

Universidade de Trás-os-Montes e Alto Douro

## **Big Data em ambiente Cloud**

### **Análise da Migração para serviços da Cloud**

Estágio de Mestrado em Engenharia Informática

Cátia Alexandra Mourão Queirós

*Sob orientação do Professor Doutor Frederico Augusto dos Santos Branco e do  
Engenheiro Luís Guilherme Martins Ferreira Coutinho*



Vila Real, dezembro de 2022



Relatório de Estágio apresentada por Cátia Alexandra Mourão Queirós à Universidade de Trás-os-Montes e Alto Douro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática, sob a orientação do Prof. Doutor Frederico Augusto dos Santos Branco, Professor Auxiliar do Departamento de Engenharias da Escola de Ciências e Tecnologia da Universidade de Trás-os-Montes e Alto Douro e do Engenheiro Luís Guilherme Martins Ferreira Coutinho, Mestre em Engenharia Eletrotécnica e de Computadores pela Universidade de Trás-os-Montes e Alto Douro e atual programador sénior de *back-end*.





Este trabalho foi escrito ao abrigo do novo Acordo Ortográfico.



# Ag

## **Agradecimentos**

O presente relatório marca o fim de uma mais uma etapa da minha vida com a conclusão do meu percurso académico e marca também o início de uma carreira no mercado de trabalho. No decorrer deste percurso foram várias as pessoas que se cruzaram no meu caminho e que de forma positiva o marcaram ao longo destes anos.

Um agradecimento muito especial à minha mãe, e à minha irmã que sempre me acompanharam neste percurso e sempre me apoiaram, sem elas não seria possível a concretização desta etapa. Um agradecimento enorme e muito sentido ao meu pai, que me ensinou todos os valores e os quais aplico diariamente e me levaram até onde me encontro hoje. Obrigada aos três que são a minha principal motivação.

Ao meu namorado, que sempre me apoiou e ajudou nas alturas mais difíceis e me motivou sempre a concluir todos os meus objetivos, sendo este um deles.

Ao Professor Doutor Frederico Augusto dos Santos Branco, pelo acompanhamento e orientação ao longo de todo o percurso em que sempre demonstrou uma enorme disponibilidade.

À empresa Openvia, pela oportunidade que me foi dada, pelo apoio diário e pela confiança depositada nos últimos meses, em particular ao Team Leader da equipa de Suporte, o Diamantino Abobeira, por toda a disponibilidade, conhecimento passado e confiança. Um agradecimento também ao arquiteto de software, o Ricardo Vilela, por todo o conhecimento passado e pela oportunidade.

Aos restantes elementos da empresa com os quais interagi no decorrer deste projeto, e também os elementos externos, obrigada pelo conhecimento transmitido e pela disponibilidade em todos os momentos.

A todas as restantes pessoas que ao longo destes anos se cruzaram comigo e contribuíram positivamente para o meu percurso.

# Jú

## Júri

Os membros do Júri recomendam à Universidade de Trás-os-Montes e Alto Douro a aceitação da dissertação intitulada “**Big data em ambiente de cloud**” realizada por **Cátia Alexandra Mourão Queirós** para satisfação parcial dos requisitos do grau de **Mestre em Engenharia Informática**.

**Dezembro de 2022**

Presidente: **Professor Doutor Luís Filipe Leite Barbosa**  
Professor Auxiliar da Escola de Ciências e Tecnologia da  
Universidade de Trás-Os-Montes e Alto Douro

Vogais do Júri: **Professor Doutor José Luís Mota Pereira**  
Professor Auxiliar da Escola de Engenharia da Universidade do  
Minho

**Professor Doutor Arsénio Monteiro dos Reis**  
Professor Auxiliar da Escola de Ciências e tecnologias da  
Universidade de Trás-os-Montes e Alto Douro

**Professor Doutor Frederico Augusto dos Santos Branco**  
Professor Auxiliar do Departamento de Engenharias da Escola de  
Ciências e Tecnologia da Universidade de Trás-os-Montes e Alto  
Douro

# Re

## **Resumo**

Os sistemas de *Big Data* são cada vez mais uma ferramenta importante para o negócio da era digital, dada a enorme quantidade de dados que é gerada nos dias de hoje e a importância que estes representam para a sociedade também. Dado o crescimento exponencial da área da mobilidade e de forma a melhorar a qualidade dos produtos, surge a necessidade de criar sistemas coesos e eficazes capazes de extrair, transformar e analisar os dados com a máxima eficácia possível. Daí, ser necessário procurar as mais recentes tecnologias, verificar o que melhor se enquadra nas necessidades de cada um seguindo sempre a filosofia da entidade. O objetivo deste estágio focou-se na análise do sistema atualmente utilizado, acompanhamento e gestão de uma possível alteração para tornar ainda melhor o sistema da empresa, sistema esse que é responsável pelo tratamento de um milhão de dados , todos os dias, e também pelo estudo e desenvolvimento dos *dashboards* que seriam utilizadas pelos diferentes clientes da empresa. Foram analisadas várias métricas a fim de verificar se a migração seria positiva/possível ou não para a empresa. O estágio teve a duração de nove meses e foi realizado na empresa Openvia Mobility, empresa tecnológica do grupo Globalvia, com o apoio do arquiteto de software e o *team leader* responsável pelo sistema atual, bem como uma equipa externa com experiência na área.





## **Abstract**

Big Data systems are increasingly an essential tool for business in the digital age, given the enormous amount of data generated today and the importance that data has in society. Given the exponential growth in mobility and to improve the quality of products, the need arises to create cohesive and effective systems capable of extracting, transforming, and analyzing the data as effectively as possible. Hence, it is necessary to search for the latest technologies, check what best fits the needs of each one, and always follow the entity's philosophy. The objective of this internship focused on the analysis of the system currently used, monitoring, and managing a possible change to make the company's system even better, a system that is responsible for processing about a million of data every day and the study and development of visualizations that different customers of the company would use. Several metrics were analyzed to verify if the migration would be positive/possible or not for the company. The internship lasted nine months and was carried out at Openvia Mobility, a technology company of the Globalvia group, with the support of the software architect and the team leader responsible for the current system, as well as an external team with experience in the area.



# Pc

## **Palavras-chave**

- Análise de dados;
- Big Data;
- Migração.



# Kw

## **Keywords**

- Data analysis;
- Big Data;
- Migration.



*"The only impossible  
journey is the one  
you never begin"*

**Tony Robbins**



# ÍG

## Índice Geral

Agradecimentos .....	viii
<b>Júri</b> .....	ix
Resumo .....	x
Abstract.....	xii
Palavras-chave .....	xiv
Keywords.....	xvi
Índice de tabelas .....	xxii
Índice de figuras .....	xxiii
1. Introdução.....	1
1.1. Enquadramento .....	3
1.2. Planeamento e objetivos .....	4
1.3. A empresa e o seu produto .....	5
1.4. Organização do documento .....	7
2. Metodologia e ferramentas de trabalho .....	9
2.1. Metodologia de trabalho .....	9
2.2. Planeamento do trabalho .....	10
2.3. Ferramentas de trabalho.....	12
3. Enquadramento Teórico .....	15
3.1. <i>Big Data</i> .....	15
3.2. <i>Cloud Computing</i> .....	17

3.3.	Arquitetura <i>serverless</i> .....	18
4.	Sistema <i>As-Is</i> .....	20212121
4.1.	Arquitetura.....	20212121
4.2.	Plano de escalonamento.....	28292929
4.3.	Plano de custos .....	32333333
4.4.	Análise do sistema <i>As-Is</i> .....	33343434
5.	Prova de conceitos.....	35363636
6.	Sistema <i>To-Be</i> .....	38393939
6.1.	Arquitetura.....	38393939
6.2.	Plano de escalonamento.....	42434343
6.3.	Plano de custos .....	44464646
6.4.	Análise do sistema <i>To-Be</i> .....	45474747
7.	Conclusões .....	47494949
	Referências bibliográficas .....	52555555



## **Índice de tabelas**

Tabela 1: Plano de escalonamento do sistema <i>As-Is</i> .....	30313131
Tabela 2: Custos do sistema <i>As-Is</i> .....	32333333
Tabela 3: Etapas e intervenientes da prova de conceitos .....	36373737
Tabela 4: Plano de escalonamento do sistema <i>To-Be</i> .....	43444444
Tabela 5: Custos do sistema <i>To-Be</i> .....	45464646
Tabela 6: Dados de comparação dos dois sistemas .....	47494949



## Índice de figuras

Figura 1: Gráfico de Gantt.....	5
Figura 2: Presença da empresa a nível mundial .....	6
Figura 3: <i>Openvia Tolling Backoffice</i> .....	6
Figura 4: Passos realizados em cada etapa do projeto.....	11
Figura 5: Canal de comunicação do POC.....	13
Figura 6: Tipos de Cloud.....	18
Figura 7: Arquitetura do sistema <i>As-Is</i> .....	21222222
Figura 8: Modelo em estrela.....	24252525
Figura 9: Relatório criado com a ferramenta Power BI .....	26272727
Figura 10: Fluxo de tráfego por hora de um determinado cliente .....	27282828
Figura 11: Utilização de CPU.....	34353535
Figura 12: Arquitetura do sistema <i>To-Be</i> .....	39404040

## Lista de abreviaturas

<b>Abreviatura</b>	<b>Significado(s)</b>
AD	<i>Active Directory</i>
AWS	<i>Amazon Web Services</i>
BI	<i>Business Intelligence</i>
CPU	<i>Central Process Unit</i>
DB	<i>Database</i>
DNS	<i>Domain Name Server</i>
EC2	<i>Elastic Cloud Computing</i>
ETL	<i>Extract, Load and Transform</i>
LDA	<i>Lightweight Directory Access Protocol</i>
ODB	<i>Open Database Connectivity</i>
OLA	<i>Online Analytical Processing</i>
POC	<i>Proof of concept</i>
S3	<i>Simple storage service</i>
SNS	<i>Simple Notification Service</i>
SQS	<i>Simple Queue Service</i>
SSAS	<i>SQL Server Analysis Service</i>
VPC	<i>Virtual Private Cloud</i>
VPN	<i>Virtual Private Network</i>
SaaS	<i>Software as a Service</i>

# 1

## 1. Introdução

Atualmente as empresas e as pessoas têm à sua disposição uma elevada quantidade de dados, o que pode criar dificuldades em maximizar os proveitos que podem resultar da análise da riqueza dos mesmos(Côrte-Real, 2022). Por vezes não têm o conhecimento ou até mesmo os recursos tanto tecnológicos como financeiros para processar os dados da forma mais correta, perdendo assim capacidade para suportar decisões de carácter operacional (Abiodun et al., 2021), tático ou estratégico, mas também podem perder a capacidade de apresentar as melhores soluções aos clientes, prejudicando diretamente a qualidade do serviço que estão a oferecer (Côrte-Real, 2022)

Assim, verifica-se a necessidade de soluções que permitam a recolha e o processamento da forma mais eficaz e rápida possível, que consiga lidar com o enorme fluxo de dados recebido, mas também ferramentas que tenham a capacidade de analisar a quantidade de dados existentes, em qualquer lugar e momento, pois outro ponto crucial é o rápido e fácil acesso aos dados e o valor que conseguimos extrair das análises(Marr, 2015). Devido à grande diversidade de dados surge assim o termo *Big Data* que engloba tanto os diferentes tipos de dados como as suas origens(Abiodun et al., 2021).

