjunto de dados representado pelo dataframe df_raid, prescindindo-se da inclusão de variáveis suplementares. Nesse cenário, restringe-se apenas à definição da variável temporal contendo as datas como o índice da série temporal em questão. Essa variável primordial, utilizada como entrada no modelo, é comumente denominada de "série temporal univariada".

No intuito de efetuar uma investigação mais minuciosa e aprofundada da série temporal, procedeu-se ao cálculo e à posterior visualização das funções ACF e PACF. Esta análise é apresentada de forma ilustrativa nas figuras subsequentes.

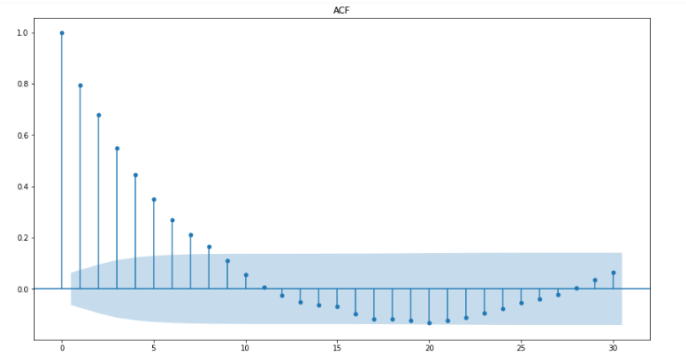

FIGURA 3.16: Função ACF

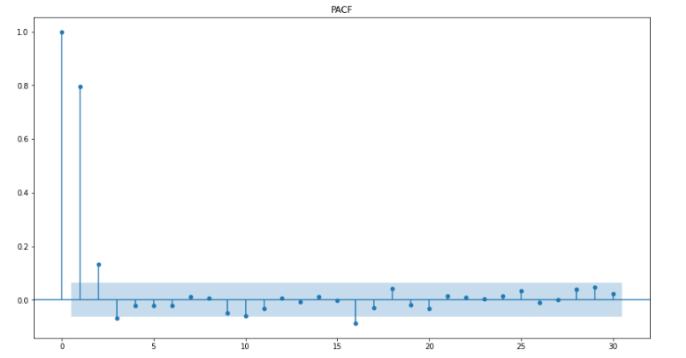

FIGURA 3.17: Função PACF

Conforme evidenciado, a ACF exibe um rápido decaimento em direção a zero, ressaltando assim a estacionaridade da série, como anteriormente mencionado na seção 3.2.3.

No entanto, é de notar que a ACF mantém-se acima da faixa de significância (faixa azul) para *lags* superiores a 1, notadamente nos *lags* 1, 2, 3, 4, 5, 6, 7 e 8. Por outro lado, a

PACF que também decai rapidamente para zero está fora da faixa de significância para *lags* iguais a 1, 2 e 3.
Com base nesses resultados, podemos concluir que embora a série temporal exiba características de estacionaridade de acordo com a rápida queda na ACF, a presença de autocorrelação significativa em *lags* posteriores (1 a 8) sugere a existência de uma estrutura de dependência temporal. Portanto, a estacionaridade da série pode não ser estritamente uniforme em todos os *lags*.

Para agilizar e automatizar o processo de seleção do modelo ARIMA mais adequado para a série temporal, utilizou-se a função *auto_arima* do *package pmdarima* na linguagem *Python*. Essa função tem a capacidade de identificar automaticamente as ordens ideais dos parâmetros p, d, q (parte não sazonal) e P, D, Q (no caso de a série apresentar sazonalidade). Além disso, é possível especificar os valores máximos para p, d, q, P, D e Q, permitindo que todas as combinações possíveis desses parâmetros sejam testadas.

A escolha do modelo ideal é baseada em critérios de informação, como o Critério de Informação AIC (*Akaike Information Criterion*) ou o BIC (*Bayesian Information Criterion*). Quanto menor o valor do AIC ou BIC, melhor é o ajuste do modelo aos dados. Portanto, a função *auto_arima* retorna o modelo ARIMA que minimiza o valor do critério escolhido previamente. Essa abordagem automatizada economiza tempo e garante que se escolhe o modelo ARIMA mais apropriado para a série temporal, com base numa análise objetiva e rigorosa dos critérios de informação.

No processo de modelação da série temporal, utilizou-se o conjunto de dados *train_set* para treinar o modelo, e o modelo que se destacou como o melhor foi o ARIMA(1,0,2). Resumidamente, esse modelo não requer diferenciação sazonal ($d = 0$) e incorpora uma dependência autoregressiva de ordem 1 ($p = 1$) e uma dependência de média móvel de ordem 2 ($q = 2$).

As previsões geradas por este modelo podem ser observadas na figura seguinte:

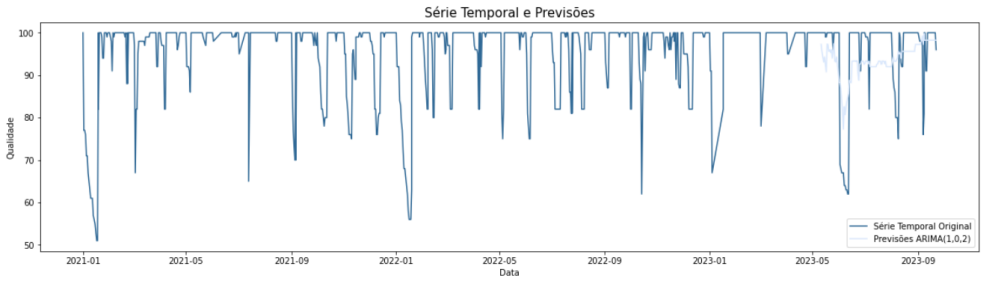

FIGURA 3.18: Previsões associadas ARIMA(1,0,2)

Estas previsões resultam da aplicação do modelo ARIMA(1,0,2) aos dados de treino e têm como finalidade estimar os valores futuros da série temporal. A representação gráfica das previsões é uma ferramenta útil para visualizar o desempenho do modelo e avaliar a sua capacidade de capturar os padrões presentes na série temporal.

A fim de comparar os valores reais com os previstos pelo foram foram computados os erros 2.12 e 2.13 associados à previsão:

$$\text{MAPE} \approx 8.12\%$$

$$\text{RMSE} \approx 8.69$$

Os erros obtidos não são consideravelmente elevados no ponto de vista matemático, com um MAPE de 8.12% e um RMSE de 8.69. Contudo, o modelo nunca conseguiu atingir o valor ideal 100 para o respetivo CQ em análise. Esses valores indicam que o modelo ARIMA(1,0,2) não aparenta ser adequado para os dados em causa. É ainda importante destacar que estes erros são relativamente altos no contexto empresarial, onde a precisão das previsões é fundamental e não é tolerável a presença de erros com esta dimensão.

O coeficiente de determinação, frequentemente representado como R^2, é uma métrica amplamente utilizada para avaliar o grau de adequação de um modelo estatístico aos dados observados. É uma medida que varia entre 0 e 1 e indica a proporção da variabilidade dos dados dependentes que é explicada pelo modelo. Quanto mais próximo de 1 for o valor de R^2, melhor o modelo se ajusta aos dados, indicando que uma porção maior da variabilidade é explicada pelas variáveis independentes do modelo. No contexto, foi computado como sendo 0.263. Isso significa que o modelo em questão conseguiu explicar aproximadamente 26.3% da variabilidade presente nos dados observados. Com base nesse valor de R^2, podemos concluir que o modelo não é ideal para os dados em análise.

A maior parte da variabilidade nos dados (cerca de 73.7%) permanece sem explicação pelo modelo atual.

De modo a visualizar os resíduos foram computados os gráficos de dispersão e histograma.

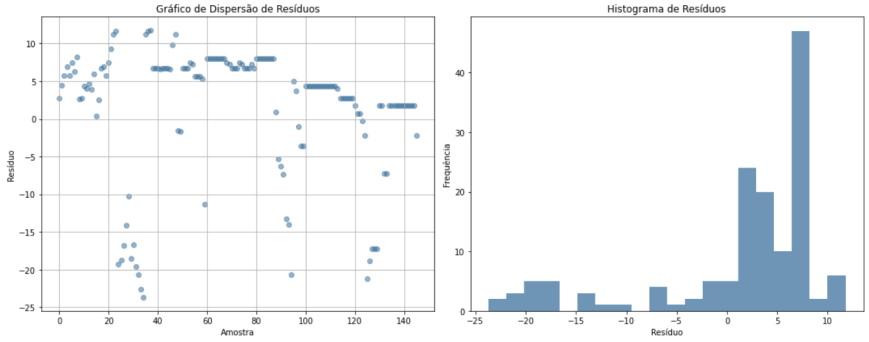

FIGURA 3.19: Gráfico de dispersão e histograma dos resíduos

No teste de Shapiro-Wilk realizado neste estudo, a estatística de teste foi calculada como 0.7796, e o valor p associado foi aproximadamente 1.525e-13, um valor muito baixo. Portanto, com base neste teste, rejeitamos a hipótese nula de que os resíduos seguem uma distribuição normal. Isso implica que os resíduos deste modelo não se ajustam à premissa da normalidade.

Além disso, foi realizado o "teste t de Student para a Média dos Resíduos". Neste estudo, a estatística de teste foi calculada como 2.1918 e o valor-p associado foi aproximadamente 0.0300. Como o valor-p foi menor que o nível de significância de 0.05, rejeitamos a hipótese nula: é assim plausível que a média dos resíduos não seja zero, para o nível de confiança a 95%. Portanto, com base neste teste, podemos concluir que o modelo apresenta um viés sistemático, tendendo a subestimar ou superestimar as previsões em relação aos valores reais.