O termo *Big Data*, engloba múltiplos conceitos (Rossi & Hirama, 2022), gerando por vezes alguma confusão e mal-entendidos. Dada a elevada quantidade de dados, há obstáculos que surgem relacionados com a gestão dos dados e da forma mais rápida de os processar, analisar e visualizar(Côrte-Real, 2022), tais como: o espaço em disco; escalabilidade do software, quando o fluxo de dados aumenta; apresentação os dados em real-time; apresentação dos dados de forma diversificada; custos das ferramentas, licenças dos serviços e custos de hardware(Abiodun et al., 2021).

Desde o surgimento do conceito *Big Data*, surgiram também três V's associados a este conceito, velocidade, variedade e volume(Côrte-Real, 2022). A inovação no conceito de *Big Data* não está apenas na necessidade de tratar muitos dados (volume), nem só em fazê-lo da

forma mais rápida (velocidade) mas foca-se também na necessidade de tratar diversos tipos de dados de diversas fontes (dados estruturados e dados não estruturados)(Tardio et al., 2020). Desta forma surge a necessidade de uma arquitetura que permita dar resposta a todas estas necessidades.

A arquitetura do sistema deve permitir uma integração de dados provenientes de diferentes fontes, sendo necessário também que sejam armazenados num repositório, para permitir depois a sua análise(Rosini Filho et al., 2020). Para se poder tomar decisões de forma mais rápida, sustentada e eficiente, é necessário que a arquitetura permita também análises de dados dinâmicas, em qualquer lugar e em qualquer momento, e tudo isto pode ser obtido utilizando um ambiente de *Cloud Computing*(Berisha et al., 2022)

O grupo Openvia, enquanto operador e gestor de múltiplas infraestruturas rodoviárias a nível mundial, controla várias portagens em diversas autoestradas por todo o mundo, com recurso a dois sistemas que se interligam e comunicam entre si, gerindo uma enorme quantidade de dados. O sistema de *Business Intelligence* (BI) da empresa é responsável por receber os dados registados no software da empresa responsável pelo armazenamento e gestão de toda a informação vinda das vias, fazer o tratamento dos mesmos e disponibilizá-los para que cada cliente consiga dissecar os seus valores e fazer as suas análises, mas também é responsável pelo tratamento de dados que permitiram internamente analisar possíveis estratégias de negócio e apoiar nas tomadas de decisão e, em alguns casos, detetar problemas em ambos os sistemas.

Dada a diversidade de serviços da Amazon Web Services (AWS) que a empresa utiliza e sendo a AWS o seu fornecedor de serviços da *Cloud* e também um dos melhores fornecedores de serviços da *Cloud* à data segundo(Rahman & Rana, 2021), suscitou o interesse por parte da Openvia em explorar serviços capazes de responder às necessidades dentro da área de *Big Data* e migrar a parte do sistema de BI para os serviços desta entidade.

Assim o objetivo deste trabalho é analisar a arquitetura do sistema atual da empresa e posteriormente realizar uma prova de conceitos, analisando os seus resultados e, caso o resultado seja positivo e satisfatório para a empresa, proceder à alteração da arquitetura do sistema de *Big Data*, mudando para uma arquitetura *serverless*, utilizando apenas serviços da AWS, pois neste momento o sistema *As-Is* utiliza tanto serviços da AWS como da Microsoft, o que dificulta o seu suporte.

## 1.1. Enquadramento

O presente documento relata a experiência vivida no ambiente empresarial enquanto se desenvolvia o projeto, bem como todo o conhecimento adquirido e trabalho desenvolvido, no decorrer um estágio na área de Engenharia Informática, com a duração de nove meses, para a obtenção do grau académico de Mestre em Engenharia Informática pela Universidade de Trás-Os-Montes e Alto Douro (UTAD). Todo o trabalho desenvolvido e descrito neste documento foi realizado na sede da Openvia Mobility, empresa do grupo tecnológico pertencente ao grupo de empresas Globalvia, com suporte e colaboração de uma equipa externa que colaborou com a empresa e também com o *Support Leader* da empresa, responsável pelo departamento de BI da empresa, sistema que faz o tratamento dos dados enviados e processados pelo software *Openvia Tolling Backoffice*, sistema este que é um *BackOffice* que permite gerir e visualizar as transações recebidas da via em tempo real, entre outras funcionalidades, que será apresentado mais detalhadamente na secção 1.3.

Para cumprir o objetivo do projeto, ou seja, a análise de uma possível migração do sistema atual para serviços da *Cloud* do principal fornecedor de serviços da *Cloud* da empresa, construindo assim uma arquitetura *serverless*, foi desenvolvido um *Proof of Concept* (POC) recorrendo a uma equipa externa com experiência na área e que tem parceria com a AWS, a fim de apresentarem uma solução viável, e por esse motivo, como a equipa não se encontra perto da sede da empresa foi necessário recorrer a tecnologias que permitissem a comunicação e acompanhamento do PoC. Para tal, utilizamos o Microsoft Teams, o Slack e o AWS Chime, sendo este último onde se efetuavam as reuniões de acompanhamento e apresentação das diferentes etapas da prova de conceitos. As restantes foram utilizadas para reuniões informais, para manter o contacto, saber qual o ponto de situação da prova de conceitos e também esclarecimento de dúvidas que a equipa pudesse ter em relação ao modelo de negócio e a estrutura dos dados.

A Openvia possui um sistema que permite a cada concessão gerir as suas transações, tanto a nível operacional como comercial. A vertente operacional do sistema permite operar e visualizar as transações que estão a receber da via, bem como o estado em que se encontram essas transações. Por outro lado, a vertente comercial permite gerir a cobrança dessas mesmas transações. Assim que a informação é enviada para o *BackOffice* ou é atualizada no mesmo, esta é enviada também para o sistema de BI onde é feito o armazenamento, tratamento, leitura, análise e visualização dos dados.

O sistema de BI que é oferecido aos clientes, é utilizado maioritariamente para efetuar análises e relatórios recorrendo a um cubo Online Analytical Processing (OLAP), utilizando a ferramenta Excel e, mais recentemente, o Microsoft Power BI. Diariamente, conectam-se aos cubos mais de 30 utilizadores das diferentes concessões para realizarem as suas análises. É notável a vontade de evoluir e apresentar da melhor forma possível as análises aos clientes e denota-se uma vontade destes em crescer nesta vertente, daí se ter evoluído para o Power BI, onde é possível criar *dashboards* interativos que podem auxiliar as tomadas de decisão e identificação de problemas nas vias.

Tendo em conta que grande parte dos sistemas da empresa estão construídos com base numa arquitetura de microserviços e arquitetura *serverless*, estes sistemas são constituídos maioritariamente por serviços que a AWS oferece, e deste modo surge a oportunidade e o interesse em explorar a vertente de *Big Data* que a AWS disponibiliza e uma possível migração dos serviços do sistema *As-Is* que atualmente se utilizam.

Assim, o objetivo principal do estágio descrito neste documento consiste na análise do sistema atual, das dificuldades que apresenta para a empresa, o acompanhamento da prova de conceitos e análise da mesma, e com base nos resultados obtidos determinar se a migração para os novos serviços é vantajosa, caso seja, apresentar as novas possibilidades e estudar visualizações vantajosas consoante determinados casos de uso, para as concessões e internamente.

## **1.2. Planeamento e objetivos**

O trabalho desenvolvido no decorrer do estágio visou a análise detalhada do sistema atual, o acompanhamento e gestão de uma prova de conceitos relativa a uma solução para um sistema de *Big Data* em ambiente de *Cloud*, utilizando uma arquitetura *serverless* e posterior análise de uma possível migração do sistema *As-Is* para o sistema *To-Be*. Assim, o objetivo principal definido pela Openvia Mobility passa pela gestão da prova de conceitos e verificar se efetivamente a migração do sistema é vantajosa não só para a empresa, mas também para os clientes que utilizam o sistema atual de *Big Data* da empresa.

Numa perspetiva geral, o projeto/estágio integra-se apenas na área de BI, ou seja, dedicado fundamentalmente à análise dos diferentes serviços de *Big Data* e qual será mais eficaz para as necessidades e casos de uso da empresa. No entanto, também se deve ter em consideração a necessidade de se procurar um novo serviço com garantias de escalabilidade

futura e integração com ferramentas atuais.. Neste sentido estudou-se a versatilidade da solução atual em vigor na empresa e da solução desenvolvida na prova de conceitos, bem como os planos de escalonamento e custos de cada opção. Com base numa análise inicial feita para desenvolvimento e conclusão do mesmo, foi desenhado um cronograma apresentado na Figura 1111Figura 11 com o objetivo de organizar as diversas etapas do projeto e ao mesmo tempo servir como indicador ao longo do desenvolvimento do estágio.

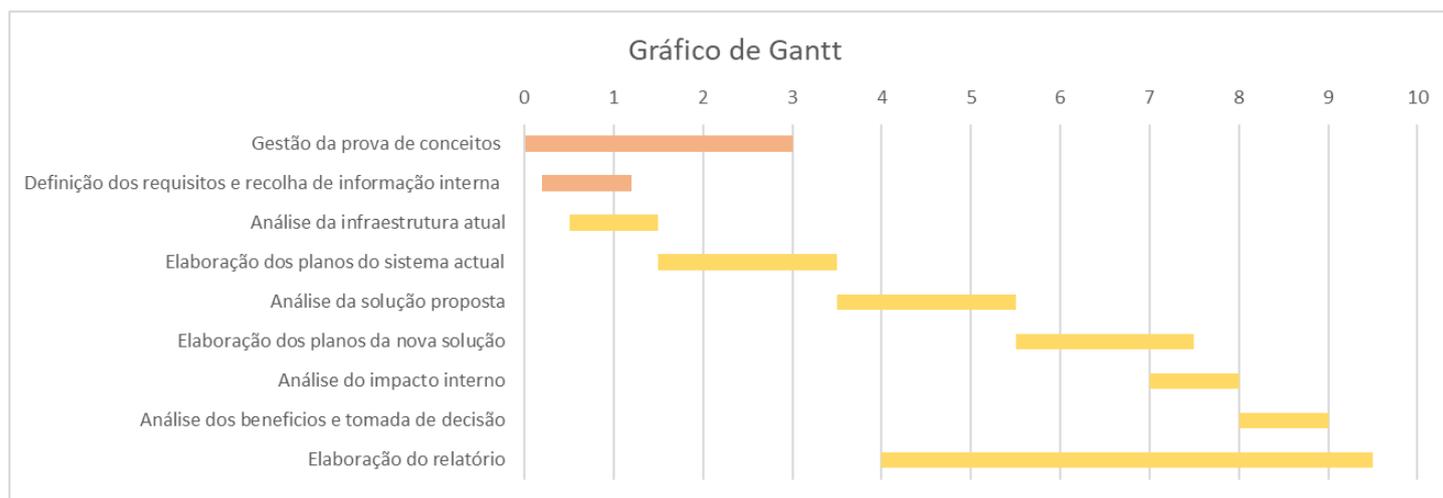


Figura 111111: Gráfico de Gantt

Relativamente à Figura 1111Figura 11, foram incorporados todos os passos para a concretização do presente relatório, onde também foi contemplado todo o período de elaboração do respetivo documento. Marcado a laranja estão as etapas referentes à prova de conceitos e a amarelo as etapas relativas a toda a análise realizada e os planos elaborados de cada um dos sistemas, nomeadamente, a solução atual e a solução resultante da prova de conceitos.