Estes resultados levam a inferir que o modelo não se ajusta aos dados do CQ em análise.

Capítulo 4

Considerações finais e trabalho futuro

O *dashboard*, juntamente com o modelo único centralizado em *Hadoop*, proporcionam acesso simplificado às informações referentes à qualidade dos dados utilizados pelos analistas e pelas equipas de negócio. A utilização deste dashboard constitui, assim, um procedimento eficiente sempre que se torne necessário obter informações atualizadas. A sua utilidade é ainda evidenciada quando se identificam valores inferiores aos padrões estabelecidos pela organização, possibilitando uma rápida deteção e resolução das lacunas.

No âmbito do desenvolvimento do *dashboard* em questão, é imperativo salientar que se encontra, neste momento, numa fase de prototipagem. A sua finalidade primordial nesta etapa reside na submissão deste protótipo à apreciação das equipas envolventes, com o intuito de recolher opiniões e sugestões que permitam validar se a solução vai de encontro com as necessidades. Esta abordagem colaborativa visa assegurar a qualidade intrínseca e a eficácia do *dashboard*, bem como garantir a sua coerência com as necessidades e expectativas das várias equipas envolvidas no processo. É crucial realçar que, após a incorporação do feedback proveniente das mencionadas equipas, o *dashboard* transitará da fase de desenvolvimento para produção.

Com base nos resultados obtidos na modelação da série temporal, e considerando que foi selecionado o modelo ARIMA(1,0,2), outros modelos não foram considerados devido a restrições de tempo. No entanto, é fundamental destacar que, após análise minuciosa, concluiu-se que tanto os modelos ARIMA quanto os modelos SARIMA não são adequados para os dados em análise, devido à sua natureza irregular e aleatória.

Consequentemente, esses modelos não resultaram em previsões precisas ou apropriadas.

De acordo com o modelo implementado em *Hadoop*, existem oportunidades para melhorias futuras. A primeira abordagem consiste em aumentar a quantidade de dados históricos armazenados. Atualmente, só se dispõe de informações a partir de 2020, mas ao acumular um volume maior de dados, será possível obter uma visão mais completa da evolução e identificar possíveis tendências e regularidades nos dados. A segunda abordagem consiste em enriquecer os dados atualmente disponíveis. Atualmente, as informações estão limitadas aos CQ's, no entanto, é factível integrar informações relacionadas com a qualidade dos processos. Os processos, como, por exemplo, a ingestão dos dados de vendas no DW, estão definidos para ocorrer em períodos de tempo específicos, acontecendo diariamente a uma hora determinada. Portanto, sempre que ocorre um atraso no processo, isso resulta na diminuição da qualidade do mesmo. Dessa forma, é possível introduzir um KPI como métrica adicional para avaliar e compreender a percentagem de qualidade dos processos.

Com base nos resultados obtidos na modelação da série podemos também identificar áreas para trabalhos futuros. Dado que os dados não possuem regularidades nem sazonalidade, uma possível abordagem consiste em adotar a previsão com base numa série de valores binários, ou seja, 0 e 1. Nesse contexto, esses valores são atribuídos quando as amostras assumem valores iguais ou superiores a 96% ou, alternativamente, valores inferiores a esse limiar. Com esta metodologia, a intenção é utilizar modelos de séries temporais caudais com a finalidade de prever eventos em que as observações se situem abaixo dos limites previamente estabelecidos pela empresa. Isso representa uma mudança em relação à abordagem inicial, que se concentrava na previsão exata dos valores futuros.

É relevante destacar que o percurso até à obtenção da estrutura da tabela final foi uma tarefa que exigiu um período de tempo considerável. A identificação das BD que contêm informações relacionadas com os CQ's foi um processo meticuloso, seguido pela análise dos dados e pela subsequente seleção das informações relevantes de acordo com as necessidades dos utilizadores. A ausência de documentação adequada acrescentou complexidade a este processo.

Bibliografia

[1] M. Suer, "What is data quality and why is it important?" https://www.alation.com/blog/what-is-data-quality-why-is-it-important/, Aug. 2021. [1]

[2] P. P. Tallon, "Corporate governance of big data: Perspectives on value, risk, and cost," *Computer (Long Beach Calif.)*, vol. 46, no. 6, pp. 32–38, 2013. [2]

[3] J. Salido and P. Voon, *A Guide to Data Governance Privacy, Confidentiality and Compliance*. Microsoft Corporation, 2010. [2]

[4] S. Earley, *DAMA Guide to the Data Management Body of Knowledge*. Technics Publications, 2017, vol. 2nd. [5, 6, 7, 12, 16 e 20]

[5] F. H. A. Krishnan, "TDWI big data maturity model guide interpreting your assessment score," Tech. Rep., 2013. [6]

[6] R. B. Gartner, *Metadata: Shaping knowledge from antiquity to the semantic web*, 1st ed. Cham, Switzerland: Springer International Publishing, 2016. [7]

[7] R. Kimball and M. Ross, *The data warehouse toolkit: The definitive guide to dimensional modeling*, 3rd ed. Nashville, TN: John Wiley & Sons, 2013. [8]

[8] "What is a data warehouse?" https://www.oracle.com/pt/database/what-is-a-data-warehouse/. [8]

[9] Adobe, "Data warehouse vs. database — key differences explained," https://business.adobe.com/blog/basics/data-warehouse-vs-database, 08 2023. [9]

[10] W. W. Eckerson, *Performance dashboards: Measuring, monitoring, and managing your business*, W. W. Eckerson, Ed. Chichester, England: John Wiley & Sons, 2012. [10]

[11] G. Blokdyk, *Key performance indicator A complete guide - 2020 edition*. 5starcooks, 2020. [11]

[12] B. Marr, *Key Performance Indicators (KPI): The 75 measures every manager needs to know*. Harlow, England: Financial Times Prentice Hall, 2012. [11]

[13] Gartner, "Build a data quality operating model to drive data quality assurance," https://www.gartner.com/document/3980235?ref=hp-wylo, 2020. [12]

[14] Z. Panian, "Some practical experiences in data governance," *Journal of Decision Systems*, vol. 62, pp. 468–475, 02 2010. [12 e 16]

[15] C. Batini and M. Scannapieco, *Data and Information Quality: Dimensions, Principles and Techniques*. Springer, 2016. [13 e 16]

[16] A. Kumar, "What is data quality management? concepts examples," https://vitalflux.com/what-is-data-quality-management-concepts-examples/, Maio 2022. [15 e 24]

[17] Collibra. (2022) The 6 Dimensions of Data Quality. https://www.collibra.com/us/en/blog/the-6-dimensions-of-data-quality. [15]

[18] IBM. (2021) Data quality. https://www.ibm.com/topics/data-quality. [15]

[19] H. Moreno, "The importance of data quality – good, bad or ugly," https://www.forbes.com/sites/forbesinsights/2017/06/05/the-importance-of-data-quality-good-bad-or-ugly/, Jun. 2017. [17]

[20] M. L. Jibril, I. A. Mohammed, and A. Yakubu, "Social media analytics driven counterterrorism tool to improve intelligence gathering towards combating terrorism in nigeria," *Int. J. Adv. Sci. Technol.*, vol. 107, pp. 33–42, 2017. [17]

[21] Gartner, "Measuring the business value of data quality," https://www.gartner.com/en/documents/1819214, 2011. [17]

[22] S. Earley, *DAMA Guide to the Data Management Body of Knowledge*. Technics Pubns Llc, 2009, vol. 1st. [18]

[23] Knowledgent, "Building a successful data quality management program," https://knowledgent.com/whitepaper/building-successful-data-quality-management-program/, 2014, (acedido em 10/01/2023). [18]

[24] "IBM InfoSphere information server for data quality," https://www.ibm.com/products/infosphere-info-server-for-datamgmt. [20]

[25] "RAID is a platform designed for communication service providers that want to leverage their data assets to improve business processes and gain business insights, while at the same time simplify their IT environment?" https://www.mobileum.com/media/1855/raid_brochure_may2017.pdf, 2020. [20]

[26] IBM. (2010) What is batch processing? https://www.collibra.com/us/en/blog/the-6-dimensions-of-data-quality. [23]

[27] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. Springer New York, 1991. [24 e 31]

[28] C. Chatfield, *Time-Series Forecast*, 2000, vol. 1. [24]

[29] R. P. Silva, B. B. Zarpelão, A. Cano, and S. B. Junior, "Time series segmentation based on stationarity analysis to improve new samples prediction," *Sensors*, vol. 21, 11 2021. [25, 27 e 28]

[30] T. Kley, P. Preuß, and P. Fryzlewicz, "Predictive, finite-sample model choice for time series under stationarity and non-stationarity," *Electronic Journal of Statistics*, vol. 13, pp. 3710–3774, 2019.