### 1.3. A empresa e o seu produto

A Openvia Mobility é uma empresa do âmbito tecnológico com foco na área da mobilidade com mais de 10 anos de experiência na gestão de infraestruturas e na transformação da mobilidade através das suas soluções inovadoras. Esta representa a inovação tecnológica do grupo Globalvia, um dos líderes mundiais na gestão de infraestruturas rodoviárias e responsável pela gestão de inúmeras concessões localizadas em diferentes países. Essa diversidade geográfica pode se visualizada na Figura 2, que representa os países onde a empresa está presente.

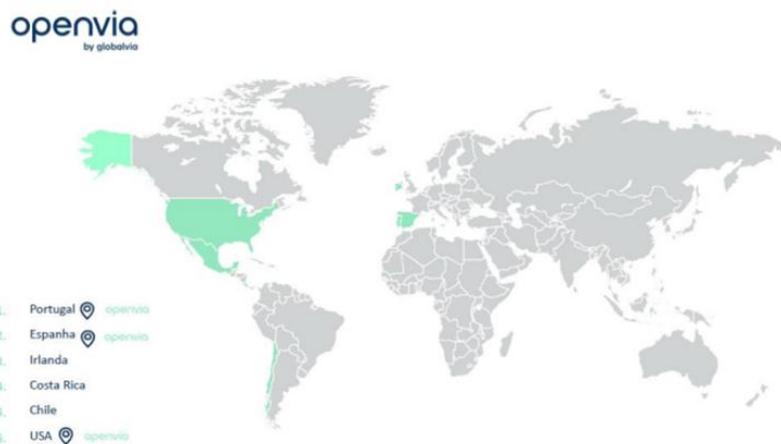


Figura 222222: Presença da empresa a nível mundial

A visão da empresa baseia-se em aumentar a mobilidade dos utilizadores através de infraestruturas digitalmente preparadas num ecossistema colaborativo. A missão passa por um mundo onde se pode mover sem limites, de forma fácil e sustentável, onde a mobilidade perfeita torna a sua vida mais fácil e ajuda na utilização mais eficiente do seu tempo.

Tendo como base esta visão, a empresa fornece diversos serviços aos seus clientes, sendo um deles o sistema de Big Data, o qual reflete os dados de um dos produtos internos, o *Tolling Backoffice*. Este produto consiste numa ferramenta de gestão de processos relativos a portagens, tais como, aquisição de dados, qualificação de portagens, consolidação de tráfego, revisão manual de OCR para reconhecimento de matrículas, construção de viagens, gestão de contas de clientes, faturação, entre outros.

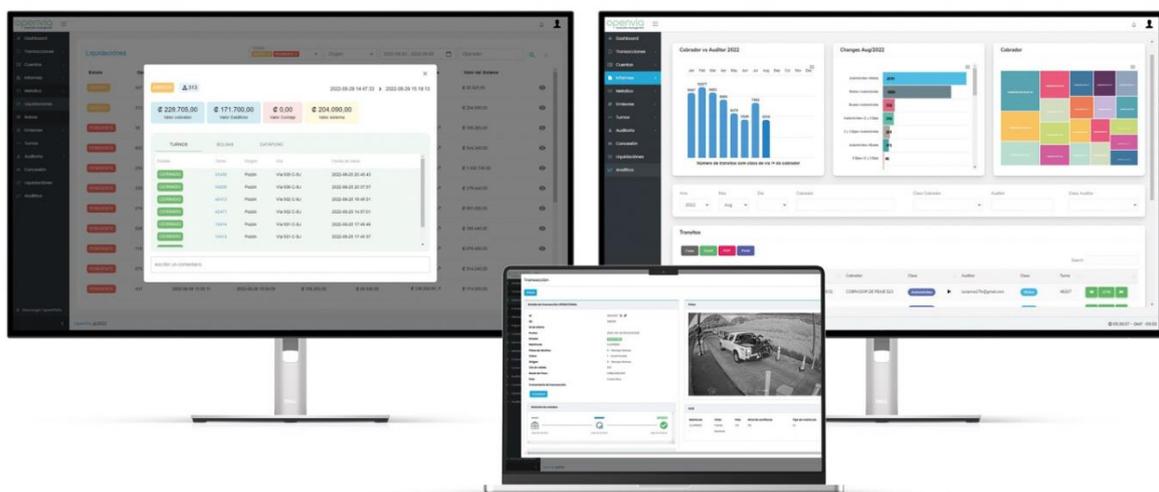


Figura 333333: Openvia Tolling Backoffice

Este software está dividido em duas vertentes, a vertente operacional e a vertente comercial. No que diz respeito à vertente operacional, este oferece aos operadores todas as funcionalidades para monitorizar e realizar auditorias a todas as transações provenientes dos pódicos. Quanto à vertente comercial, permite a automatização de qualquer fluxo de trabalho, desde a cobrança das portagens, passando por todas as suas fases.

A ferramenta em questão segue uma abordagem *Software as a Service* (SaaS), baseada numa arquitetura orientada ao software em *Cloud Computing*, com evolução contínua e as versões mais recentes.

Toda a informação que passa pelo backoffice, nomeadamente, as transações e os seus pagamentos, são utilizados no sistema de *Big Data* da empresa, a fim de cada cliente analisar os seus próprios dados.

#### **1.4. Organização do documento**

O presente documento visa refletir o trabalho desenvolvido durante o estágio, com a duração de nove meses, na empresa tecnológica Openvia Mobility. O objetivo do mesmo passou pela análise do sistema de BI atualmente utilizado na empresa e a análise da migração para novos serviços na *Cloud*, acompanhando a POC e posteriormente análise e elaboração da analítica para consumo interno e para os clientes. Todas as tarefas realizadas no âmbito deste projeto encontram-se aqui descritas e, de modo a tornar clara a visão sobre todo o processo, encontram-se descritas no documento desde a integração na empresa e comunicação com as entidades externas à análise de ambas as soluções e conclusões dos resultados. O documento encontra-se dividido em 7 secções, subdivididos apropriadamente.

No capítulo um, encontra-se descrita a apresentação do tema do estágio bem como em que se enquadra o projeto e como surgiu a necessidade de desenvolvimento do mesmo e também é apresentada a planificação das tarefas principais, no qual se recorreu a um diagrama de Gantt para clarificar o planeamento, é apresentado também como está organizado/estruturado o documento do relatório de estágio e por último, é apresentada a empresa onde decorreu o estágio.

O capítulo dois consiste na apresentação dos conceitos chave do projeto e a metodologia de trabalho utilizada na empresa. São apresentadas noções relacionadas com *Big Data* e *Cloud Computing*, e também será feita a apresentação da metodologia Agile, metodologia utilizada

pela empresa, bem como uma breve apresentação das tecnologias utilizadas no decorrer do projeto e, em alguns casos, são utilizadas pelos restantes membros das equipas da empresa.

O capítulo três, apresenta os conceitos teóricos abordados ao longo do relatório, tais como, *Big Data*, *Cloud Computing*, arquitetura serverless e *Data Warehouse*.

O capítulo quatro, tem como objetivo apresentar o sistema atualmente utilizado na empresa, dando a conhecer a sua arquitetura, os serviços que a constituem bem como as diferentes tecnologias. Também serão apresentados e descritos o plano de escalonamento e o plano de custos do sistema atual. E por último, será feita uma análise de todos os dados do sistema *As-Is*.

No capítulo cinco, será feita a apresentação do POC, como decorreu, quais os passos até à conclusão do mesmo, quais os resultados obtidos desta prova de conceitos.

O capítulo seis descreve a arquitetura da solução apresentada na prova de conceitos, o seu plano de escalonamento e o plano de custos do mesmo. Serão analisadas as vantagens que este sistema pode trazer em relação ao sistema que já existe, bem como as preocupações inerentes a esta alteração.

No capítulo sete serão discutidas as conclusões obtidas com base nos resultados, nas métricas escolhidas e casos de uso, visando identificar os pontos fortes e os pontos fracos das duas soluções defendidas ao longo do projeto. Será ainda apresentada uma visão futura para o projeto, bem como uma retrospectiva sobre o trabalho realizado.

# 2

## 2. Metodologia e ferramentas de trabalho

Nesta secção será apresentada a metodologia de trabalho utilizada por todos os colaboradores da empresa e no desenvolvimento dos produtos internos, que no decorrer deste projeto se teve como base todos os valores que a metodologia Agile defende. Será também apresentado o planeamento do trabalho e por último serão descritas todas as ferramentas utilizadas para a elaboração e extração de toda a informação aqui detalhada neste relatório.

### 2.1. Metodologia de trabalho

Habitualmente o desenvolvimento de um produto é uma tarefa complexa, na qual é necessário definir os requisitos funcionais do produto final e daí pode surgir o primeiro desafio na implementação de um produto novo (González Moyano et al., 2022)

Um processo comum de implementação de um novo produto assenta no modelo tradicional, o modelo em cascata, em que o desenvolvimento é visto como um conjunto de tarefas cuja execução é feita sequencialmente, ou seja, apenas se implementa o passo seguinte quando a etapa anterior estiver concluída na sua totalidade. Uma das desvantagens do modelo tradicional, é que tem como base um trabalho muito mais demorado e conseqüentemente mais pesado para a equipa, tal como (Rumetna et al., 2022) indica alguns dos passos que é necessário seguir ao adotar o modelo em cascata

- O plano do projeto deve ser desenvolvido com antecedência;
- Os requisitos do projeto têm de estar definidos, clarificados e acordados (contratos);
- Implementação do projeto que cumpra todos os requisitos escritos;
- Testar completamente o software e verificar se este satisfaz os requisitos.

Muitos projetos que seguem esta metodologia, deparam-se com diversos problemas, especialmente na manutenção e nas alterações baseadas nos pedidos dos utilizadores.