[31] . A. G. Hyndman, R.J., "Forecasting: Principles and practice (3rd ed)," *OTexts: Melbourne, Australia. OTexts.com/fpp3*, 2021. [27, 28, 32 e 33]

[32] P. C. Phillips and P. Perron, "Testing for a unit root in time series regression," *Biometrika*, vol. 75, 1988. [28]

[33] R. I. Harris, "Testing for unit roots using the augmented dickey-fuller test. some issues relating to the size, power and the lag structure of the test," *Economics Letters*, vol. 38, 1992. [28]

[34] K. E. A. et al., "Forecasting the dynamics of cumulative covid-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-regressive integrated moving average (arima) and seasonal auto-regressive integrated moving average (sarima)," *Applied Soft Computing*, vol. 103, 2021. [28 e 29]

[35] W. Enders, *Applied Econometric Time Series (4.a ed.)*. W John Wiley Sons, 2014. [29]

[36] P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, and D. N. Rutledge, "New data preprocessing trends based on ensemble of multiple preprocessing techniques," *TrAC - Trends in Analytical Chemistry*, vol. 132, 11 2020. [30]

[37] S. García, J. Luengo, and F. Herrera, "Data preprocessing in data mining," *Intelligent Systems Reference Library*, vol. 72, 2015. [30]

[38] R. Hyndman and Y. Khandakar, "Automatic time series forecasting: The forecast package for r," *Journal of Statistical Software*, vol. 26, 07 2008. [33]

[39] W. N. Junior, O. Resende, G. K. Pinheiro, L. C. M. Silva, and E. R. Costa, "Use of aic and bic in desorption isotherms of tamarind seeds (tamarindus indica l.)," *Engenharia Agricola*, vol. 40, 2020. [33]

[40] T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)? -arguments against avoiding rmse in the literature," *Geoscientific Model Development*, vol. 7, pp. 1247–1250, 6 2014. [34]

[41] P. Ramos and J. M. Oliveira, "A procedure for identification of appropriate state space and arima models based on time-series cross-validation," *Algorithms*, vol. 9, 2016. [34]

[42] J. A. Mauricio, "Computing and using residuals in time series models," *Computational Statistics Data Analysis*, vol. 52, no. 3, pp. 1746–1763, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167947307002344 [35]

Apêndice A

Relativo às tabelas

Atributo	Descrição
rid	identificação única para cada execução dos cq's
name	Nome do cq
shortdescription	Descrição do cq
createdon	Data da execução
createdby	'InformationServerSystemUser'
exceptionrecordscount	Número total de linhas inconsistentes
projectname	Nome do projeto: DM_DATA_QUALITY
owner	'InformationServerSystemUser'
currentstatus	Estado do cq: 'created'

TABELA A.1: Descrição da tabela XMETA_EMEXCEPTIONSET

Atributo	Descrição
exceptionsetrid	identificação única da execução do cq
propertyname	Contém o nome dos atributos
stringvalue	Contém a informação dado o nome do atributo em propertyname

TABELA A.2: Descrição da tabela EMSETPROPERTY

Atributo	Descrição
ctr_id	Identificação única de cada CQ
dom_id	Identificação única do dominio do CQ
description	Descrição do CQ
amount	Número de diferenças
created_date	Data de execução do CQ
created_by	Nome do projeto: DM_DATA_QUALITY
amount_source	Número de dados na fonte
amount_destiny	Número de dados no destino

TABELA A.3: Descrição da tabela CQA_T_CASES

Atributo	Descrição
ctr_id	identificação única para cada execução dos cq's
aow_id	Identificação única da área
utl_id	Identificação única do utilizador
dom_id	Identificação única do dominio do CQ
srv_id	Identificação única do serviço
ctr_nome	Nome do CQ
ctr_descricao	Descrição do CQ
ctr_data	Data de execução do CQ
ctr_tp_alerta	Tipo de alerta gerado; E:Email ou R: Registo de ITSM
ctr_aplicacao_origem	Sistema operacional fonte
ctr_aplicacao_destino	Sistema operacional de destino
ctr_tipo_fluxo	Tipo do controlo; 1: CQ ou 2: monitorização
ctr_sentido_validacao	Sentido de validação do CQ; 1: validação nos dois sentidos ou 2: validação sentido da origem para o destino
ctr_fecho_contas	Fecho de contas; S: sim ou N: não

TABELA A.4: Descrição da tabela CQA_T_CONTROLOS

Atributo	Descrição
aow_id	Identificação única da área
aow_nome	Nome da área; ex: BI, People, Logistica, Negócio, Digital, entre outras
aow_descricao	Descrição com nome da equipa envolvente
created_date	Data em que a informação foi inserida na tabela CQA_T_AREAS_OWNER
modified_date	Data da ultima atualização da informação

TABELA A.5: Descrição da tabela `CQA_T_AREAS_OWNER`

Atributo	Descrição
utl_id	Identificação única da área
utl_nome	Nome do utilizador
utl_email	Email do utilizador
created_date	Data em que a informação foi inserida na tabela CQA_T_UTILIZADORES
modified_date	Data da ultima atualização da informação
utl_ativo	Atividade do utilizador; S: Ativo ou N: Desativo

TABELA A.6: Descrição da tabela `CQA_T_UTILIZADORES`

Atributo	Descrição
srv_id	Identificação única do serviço
srv_nome	Nome do serviço; ex: Fecho contas, Preço, Descontos, Receitas, Preço custo e vendas, entre outros
utl_email	Email do utilizador
created_date	Data em que a informação foi inserida na tabela CQA_T_SERVICOS
modified_date	Data da ultima atualização da informação
utl_id	Identificação única da área
dom_id	Identificação única do dominio do CQ

TABELA A.7: Descrição da tabela `CQA_T_SERVICOS`

Atributo	Descrição
dom_id	Identificação única do dominio do CQ
dom_nome	Descrição do dominio; ex: Loja, Cadeia de abastecimento, Contabilidade, Gestão, Recursos Humanos, entre outros
created_date	Data em que a informação foi inserida na tabela CQA_T_DOMINIOS
modified_date	Data da ultima atualização da informação

TABELA A.8: Descrição da tabela `CQA_T_DOMINIOS`

Apêndice B

Relativo às *queries*

```
SELECT
    CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(ES.EXCEPTIONRECORDSCREATEDON, 'yyyy-MM-dd'), 'yyyyMMdd') AS DECIMAL(8)) AS time_key,
    'IBM DQ' as origem,

    REGEXP_REPLACE(CASE
    WHEN ID.STRINGVALUE IS NULL THEN 'novalue'
    ELSE LOWER(ID.STRINGVALUE)
    END, '[\\n\\r\\|]', ' ') AS cq_id,

    REGEXP_REPLACE(CASE
    WHEN SPLIT(ES.Shortdescription, ' RuleDescription:')[1] = ES.Shortdescription THEN
        ES.Shortdescription
    ELSE
        SPLIT(ES.Shortdescription, ':')[-1]
    END, '[\\n\\r\\|]', ' ') AS cq_name,
    'Ativo' AS status,
    'DQ' AS cq_type,
    REGEXP_REPLACE(CASE
    WHEN LOWER(CQT.STRINGVALUE) LIKE '%business%' THEN 'Business Data Rule Definition'
    WHEN LOWER(CQT.STRINGVALUE) LIKE '%technical%' THEN 'Technical Data Rule Definition'
    ELSE NULL
    END, '[\\n\\r\\|]', ' ') AS data_rule,

    REGEXP_REPLACE(CASE
    WHEN CQD.STRINGVALUE IS NULL
    THEN 'novalue' ELSE LOWER(CQD.STRINGVALUE)
    END, '[\\n\\r\\|]', ' ') AS cq_rule,

    REGEXP_REPLACE(CASE
        WHEN INSTR(ES.Shortdescription, ' RuleDescription:') = 0 THEN ES.Shortdescription
        ELSE
            SUBSTR(ES.Shortdescription, LENGTH(ES.Shortdescription) - INSTR(REVERSE(ES.Shortdescription), ':') + 2)
    END, '[\\n\\r\\|]', ' ') AS short_description,
    REGEXP_REPLACE(CASE
    WHEN upper(CQD.STRINGVALUE) = 'CONSISTENCY' THEN
    CASE
    WHEN S1.STRINGVALUE LIKE '%;% vs. %'
    THEN lower(SUBSTR(S1.STRINGVALUE, INSTR(S1.STRINGVALUE, ';') + 2, INSTR(S1.STRINGVALUE, ' vs. ') - (INSTR(S1.STRINGVALUE, ';') + 2)))
    WHEN S1.STRINGVALUE LIKE '%vs %'
    THEN lower(SUBSTR(S1.STRINGVALUE, INSTR(s1.STRINGVALUE, 'between') + 8, INSTR(s1.STRINGVALUE, ' vs') - (INSTR(S1.STRINGVALUE, 'between') + 8)))
    WHEN S1.STRINGVALUE LIKE '%between%and%'
    THEN lower(SUBSTR(S1.STRINGVALUE, INSTR(S1.STRINGVALUE, 'between') + 8, INSTR(SUBSTR(S1.STRINGVALUE, INSTR(S1.STRINGVALUE, 'between') + 8), 'and') - 1))
    ELSE NULL
        END
    ELSE NULL
    END, '[\\n\\r\\|]', ' ') AS sourcel,
```