A Openvia Mobility, tem implementado em todos os seus projetos uma metodologia de trabalho que contraste com o modelo tradicional, a metodologia Agile. A metodologia *Agile* é uma metodologia que começa com uma fase inicial de planeamento e segue o caminho para a fase de implementação com interações iterativas e incrementais ao longo de todo o ciclo de vida do projeto (González Moyano et al., 2022). O objetivo inicial desta metodologia passaria por reduzir as despesas gerais dos processos de desenvolvimento de um produto com a capacidade de mudar ou adaptar o produto sem ariscar todo o produto. A metodologia adotada pela empresa está assente em quatro valores e doze princípios fundamentais, valores e princípios estes que são a base na orientação do processo de desenvolvimento. Os quatro valores são os seguintes (González Moyano et al., 2022):

- Indivíduos e interações acima de processos e ferramentas;
- Trabalhar o software acima de documentação;
- Colaboração com o cliente acima de negociação contractual;
- Possibilidade de mudar acima de seguir um plano delineado.

Sendo esta a metodologia utilizada internamente, aplicada no desenvolvimento de todos os produtos internos e tendo sempre presente os valores da metodologia no decorrer do trabalho, foi sob os valores desta metodologia que se orquestrou todo o processo de elaboração do POC, mantendo-se sempre o contacto entre equipas, definindo-se pequenas tarefas e dando sempre a possibilidade de modificar o plano inicial caso se verificasse ser mais vantajoso.

Em suma, no decorrer de todo o trabalho, estiveram presentes os princípios gerais da metodologia Agile apresentados anteriormente, mas no decorrer do projeto não foi utilizada a metodologia em específico, apenas se teve em conta os princípios da mesma.

## **2.2. Planeamento do trabalho**

A planificação do trabalho foi dividida em três etapas, gestão da prova de conceitos, análise do sistema de *Big Data* atual e análise da solução proveniente da prova de conceitos. Através da Figura 44, é possível verificar todas as etapas realizadas em cada uma das etapas.

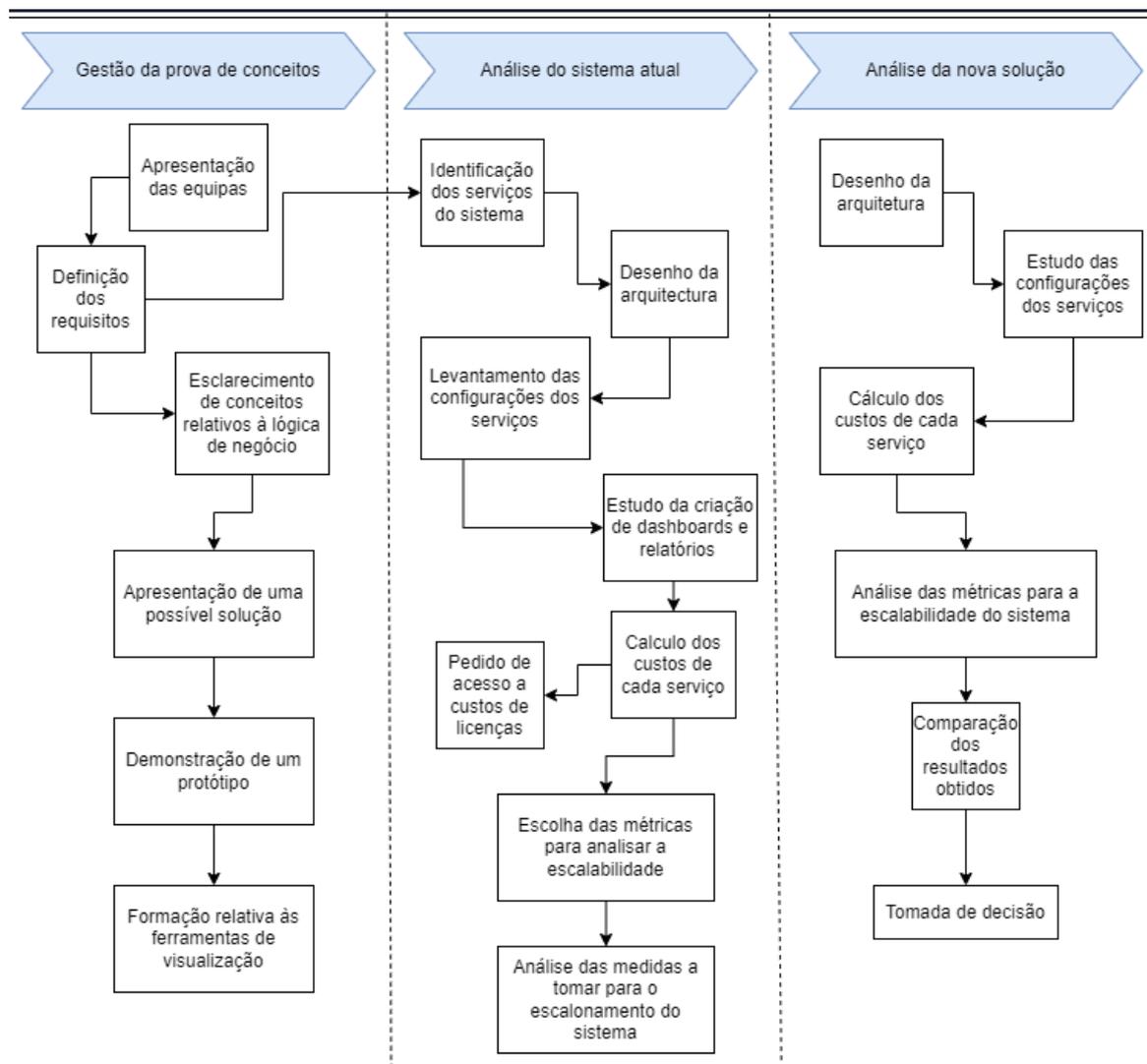


Figura 444444: Passos realizados em cada etapa do projeto

Inicialmente, para avançar com a prova de conceitos, foi necessário definir os requisitos que se pretendia que a nova solução cumprisse; após a elaboração dos requisitos, toda a informação foi transmitida à equipa responsável. Quando a equipa iniciou o desenvolvimento do PoC, dúvidas surgiram, como por exemplo, o significado de transação, viagem, como se identificava uma transação, etc. Todas estas dúvidas foram esclarecidas bem como qualquer informação extra que a equipa solicitou.

De seguida, após o estudo e elaboração da solução, foi apresentada a solução final e foi feita uma pequena demonstração utilizando os serviços em questão; também houve a oportunidade de realizar uma formação com especialistas AWS sobre a ferramenta AWS Quicksight.

Em paralelo ao desenvolvimento do PoC, enquanto se aguardava pela apresentação da solução, iniciou-se o estudo dos serviços que constituíam o sistema atual da empresa. Posteriormente a ter reunida toda a informação dos serviços em uso, desenhou-se a arquitetura do sistema.

Para a elaboração do plano de custos do sistema, foi necessário verificar todas as configurações de cada serviço, verificar o número de pedidos realizados mensalmente, verificar quanto espaço estava a ser utilizado e de seguida verificar os custos reais de todo o processo. Um dos últimos passos da análise do sistema atual, teve como objetivo definir métricas para estudar a escalabilidade, estas métricas foram utilizadas tanto para o sistema atual como para a nova solução. Logo após a definição, analisou-se as medidas a tomar para o plano de escalonamento.

Após a apresentação da solução na etapa “Gestão da prova de conceitos”, iniciou-se a análise da nova solução. Foi desenhada toda a arquitetura do possível novo sistema, bem como o estudo dos custos, tendo como base os mesmos valores utilizados para a análise do sistema atual da empresa, e também se analisou as medidas a tomar em relação ao escalonamento dos serviços.

Para finalizar, foram comparados os dois sistemas, os custos e a dificuldade de escalonamento de cada um, e com base nestes valores e na visão e perspetiva da empresa, foi tomada a decisão de qual solução adotar, manter a atual ou avançar com a alteração da arquitetura.

### **2.3. Ferramentas de trabalho**

Toda a equipa do Openvia tem objetivos bem definidos e o foco passa essencialmente pela eficiência do trabalho interno. Tendo em conta esta premissa, os pilares que sustentam a empresa assentam num esforço coletivo em que existe uma constante comunicação entre as diferentes equipas, mesmo sendo externas. De modo a conseguir agilizar todo o processo de acompanhamento do POC, a empresa disponibiliza um conjunto de tecnologias como suporte para a comunicação contínua entre equipas, sendo que toda a prova de conceitos foi acompanhada via email, sem recurso a ferramentas como o *Jira* ou o *Trello*, pois todas as tarefas relativas ao desenvolvimento da solução foram da responsabilidade da equipa externa.

## Slack

O Slack é uma plataforma de comunicação com foco na atividade empresarial e nas suas necessidades organizacionais. No cotidiano da empresa, esta ferramenta é utilizada diariamente por todos os elementos que interagem com as equipas. Esta plataforma permite a troca de mensagens diretas entre elementos, bem como a criação de canais de comunicação específicos de um tema, públicos ou privados, nos quais podem ser adicionados inúmeros elementos.

Na Figura 555555Figura 5555Figura 55, é possível ver o canal que foi criado para o acompanhamento informal da prova de conceitos. Neste canal foram adicionados todos os intervenientes internos e externos envolvidos no projeto.

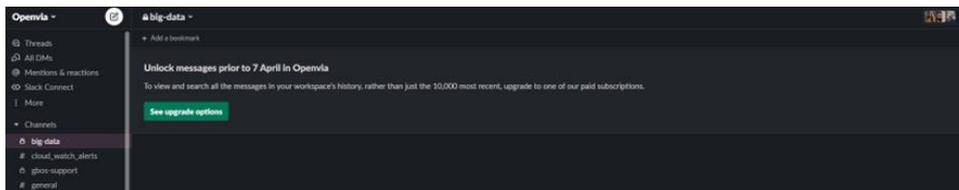


Figura 555555: Canal de comunicação do POC

Esta ferramenta permite inúmeras integrações com outros serviços, que internamente são utilizadas para alertar de possíveis problemas nos diferentes serviços que a empresa tem, neste caso específico, essas integrações não estão diretamente relacionadas com a prova de conceitos nem com o sistema atual de BI da empresa.

## Microsoft Teams

O Microsoft Teams é a ferramenta de comunicação oficial da empresa, sendo utilizada por todos os elementos internos da corporação. A principal limitação é que apenas permite marcar reuniões com os externos, não permitindo trocar mensagens diretamente com cada um (caso seja externo). Internamente, a ferramenta foi utilizada para esclarecimento de dúvidas com o Team Leader e o Arquiteto de Software da empresa e constante comunicação com os mesmos e também marcação de reuniões para esclarecimento de dúvidas que a equipa pudesse ter em relação ao sistema de *Big Data* pretendido.

### **Amazon Chine**

Uma vez que o principal fornecedor de serviços da *Cloud* da empresa é a AWS, e visto que eles são a ponte entre a empresa e a equipa externa responsável por elaborar o protótipo, todas as reuniões formais para apresentação de equipas, definição de objetivos e apresentação das diferentes etapas e, conseqüentemente, o resultado, foram realizadas através deste serviço da AWS.

### **Amazon Elasticsearch**

O serviço Amazon Elasticsearch permitiu investigar e reunir informação útil para a elaboração do plano de custos, uma vez que neste serviço são registadas todas as entradas no sistema, o que permitiu calcular o número de pedidos feitos para conseguir calcular o número de vezes que os serviços do Sistema *As-Is* eram utilizados.

### **Microsoft SQL Server**

A ferramenta Microsoft SQL Server permitiu complementar toda a informação recolhida através do Elasticsearch e também auxiliou no cálculo do espaço ocupado e o crescimento do mesmo.

### **Microsoft Excel**

Tendo em conta que a conexão aos cubos OLAP era realizada através de um ficheiro Excel, esta ferramenta foi bastante útil na realização da prova de conceitos, para facultar explicações da lógica que a empresa gostaria de ver implementadas no novo sistema e esclarecer conceitos.

### **Dbeaver**

O Dbeaver é uma ferramenta utilizada pela empresa que permite a conexão a diferentes bases de dados. Por ser uma ferramenta utilizada na empresa, foi a ferramenta escolhida para realizar a conexão à base de dados de BI, bem como para realizar qualquer pesquisa relevante sobre os dados para a análise do sistema atual.

### 3. Enquadramento Teórico

Este capítulo reserva-se à apresentação teórica de temas relevantes ligados ao projeto. Sendo assim, faz-se referência ao tema do Big Data, sua designação e caracterização, a *Cloud Computing* e a sua importância no mundo atual, à arquitetura *serverless*, sendo um dos requisitos da nova solução.

#### 3.1. *Big Data*

O termo *Big Data* é relativamente recente, começou a ser trabalhado no início dos anos 2000, e representa conjuntos de dados complexos e de grandes dimensões e, dada a enorme quantidade de dados que há nós dias de hoje, os processos tradicionais têm dificuldades em conseguir gerir toda a informação (Tardio et al., 2020). Inicialmente o termo *Big Data* era definido por “3Vs”, mas neste momento utilizam-se “5Vs” para a sua caracterização (Tardio et al., 2020), denominados da seguinte forma:

- Volume → Tal como o próprio nome indica, o termo está relacionado com uma enorme quantidade de dados, e ao utilizar *Big Data*, pressupõe-se a necessidade de processar e manipular grandes quantidades de dados. Os dados podem ser de fontes desconhecidas, como por exemplo, *feeds* de dados do Facebook ou fluxos de ações num website ou aplicação móvel.
- Velocidade → A característica velocidade está relacionada com a velocidade que os dados são recebidos, armazenados e tratados.
- Variedade → A variedade representa a diversificação das fontes de dados, incluindo dados não estruturados, semiestruturados e dados estruturados.
  - Dados não estruturados – representam dados não organizados e normalmente refere-se a dados que não se encaixam na estrutura

tradicional de dados do modelo relacional. Alguns exemplos de dados não estruturados, textos, fotos ou vídeos.

- Dados semiestruturados – estes dados são semi-organizados e usualmente não estão em conformidade com a estrutura formal dos dados, como por exemplo os ficheiros que armazenam os *logs*.
  - Dados estruturados – Representam dados estruturados e são dados que cumprem o formato do modelo relacional e podem ser armazenados no formato de linhas e colunas.
- Veracidade → Está relacionado com a possível inconsistência dos dados, ou seja, dada a quantidade enorme de dados, pode ficar confuso e não ser relativamente fácil encontrar a qualidade, precisão e valor dos dados.
  - Valor → Este ponto está correlacionado com a qualidade e o valor que conseguimos extrair dos dados, uma vez que dados sem valor não são úteis para a empresa a menos que sejam transformados em algo útil.

Atualmente a área de *Big Data* traz múltiplos benefícios para as empresas e para o mundo em geral (Rosini Filho et al., 2020; Trovati et al., 2015), pois possibilita que se extraia mais informação e se tire mais vantagens dessa mesma informação, através de análises mais complexas dada a complexidade e a capacidade das ferramentas em tratar os dados, permitindo uma abordagem completamente diferente perante os problemas dada a confiança nas análises dos dados para lidar com determinadas situações (Côrte-Real, 2022). O *Big Data* ajuda a lidar e validar uma série de componentes, como por exemplo, desenvolvimento de produtos, experiência dos clientes ou eficiência operacional.

Do mesmo modo que qualquer outro sistema, os sistemas de *Big Data* também têm desafios, pois embora existam cada vez mais tecnologias para esta área, no caso do armazenamento, a quantidade de dados que está a ser criada regularmente ultrapassa os 2 mil biliões (Abiodun et al., 2021) e as organizações ainda lutam para conseguir lidar com este crescimento dos dados exponencial. É necessário ter em conta que não basta apenas armazenar os dados. Os dados devem ser trabalhados para garantir valor e os recursos são muito dispendiosos para este tratamento, até que os dados tenham valor (Oussous et al., 2018). Por último, a tecnologia relacionada com *Big Data* está em constante mudança e assim surge o maior desafio de todos nesta área que é manter-se atualizado com a tecnologia que está a ser utilizada.

### 3.2. *Cloud Computing*

A *Cloud Computing* é a entrega dos recursos de tecnologias de informação (TI) sob demanda utilizando a internet (McHaney, 2021), ou seja, em vez de comprar e gerir todo o hardware, comprar, ter e manter os servidores físicos, há a possibilidade de aceder a esses serviços de tecnologia remotamente, com a mesma capacidade de processamento e armazenamento utilizando um fornecedor de serviços da *Cloud*, permitindo que as pessoas se foquem nos softwares e se abstraiam da gestão do hardware físico (Baldini et al., 2017).

Atualmente, cada vez mais empresas procuram os serviços disponibilizados na *Cloud*, e isso apresenta múltiplos benefícios, tais como: a agilidade, uma vez que a *Cloud* oferece um acesso fácil a uma grande variedade de tecnologias permitindo uma inovação rápida (McHaney, 2021); a elasticidade que os serviços de *Cloud* oferecem, permite às entidades não gastarem recursos em excesso e gerir os recursos consoante as suas necessidades, permitindo aumentar ou diminuir instantaneamente os recursos para ajudar a capacidade de acordo com a evolução das necessidades empresariais (Abhishek & Rajeswara Rao, 2022); ou ainda, a implementação global em minutos, que através da *Cloud*, permite expandir as atividades das entidades para múltiplas zonas geográficas e implementar soluções a nível global (Avram, 2014)

Segundo (Rahman & Rana, 2021) (Ferreira, sem data) há 3 principais modelos de serviço em *Cloud Computing*:

- **Infraestrutura como serviço (IaaS)** → É a camada mais profunda da *Cloud* e permite que os utilizadores tenham acesso a recursos de rede, computadores e armazenamento dos dados. O IaaS oferece o mais alto nível de flexibilidade e controlo de gestão dos recursos de TI.
- **Plataforma como Serviço (PaaS)** → É a camada intermédia entre os tipos de *Cloud Computing*, e permite aos utilizadores não terem de gerir mais a infraestrutura subjacente, o que possibilita manter o foco na implementação e no software em si.
- **Software como Serviço (SaaS)** → Este tipo oferece um produto completo, executado e gerido pelo fornecedor dos serviços da *Cloud*. Com este tipo, não há a necessidade de pensar em manutenção dos serviços ou a gestão da infraestrutura, podendo apenas focar nos serviços que irá utilizar.

Para além dos diferentes modelos de serviço de *Cloud Computing*, há também diferentes tipos de *Cloud*, como se pode verificar na Figura 666666 Figura 666666 Figura 66 que se pode optar (Bello et al., 2021), tais como, *Cloud* pública, que significa que toda a

infraestrutura é compartilhada por todos os utilizadores, o que representa, por um lado custos mais baixos, mas por outro menor nível de segurança e menor flexibilidade; *Cloud* privada, os recursos são reservados a um único utilizador, de forma a que toda a infraestrutura de servidores, por exemplo, seja gerida exclusivamente pelo cliente que contratou. Entre algumas vantagens estão a maior segurança e flexibilidade, mas como desvantagem tem o custo, que pode ser muito elevado; por último, a *Cloud* híbrida é a junção dos dois modelos anteriores, porque permite que as informações, plataforma e aplicações sejam compartilhados tanto no ambiente físico como à distância(Bello et al., 2021).



Figura 666666: Tipos de Cloud

### 3.3. Arquitetura *serverless*

A arquitetura *serverless*, é uma arquitetura orientada a eventos e que desafia o *status quo* do design de software, embora herde conceitos elementares da arquitetura de micros serviços e permite criar e executar aplicações e serviços sem precisar de gerir qualquer infraestrutura(Sbarski et al., 2022) As aplicações continuam a ser executados em servidores, mas os servidores são geridos apenas pelo fornecedor de serviços da *Cloud*(Castro et al., 2022)

Este conceito é relativamente novo e ganhou popularidade devido a serviços de diversas entidades *Cloud*, tais como, AWS Lambda, Microsoft Azure, Google Cloud, entre outros(Baldini et al., 2017)

Antes de surgir esta nova arquitetura, as implementações iniciais de aplicações de softwares, executadas em servidores físicos, não poderiam utilizar ao máximo a capacidade de computação do hardware subjacente, uma vez que só poderia haver uma instância do sistema operativo em execução num determinado momento(Jiang et al., 2020). Posteriormente, após identificar as capacidades de partilhar os recursos computacionais, vários computadores

virtuais foram capazes de executar o mesmo hardware ao mesmo tempo, ao trocar operações de CPU e I/O entre eles(Jiang et al., 2020). Esta evolução tecnológica levou a muitas inovações no mundo tecnológico e na indústria, e ainda mais importante, levou ao aparecimento e crescimento de *Cloud Computing*(McHaney, 2021).

A arquitetura *serverless*, segundo (Baldini et al., 2017)apresenta como principais vantagens a escalabilidade automática e a redução de custos A escalabilidade é realizada de forma automática, ou seja, quanto mais consumo dos recursos há nos serviços, mais disponibilidade terá. Deste modo, pode-se afirmar que a escalabilidade é quase infinita, mas, de notar que, dependendo do fornecedor de serviços da *Cloud*, poderá existir limitações e bloqueios de segurança(Castro et al., 2022). Quanto à redução de custos, apenas se paga quando as funções são executadas, ou seja, cada vez que a função é executada, o utilizador é cobrado apenas pelo processamento consumido(Baldini et al., 2017) . Sendo assim, o utilizador não paga pelo tempo em que as funções não estão a ser executadas(Abiodun et al., 2021).

Evidente que, como em qualquer caso, um produto tem vantagens, mas também apresenta desvantagens, pois a utilização da arquitetura *serverless* requer muita cautela e profissionalismo na sua implementação(Rajan, 2018). Por exemplo, configurar uma arquitetura deste tipo apenas num fornecedor de serviços garante confiança, mas pode dificultar a troca para fornecedores diferentes no futuro e isso pode acontecer porque nem todos os recursos são iguais e os fluxos podem ser diferentes. Outro exemplo são as preocupações com segurança(Baldini et al., 2017), porque os fornecedores dos serviços frequentemente executarão código de vários clientes no mesmo servidor a qualquer momento e como as empresas não possuem os servidores físicos próprios, surgem questões relacionadas com a partilha de máquinas, o que significa que se as máquinas não forem bem configuradas, os dados podem acabar por ser expostos(Castro et al., 2022)

Em suma, para escolher esta arquitetura, é necessário que a entidade tenha uma confiança muito elevada no fornecedor dos serviços.



visualização dos dados e por último como os utilizadores podem conectar-se e ter acesso aos cubos, podendo assim criar as suas visualizações e realizar as suas análises de dados. São utilizados diversos serviços/ferramentas, tanto da AWS como da Microsoft. Ainda nesta secção, será descrito todo o processo e comunicações, bem como as tecnologias/ferramentas utilizadas.