FIGURA B.1: Query *working_xmeta* em DataStage - Parte 1

```
REGEXP_REPLACE(CASE
    WHEN UPPER(CQD.STRINGVALUE) = 'CONSISTENCY' THEN (
        CASE
        WHEN S1.STRINGVALUE LIKE '% vs. %' THEN
            lower(SPLIT(SUBSTR(S1.STRINGVALUE, INSTR(S1.STRINGVALUE, 'vs.') + 4), ' ')[0])

        WHEN S1.STRINGVALUE LIKE '%between % vs %' THEN
        lower(SUBSTR(S1.STRINGVALUE, INSTR(S1.STRINGVALUE, 'vs ')+3,
               INSTR(SUBSTR(S1.STRINGVALUE,INSTR(S1.STRINGVALUE,'vs ')+3),' ')-1))

        WHEN S1.STRINGVALUE LIKE '%between % and %' THEN
        lower(SPLIT(SPLIT(SUBSTR(S1.STRINGVALUE, INSTR(S1.STRINGVALUE,'between')+8),' and ')[1],' ')[0])
        ELSE NULL
        END)
END, '[\\n\\r\\|]', ' ') AS source2,
'Information Quality' AS domain,
REGEXP_REPLACE(CASE
WHEN PA.STRINGVALUE IS NULL
THEN 'novalue'
ELSE LOWER(PA.STRINGVALUE)
END, '[\\n\\r\\|]', ' ') AS area,
REGEXP_REPLACE(CASE
WHEN TN.STRINGVALUE IS NULL THEN 'novalue'
ELSE LOWER(TN.STRINGVALUE)
END, '[\\n\\r\\|]', ' ') AS table_name,
REGEXP_REPLACE(CASE
WHEN PP.STRINGVALUE IS NULL
THEN 'novalue'
ELSE LOWER(PP.STRINGVALUE)
END, '[\\n\\r\\|]', ' ') AS project_name,
REGEXP_REPLACE(CASE
WHEN OWN.STRINGVALUE IS NULL THEN 'novalue'
ELSE LOWER(OWN.STRINGVALUE)
END, '[\\n\\r\\|]', ' ') AS owner,
null as respons_name ,
NULL AS service_name,
'N' As fecho_contas,
'Email' AS alert_type,
REGEXP_REPLACE(CASE
            WHEN SDT.STRINGVALUE IS NULL
            THEN 'novalue'
            ELSE LOWER(SDT.STRINGVALUE)
END, '[\\n\\r\\|]', ' ') AS SDTicket,
null AS threshold,
REGEXP_REPLACE(CASE
                WHEN S.STRINGVALUE IS NULL
                THEN 'novalue'
                ELSE LOWER(S.STRINGVALUE)
END, '[\\n\\r\\|]', ' ') AS stage,
```

FIGURA B.2: Query *working_xmeta* em DataStage - Parte 2

```
REGEXP_REPLACE(CASE WHEN ES.EXCEPTIONRECORDSCOUNT  > 0 THEN 'NOK' ELSE 'OK' END, '[\\n\\r\\|]', ' ') AS cq_result,
CASE ES.EXCEPTIONRECORDSCOUNT WHEN 0 THEN 1 ELSE 0  END AS success_cq_nrs,
CASE WHEN ES.EXCEPTIONRECORDSCOUNT  > 0 THEN 1 ELSE 0  END AS unsuccess_cq_nrs,
ES.EXCEPTIONRECORDSCOUNT AS alarms_nrs,
CASE
    WHEN ES.EXCEPTIONRECORDSCOUNT > (CASE WHEN TR.STRINGVALUE IS NULL THEN 1 ELSE CAST(TR.STRINGVALUE AS INT) END)
    THEN ES.EXCEPTIONRECORDSCOUNT
    ELSE CASE WHEN TR.STRINGVALUE IS NULL THEN 1 ELSE CAST(TR.STRINGVALUE AS INT) END
END AS total_records,
ES.EXCEPTIONRECORDSCREATEDON AS create_date,
ES.EXCEPTIONRECORDSCREATEDON AS last_updt_date ,
CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(ES.EXCEPTIONRECORDSCREATEDON ,'yyyy-MM-dd'),'yyyy') AS DECIMAL(8)) AS year_key

FROM xmeta_emexceptionset ES

LEFT OUTER JOIN xmeta_emsetproperty ID
ON lower(ID.PROPERTYNAME)='dq_id'
AND ID.EXCEPTIONSETRID = ES.RID
LEFT OUTER JOIN xmeta_emsetproperty CQT
ON lower(CQT.PROPERTYNAME)='dqtype'
AND CQT.EXCEPTIONSETRID = ES.RID
LEFT OUTER JOIN xmeta_emsetproperty CQD
ON lower(CQD.PROPERTYNAME)='dqdimension'
AND CQD.EXCEPTIONSETRID = ES.RID
LEFT OUTER JOIN xmeta_emsetproperty TN
ON lower(TN.PROPERTYNAME)='table'
AND TN.EXCEPTIONSETRID = ES.RID
LEFT OUTER JOIN xmeta_emsetproperty OWN
ON lower(OWN.PROPERTYNAME)='owner'
AND OWN.EXCEPTIONSETRID = ES.RID
LEFT OUTER JOIN xmeta_emsetproperty SDT
ON lower(SDT.PROPERTYNAME)='sdticket'
AND SDT.EXCEPTIONSETRID = ES.RID
LEFT OUTER JOIN xmeta_emsetproperty PA
ON lower(PA.PROPERTYNAME)='area'
AND PA.EXCEPTIONSETRID = ES.RID
LEFT OUTER JOIN xmeta_emsetproperty TR
ON lower(TR.PROPERTYNAME)='totalrec'
AND TR.EXCEPTIONSETRID = ES.RID
LEFT OUTER JOIN xmeta_emsetproperty PP
ON lower(PP.PROPERTYNAME)='project'
AND PP.EXCEPTIONSETRID = ES.RID
LEFT OUTER JOIN xmeta_emsetproperty S1
ON  S1.EXCEPTIONSETRID = ES.RID
AND UPPER(S1.PROPERTYNAME) = 'RULEDESCRIPTION'
LEFT OUTER JOIN xmeta_emsetproperty S
ON lower(S.PROPERTYNAME)='stage'
AND S.EXCEPTIONSETRID = ES.RID
WHERE ES.EXCEPTIONRECORDSCREATEDON >= FROM_UNIXTIME(UNIX_TIMESTAMP('2020-01-01', 'yyyy-MM-dd'))
AND ES.EXCEPTIONRECORDSCREATEDON >= #DATA#
```