Nesta descrição será detalhado qual o objetivo e como se comportam os serviços, bem como os problemas que podem surgir e requisitos necessários para o bom funcionamento.

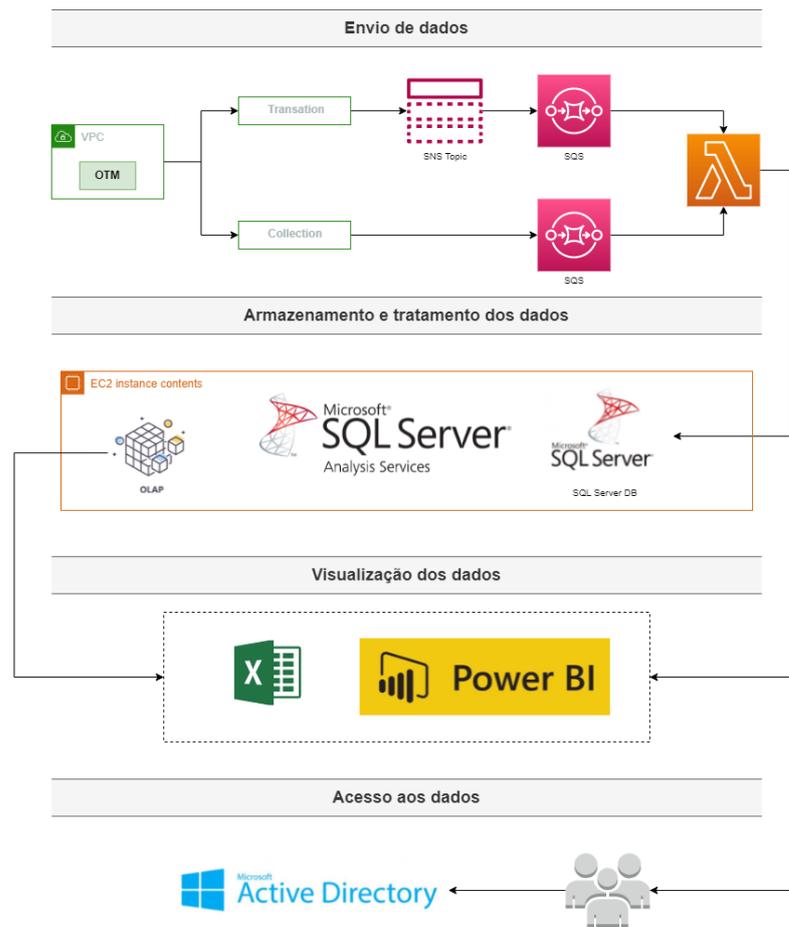


Figura 777777: Arquitetura do sistema As-Is

### Envio de dados

O processo inicia quando se recebe a informação das vias e esta é guardada na base de dados do software de gestão de transações da empresa, o OTM, mais precisamente um serviço denominado de Transaction. O Transaction é um dos múltiplos serviços que integra o OTM e é responsável pela parte operacional do sistema, ou seja, recebe toda a informação da via para que possa ser visualizada do *BackOffice* pelos operadores das diferentes concessões. Quando a informação chega a este serviço, inicia-se o processo de envio para o BI, ou seja, o Transaction

recebe a informação, de seguida envia para um SNS Topic que tem como função enviar a informação para os restantes serviços que necessitam da informação do Transaction, um desses serviços é o BI, ou seja, a informação depois de chegar ao Topic é enviada para um SQS que é destinado apenas para o BI.

Quando a informação chega ao SQS, está programada um lambda function para que, sempre que a *queue* correspondente receber informação, a lambda vai “buscar” essa informação e guarda na tabela de Staging do SQL Server destinado a guardar os dados para o sistema *As-Is*. Os dados guardados na tabela de Staging estão num estado muito cru, sem qualquer tratamento.

Existe ainda outro processo que envia dados para o BI, para além do Transaction, esse processo é responsável por enviar todas as atualizações feitas a cada transação no serviço do Collection, serviço este que é responsável pela parte da cobrança das transações. A única diferença entre este processo e o processo apresentado primeiramente, é que neste processo está definido no seu código que quando há uma alteração numa transação, tem de enviar essa correção para uma *queue* que também é responsável por enviar informação para o BI. Quando a informação enviada pelo Collection chega a *queue*, o processo é idêntico ao do Transaction, isto é, a lambda é “chamada” e envia para a tabela de Staging no SQL Server, criando uma entrada para a mesma transação.

Uma vez que este processo não engloba o ETL, é necessário ter em conta que o número de pedidos neste ponto é relativamente alto e os valores a serem guardados na base de dados são bastantes, pois, para cada iteração é criada uma nova entrada na base de dados. Isto provoca sobrecarga do sistema e dificuldade em filtrar qualquer informação diretamente na base de dados Staging.

Os serviços utilizados a ter em conta para o sistema de BI neste ponto são AWS SQS e AWS Lambda Function apenas, porque são os únicos serviços em que foi criado algo específico para o sistema *As-Is*. O Topic é utilizado para outros parâmetros do OTM, a alteração foi feita a nível das interações com o Topic, não tendo impacto nem num sistema, nem no outro. A vantagem da utilização deste processo é conseguirmos garantir que a informação é enviada em tempo real. A desvantagem deste processo é a possibilidade de se perder informação, ou seja, a *queue* do SQS não guarda a informação, apenas recebe e envia, e pode haver perdas de informação, não há forma de recuperar. O que pode originar discrepâncias de informação entre os sistemas.

### **Descrição dos serviços utilizados no primeiro nível:**

- **AWS SQS**(Buddha & Beesetty, 2019): este serviço tem como objetivo receber e enviar múltiplas mensagens de diversos tamanhos. No sistema A, o serviço é responsável por receber os dados do Kafka e do SNS Topic e armazená-los até que a Lambda Function “procure” os dados e os envie para a respetiva base de dados.
- **AWS Lambda Function**(Poccia, 2017): é um serviço que permite criar uma função na qual definimos como queremos que os dados sejam envidados e como será ativada a lambda para que seja executada a função, para isso, para além do Lambda Function é necessário ter um *trigger* que ative esta lambda. No sistema A, a Lambda Function é ativada quando chegam novos dados à *queue* e o objetivo deste serviço é enviar os dados para a base de dados SQL Server. A lambda já faz a filtragem de dados e envia apenas para o cubo a informação relevante. A informação neste ponto é armazenada numa tabela Staging.

### **Armazenamento e tratamento dos dados**

Esta fase é responsável pelo armazenamento dos dados na tabela de Staging e pelo tratamento dos mesmos e tal como se pode ver na Figura 77, os serviços utilizados são uma instância EC2 da AWS, na qual se instalaram o SQL Server e SSAS, sendo os dois últimos ferramentas da Microsoft.

Uma vez que os dados são armazenados numa forma *raw*, ou seja, sem qualquer tratamento, foi necessário implementar regras e decidir qual a informação que acrescentaria valor ao negócio e ao sistema que irá dar origem aos cubos OLAP. Após o tratamento dos dados, estes são guardados numa tabela de factos, na qual constam os dados altamente redundantes para se obter o melhor desempenho, dados que são imutáveis. Foram também criadas as tabelas dimensão e o sistema segue o modelo em estrela que pode ser consultado na figura 8, que consiste numa abordagem na qual as tabelas modelo são classificadas como tabelas de factos ou dimensões. Estas tabelas de dimensão contêm os dados de referência como a data e as classes dos veículos, por exemplo. Estas tabelas não são ingeridas regularmente com novos dados, em vez disso, todo o conteúdo da tabela é atualizado de uma só vez, uma vez que são dados estáticos, não sofrem alterações.

. Existem dois tipos de abordagens utilizadas para Data Warehouse(Kimball & Caserta, 2004), a *Snowflake* e o modelo em estrela(Sidi et al., 2016), sendo que esta última apresenta

mais vantagens para o sistema atual, pois para a otimização de dados, o modelo em estrela é aquele que utiliza menos *joins* entre as tabelas, o design é mais simples, a complexidade é menor e possui uma redundância maior. Consequentemente, a desvantagem é que as tabelas de dimensão podem ser um problema, pois não são atualizadas de forma automática e podem perder valor. Adicionalmente, este esquema utiliza mais espaço de armazenamento. No modelo em estrela, todas as dimensões relacionam-se com a tabela de factos e apresenta algumas propriedades específicas:

- Apenas uma tabela de factos;
- Uma tabela por dimensão;
- As chaves primárias da tabela de factos, correspondem a apenas uma de cada tabela de dimensão.

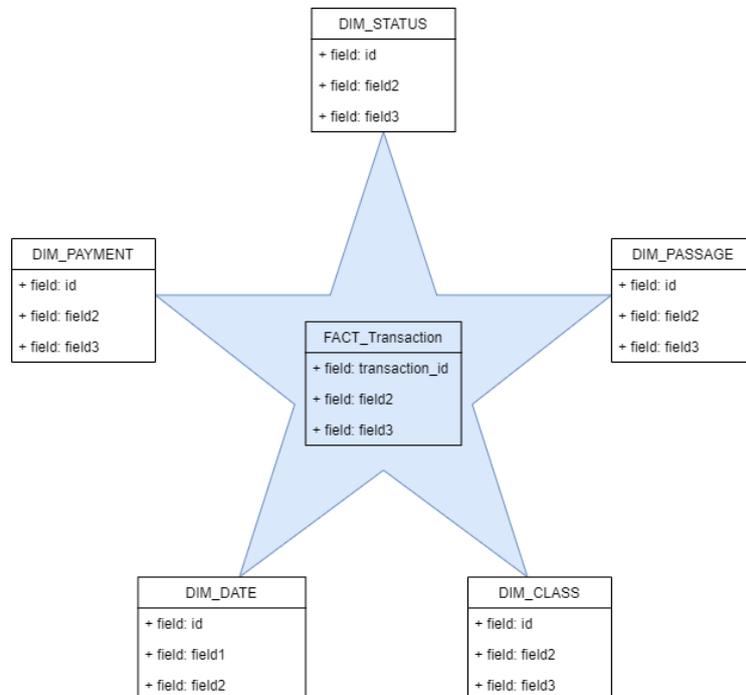


Figura 888888: Modelo em estrela

Após o tratamento dos dados, o próximo passo são os cubos OLAP para as diferentes concessões. Todos seguem o mesmo modelo, apenas possuem um *scheduler* distinto por causa do fuso horário de cada concessão. Para facilitar o processamento da informação foram ainda criadas partições para que fosse possível correr e atualizar apenas uma parte dos dados, pois as quantidades de dados têm vindo a aumentar exponencialmente.

Para este último ponto é utilizado o SSAS, para auxiliar na construção de toda a estrutura mencionada anteriormente e que no final resultarão os cubos OLAP, que posteriormente permitirão conectar e analisar os dados.

### **Descrição dos serviços utilizados no segundo nível:**

- **AWS EC2**(Ifrah, 2019): este serviço permite criar e/ou reservar instâncias com os mais recentes processadores, armazenamento, rede, sistema operacional e modelo de compra. Este serviço oferece infraestrutura de computação segura, confiável, de alto desempenho e económica para atender as necessidades do negócio. Por esse motivo, o core do sistema A encontra-se numa instância criada a partir do EC2. Nessa instância está instalado o SQL Server e todos os serviços necessários para que o sistema A funcione corretamente e retribua o que é pretendido, neste caso os cubos OLAP.
- **SQL Server DB:** é uma base de dados relacional, desenvolvida pela Microsoft, que tem como função principal armazenar e recuperar dados conforme solicitado e neste caso é responsável pelo armazenamento de todos os dados do sistema *As-Is* que são utilizados nos cubos OLAP.
- **SSAS:** é uma ferramenta que oferece suporte a modelos tabulares em todos os níveis de compatibilidade, modelos multidimensionais e Power Pivot. No sistema *As-Is*, é utilizado para a criação e implementação do modelo dos cubos OLAP.

### **Visualização dos dados**

Esta fase, em termos de negócio é a mais relevante pois é a fase que dá valor ao produto, ou seja, é o ponto em que será visualizada a informação dos cubos OLAP e o que apresenta valor para os clientes. Neste ponto é disponibilizado o produto final para que os utilizadores consigam utilizar os cubos OLAP.

As ferramentas utilizadas até a data são o Microsoft Excel e o Microsoft Power BI, nesta última, é utilizada a conexão *SQL Server Analysis Service* disponibilizada. Através destes

recursos os clientes podem realizar análises de dados, estudar KPI's ou auxiliar nas suas tomadas de decisão.

Nas imagens seguintes, é possível verificar alguns exemplos das visualizações criadas através do sistema de Big Data atual da empresa, tanto para um cliente da empresa, realizado com a ferramenta Microsoft Excel, como um relatório solicitado internamente, através da ferramenta Power BI.

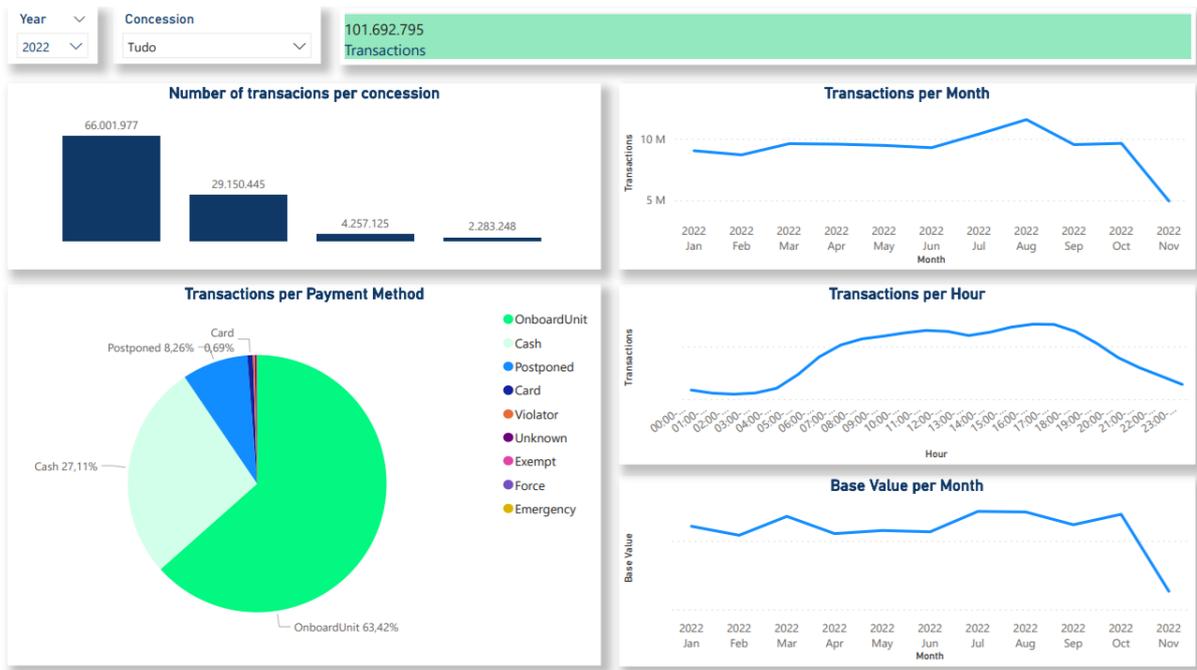


Figura 999999: Relatório criado com a ferramenta Power BI

00:00-01:00	01:00-02:00	02:00-03:00	03:00-04:00	04:00-05:00	05:00-06:00	06:00-07:00	07:00-08:00	08:00-09:00	09:00-10:00	10:00-11:00	11:00-12:00	12:00-13:00	13:00-14:00	14:00-15:00	15:00-16:00	16:00-17:00	17:00-18:00	18:00-19:00	19:00-20:00	20:00-21:00	21:00-22:00	22:00-23:00	23:00-00:01
742	455	323	302	368	734	886	1148	1539	1902	2438	2546	2907	3108	2857	2577	2235	2273	2231	1821	1351	1113	634	324
881	647	432	341	521	1178	1478	1744	2035	1910	1781	2210	2567	2640	2397	2379	2059	1915	1600	1269	2259	1616	533	175
193	106	100	196	540	1072	1095	1465	1944	2321	2371	2776	3225	3146	3445	3637	3855	3823	3376	1612	1267	114	325	
385	117	105	118	348	1641	3418	3527	3038	3056	2961	3043	2875	3050	3731	3562	3840	3928	3419	2862	2190	1522	336	404
186	106	110	119	510	1932	3360	4178	4878	3455	3534	3220	3542	3284	3513	3648	3843	3906	3253	2442	1685	1472	912	405
131	131	12	111	332	1624	3774	3530	3264	3251	3281	3433	3242	3007	3314	3840	4288	3831	3296	2345	2110	1636	1031	431
242	105	132	190	490	1896	3221	4162	4071	3467	3204	3052	3428	3224	3405	3761	3593	3842	3421	2630	1720	1337	883	410
266	114	95	120	343	1908	3774	3652	3255	3177	3033	3183	3223	3185	3554	3985	4041	3848	3288	2859	2107	1705	1186	562
231	107	162	165	454	1901	3236	3963	4186	3177	3145	3150	3353	3225	3507	3753	3864	3988	3408	2563	1924	1586	1254	523
266	114	92	150	321	1961	3529	3407	3022	3197	3036	3095	3271	3456	3981	4298	3985	3324	2950	2302	1932	1268	743	
362	209	165	211	519	1940	3198	4076	4094	3500	3418	3531	3824	3722	3522	3788	3417	3204	2833	2768	2044	1458	2068	1024
408	245	159	167	393	1553	3391	3580	3145	3038	3012	3289	3463	3593	3842	3843	3846	3434	3183	2948	2003	1143	1710	1027
588	368	239	319	583	1248	1953	2135	2570	2963	3068	3300	3811	4005	3629	3180	3011	3147	2381	2849	2105	1539	1228	910
808	528	329	272	346	710	1043	1805	2177	2511	2639	3085	3432	3180	3226	2933	2391	3051	2897	2743	2215	1965	1353	1360
622	380	314	197	282	614	848	1029	1250	1801	2189	2404	2800	2863	2633	2568	2433	2284	2182	1781	1534	903	620	290
388	674	441	265	293	414	376	747	1049	1338	1862	1939	2152	2663	2796	2146	2381	3128	3040	2571	2209	1640	1037	578
186	106	64	218	546	1903	3212	4126	3560	3288	2918	3007	3155	3038	3330	3708	3835	3986	3282	2251	1708	1438	756	332
272	140	95	132	330	1633	3584	3676	3172	2933	2937	2936	3024	3038	3393	3987	4251	3930	3267	2746	2003	1404	933	370
157	36	129	157	459	1977	3263	4140	4328	3646	3128	3058	3284	3101	3426	3631	3674	3885	3358	2392	1747	1560	892	368
228	106	100	115	358	1820	3632	3551	3180	3118	3048	2880	2982	3153	3384	4011	4238	3833	3012	2852	2034	1560	916	460
202	116	109	157	457	1981	3170	4205	4283	3468	3184	3072	3020	3172	3481	3783	3733	3854	3310	2575	1762	1626	114	431
218	129	100	131	315	1493	3978	3716	3205	2390	2327	2960	3145	3152	3593	4087	4341	3969	3247	2789	2095	1700	154	596
234	115	116	157	538	1951	3025	3803	3963	3496	3142	3177	3311	3222	3526	3761	3718	3964	3382	2582	1947	1714	939	543
292	175	32	115	360	1473	3213	3388	3003	3065	3051	3016	3059	3180	3408	4018	4314	3919	3295	2881	2206	1816	1254	635
238	180	150	235	527	1804	3173	4221	3981	3225	3356	3198	3938	3441	3592	3826	3825	3805	3320	2781	2379	1936	1404	833
401	265	208	142	365	1494	3524	3618	3330	3083	2930	3207	3409	3530	3848	4166	3863	3630	2977	2822	2341	2022	1646	1100
604	443	405	413	540	1214	1438	1821	2553	2784	3170	3243	3665	3886	3687	3326	2927	2754	2576	2341	1985	1471	1245	850
911	633	566	421	436	763	1073	1624	2030	2278	2640	2634	3187	2966	3007	3091	3189	2902	2636	2474	2030	1144	1486	1103
586	444	316	266	333	644	892	907	1104	1572	2134	2363	2518	2761	2914	2884	2934	2884	2639	1967	1663	935	789	318
177	535	461	283	305	388	686	640	831	1238	1645	1884	2187	2503	2623	2624	2629	2710	2736	2335	1945	1448	960	508
184	129	127	231	527	1936	3158	4148	4104	3324	3196	3263	3203	3209	3585	3894	3880	4023	3348	2389	1764	1571	733	345
295	150	129	113	393	1978	3558	3426	3163	2312	2849	2923	3070	3167	3354	3847	4206	3723	3189	2870	2199	1591	948	373
176	101	114	193	520	1821	3242	4211	4191	3762	3483	3271	3684	3212	3345	3451	3132	3776	3522	2547	1811	1634	852	403
224	109	94	131	343	1933	3485	3328	3256	3033	2914	2983	3169	3101	3231	3589	3964	3962	3400	2981	2135	1637	1053	480
195	102	106	115	417	1932	3226	4090	4186	3177	3333	3174	3478	3055	3111	3764	3701	3990	3298	2470	1914	1702	947	393
264	132	63	100	352	1466	3459	3619	3293	3044	2900	3089	3128	3005	3162	4061	4130	3905	3190	2904	2283	1638	1145	578
205	132	113	165	408	1905	3248	4092	4194	3776	3305	3165	3421	3224	3390	3277	3483	3770	3108	2514	1831	1141	1088	567
230	137	91	120	345	1917	3507	3607	3251	2359	3053	2929	3147	3262	3533	3540	3590	3600	3503	3059	2196	1171	1242	634
320	189	160	243	506	1892	3195	4092	4008	3495	3376	3680	3882	3578	3642	3860	3420	2963	2953	2816	2090	1524	1308	738
438	228	195	171	339	1422	3513	3526	3187	3041	3115	3210	3566	3743	3878	3443	3773	3101	3160	2737	2308	1953	1533	1006
562	364	297	240	405	1185	1973	2084	2442	2863	3172	3331	3721	3563	3183	3151	2892	3098	2628	2229	1833	1432	1349	890
174	470	386	301	321	672	1038	2003	2008	2321	2768	2935	3341	3096	3864	4293	2871	2704	2760	2479	1977	1706	1472	1153
602	369	372	244	344	721	952	1026	1121	1174	2217	2473	2998	2731	2422	2449	2235	2159	1867	1640	1230	1022	539	285
887	653	414	263	290	396	583	796	1122	1451	1841	2041	2536	2662	2701	2892	3011	2911	2745	2429	2191	1510	942	430
170	110	124	117	335	1983	3192	3821	4152	3463	3199	3032	3078	2891	3356	3977	3437	3807	3106	2966	1973	1441	734	376
224	121	63	115	341	1628	3425	3567	3226	3100	2830	2857	2913	3053	3318	3616	3421	3420	2971	2548	1739	1331	858	355
152	99	116	163	434	1938	3251	3942	4194	3838	3251	3186	3396	3185	3405	3521	3720	3732	3312	2397	1710	1653	1028	372
211	93	67	110	301	1920	3538	3640	3110	3344	3002	3162	3177	3255	3506	4095	4181	3905	3333	2786	2096	1955	1067	425
206	124	115	174	437	1891	3172	4127	4015	3446	3078	2841	3255	3152	3417	3650	3658	3866	3425	2469	1826	1615	921	404
250	120	91	101	322	1666	3536	3435	3078	3101	2888	2838	2921	2946	3550	3928	4178	3787	3091	2721	2270	1738	1231	542
216	125	120	156	432	1916	3296	4053	4388	3533	3394	3286	3408	3190	3476	3552	3276	3246	2874	2639	2410	1510	1360	596
372	153	69	112	320	1483	3457	3613	3192	3175	2963	2949	3234	3243	3524	3844	4128	3221	2713	2046	2473	1937	1314	746
277	188	180	202	432	1948	3118	4047	3324	3849	3929	3432	3714	3539	3663	3723	3675	3294	2720	2674	2064	1879	1530	747
370	236	166	161	351	1491	3482	2418	3186	2863	3102	3195	3479	3085	3966	4082	3825	3246	2874	2639	2410	1510	1360	596
145	335	246	272	476	1171	1957	2023	2449	2864	3273	3790	3822	3669	3567	3657	3066	2896	2409	1928	1501	1002	917	371
743	670	379	300	390	716	1036	2025	2278	2332	2861	3090	3335	3340	3060	2842	2707	2936	2627	2479				