FIGURA B.3: Query *working_xmeta* em DataStage - Parte 3

```
SELECT
    TO_NUMBER(TO_CHAR(A.CREATED_DATE, 'YYYYMMDD')) AS time_key,
    'Raid' AS origem,
    A.CTR_ID AS  cq_id,
    D.CTR_NOME AS cq_name,
    DECODE(A.STATUS_CONTROLO,'S','Ativo','E','Em Estabilização','Inativo') AS status,
    DECODE(D.CTR_TIPO_FLUXO,'2','MON','CQ') AS cq_type,
    null AS data_rule,
    'consistency' AS cq_rule,
    D.CTR_DESCRICAO AS short_description,
    D.CTR_APLICACAO_ORIGEM AS source1,
    D.CTR_APLICACAO_DESTINO AS source2,
    CASE
        WHEN A.DOM_ID is null THEN F.DOM_NOME
        ELSE FF.DOM_NOME
    END AS domain,
    B.AOW_NOME AS area,
    null AS table_name,
    null AS project_name,
    B.AOW_NOME AS owner,
    C.UTL_NOME AS respons_name,
    E.SRV_NOME AS service_name,
    D.CTR_FECHO_CONTAS fecho_contas,
    CASE
        WHEN D.CTR_TP_ALERTA = 'E' THEN 'Email'
        WHEN D.CTR_TP_ALERTA = 'R' THEN 'Registo de ITSM'
        ELSE NULL
    END AS alert_type,
    null as SDTicket,
    null as threshold,
    DECODE(A.STATUS_CONTROLO,'S','Ativo','E','Em Estabilização','Inativo') AS stage,
    CASE
        WHEN DECODE(A.AMOUNT,0,1,0) = '1' then 'OK'
        ELSE 'NOT OK'
    END AS cq_result,
    DECODE(A.AMOUNT,0,1,0) AS success_cq_nrs,
    DECODE(A.AMOUNT,0,0,1) AS unsuccess_cq_nrs,
    A.AMOUNT AS alarms_nrs,
    CASE
        WHEN D.CTR_TIPO_FLUXO = 2 THEN A.AMOUNT
        WHEN D.CTR_TIPO_FLUXO = 1 AND D.CTR_SENTIDO_VALIDACAO = 2 THEN A.AMOUNT_SOURCE
        WHEN D.CTR_TIPO_FLUXO = 1 AND D.CTR_SENTIDO_VALIDACAO = 1 AND A.AMOUNT_SOURCE >= A.AMOUNT_DESTINY THEN A.AMOUNT_SOURCE
        WHEN D.CTR_TIPO_FLUXO = 1 AND D.CTR_SENTIDO_VALIDACAO = 1 AND A.AMOUNT_SOURCE < A.AMOUNT_DESTINY THEN A.AMOUNT_DESTINY
        ELSE NULL
    END AS total_records,
    CAST(A.CREATED_DATE AS TIMESTAMP) AS create_date,

    A.CREATED_DATE as last_updt_date,

    TO_CHAR(A.CREATED_DATE, 'YYYY') AS year_key

FROM CQA_T_CASES A

INNER JOIN CQA_T_CONTROLOS D
ON A.CTR_ID = D.CTR_ID

INNER JOIN CQA_T_AREAS_OWNER B
ON D.AOW_ID = B.AOW_ID

INNER JOIN CQA_T_UTILIZADORES C
ON D.UTL_ID = C.UTL_ID

INNER JOIN CQA_T_SERVICOS E
ON D.SRV_ID = E.SRV_ID

INNER JOIN CQA_T_DOMINIOS F
ON D.DOM_ID = F.DOM_ID

LEFT JOIN CQA_T_DOMINIOS FF
ON A.DOM_ID = FF.DOM_ID

WHERE  trunc(A.CREATED_DATE) > to_date('2020.01.01','YYYY.MM.DD')
```

FIGURA B.4: Query *working_raid* em DataStage