- Facilita a implementação de políticas de utilização (políticas de grupos);
- Unificação do sistema de nomes baseado em DNS;
- Facilita a atribuição e manutenção de múltiplos domínios;
- Proporciona um índice dos recursos na rede.

No caso do sistema *As-Is*, a ferramenta da AD tem configurados diversos grupos e quando se trata de dar acesso ao sistema *As-Is*, há um grupo pré-definido a que os utilizadores têm de ser adicionados.

Para além do acesso à AD, é necessário adicionar os utilizadores no SSAS, para que consigam abrir o cubo OLAP e fazer as suas análises internas. Outro ponto importante na questão dos acessos é, quem não está na rede interna, para além de se adicionar um utilizador na AD, é necessário criar um acesso para a VPN.

Após efetuar estes passos, o utilizador tem acesso ao sistema *As-Is* e pode utilizar sem qualquer restrição, sendo que, cada utilizador apenas tem acesso ao cubo que deve.

#### **Descrição dos serviços utilizados no quarto nível:**

**Active Directory:** é um serviço de diretório desenvolvido pela Microsoft para redes de domínio Windows. Este está incluído na maioria dos sistemas operativos Windows Server como um conjunto de processos e serviços. Inicialmente, o Active Directory era usado apenas para gestão de domínio centralizado. No entanto, o Active Directory acabou por se tornar um título abrangente para uma ampla gama de serviços relacionados com a identidade baseada em ficheiros.

**VPN:** permite-nos controlar quem tem acesso ao sistema, ou seja, para aceder ao Cubo OLAP é necessária uma conexão à VPN da empresa. Este serviço estabelece conexões seguras entre a rede e os escritórios remotos.

## **4.2. Plano de escalonamento**

A escalabilidade é a habilidade de ajustar o sistema à capacidade desejada. O plano de escalonamento surge para apresentar uma solução que garanta essa habilidade de ajuste. O plano será construído tendo como foco duas métricas. Quando falamos do sistema *As-Is*, sabemos que as cargas de trabalho associadas ao sistema são o armazenamento de dados, os utilizadores que acedem à máquina através dos cubos OLAP e a performance da mesma quando

está a processar a informação. Tendo em conta os últimos aspetos apresentados, a escalabilidade será avaliada segundo as duas métricas seguintes:

- **Aumento do número de dados:** no mundo das autoestradas, o fluxo rodoviário é sempre instável, ou seja, num dia pode haver 1000 transações e no dia a seguir 10000, e quanto mais clientes houver mais transações existirão. Assim, lidar com grandes volumes de dados torna-se um enorme desafio porque não se trata apenas de armazenar os dados, mas também a capacidade de manipular os dados em tempo real, o que implica fazer pesquisas, ler os dados e atualizar tabelas de forma que os dados estejam sempre atualizados, sem degradar a experiência e performance para o utilizador.

Atualmente, neste novo mundo da informação, armazenar uma enorme quantidade de dados é um requisito comum, ou seja, quanto mais dados forem armazenados, melhor será o aprimoramento da precisão da análise e a expansão da amostra dos dados.

- **Aumento da concorrência e taxas de interação:** concorrência significa mais conexões, mais pedidos ao servidor, fluxos de dados a serem processados em paralelo, manipulando várias sessões ao mesmo tempo. A taxa de interação está diretamente relacionada com a concorrência, mas este mede a frequência com que os utilizadores interagem com o servidor, que no caso do sistema *As-Is*, é diariamente, inúmeras vezes no mesmo dia. O maior desafio é manter a latência baixa principalmente quando há grandes quantidades de interações com o sistema.

A escalabilidade pode ser feita a nível vertical e a nível horizontal. A **escalabilidade vertical** é relativamente simples, basta adicionar mais recursos ao hardware do servidor, como CPU e memória, ou ainda melhorar o desempenho do disco, alterando-o para um mais rápido. Esta abordagem geralmente é rápida e não requer nenhuma alteração na arquitetura do sistema, especialmente em *Cloud Computing*, onde é possível aumentar a capacidade da máquina em relativamente pouco tempo. A principal desvantagem é poder atingir o limite máximo de hardware que pode ser utilizado no servidor, neste caso, não poderemos aumentar mais o tamanho da RAM ou a quantidade de CPU infinitamente. Já a **escalabilidade horizontal** consiste em adicionar mais servidores e distribuir a carga pelos diferentes servidores em vez de concentrarmos tudo num.

Após esta breve apresentação dos diferentes escalonamentos e métricas a ter em atenção, será apresentado o plano de escalonamento utilizado para o sistema *As-Is*. Desde já fica a nota

que não é utilizado escalonamento automático, apesar de todas as vantagens que tem associadas, isto porque é uma decisão interna da empresa. Sendo assim, o plano de escalonamento contempla os diversos passos para escalar a máquina quando surge essa necessidade.

Consultando a Tabela 1, é possível verificar as medidas que é necessário aplicar em cada etapa da arquitetura.

Tabela 111111: Plano de escalonamento do sistema *As-Is*

Métricas	Envio dos dados	Armazenamento e tratamento dos dados	Visualização dos dados	Acesso aos dados
<b>Aumento do número de dados</b>	A capacidade de processamento/armazenamento da <i>queue</i> escala automaticamente	Necessário aumentar a capacidade da instância.	O aumento do número de dados afeta a visualização dos dados na medida de as análises demorarem a ficar disponíveis	Não afeta diretamente o acesso ao sistema
<b>Aumento da concorrência e taxas de interação</b>	Não afeta o funcionamento do envio dos dados, não é necessária intervenção	Não afeta o correto funcionamento do armazenamento e dos processos de ETL, não é necessária intervenção	Dado as ferramentas utilizadas, a interação é independente, ou seja, é irrelevante o número de pessoas a visualizar os dados	Aumentar a capacidade de processamento da instância a fim de diminuir a latência

#### **Métrica “Aumento do número de dados”:**

No que diz respeito à métrica “Aumento do número de dados”, é necessário ter em atenção a capacidade de armazenamento dos dados quando estes aumentam. Quanto ao envio dos dados, o aumento dos dados implica que a capacidade de envio aumente, ou seja, os dados que estão no SQS a espera que a lambda os envie para o SQL irá aumentar, mas consequentemente o número de vezes que a Lambda Function é chamada também aumenta, o que faz com que não se verifique alterações a nível de *performance*, mas sim de preço, pois quanto mais vezes é chamada a Lambda Function, maior será o valor a ser pago pelo serviço. Em suma, no envio de dados, o sistema está preparado para o aumento de dados.

Quanto ao armazenamento, nesta métrica, o que é necessário fazer é o aumento da capacidade de armazenamento da instância e esta medida é realizada pela empresa diretamente na consola da AWS, o que implica um olhar atento sobre o armazenamento e o aumento do

fluxo de dados. Em média, por mês, são armazenados cerca de 8 000 000 de transações, o que ocupa cerca de 5 GB. Tendo em conta este valor, é possível prever mais ou menos quando é necessário aumentar, com alguma margem de erro, pois tal como a métrica indica, se o fluxo aumentar mais do que o previsto, a capacidade da máquina esgotará mais rapidamente. Os parâmetros que ajudam a detetar que a capacidade atingiu o máximo é quando as partições falham no processamento por falta de capacidade. Nesse momento é necessário intervir.

No que diz respeito à visualização dos dados, o aumento dos dados tem consequências diretas na visualização dos dados, pois quanto maior o número de dados maior a demora na resposta das análises. Tendo como exemplo um cliente que tem um número médio mensal de dados de 300 000 e outro cliente que tem como média mensal 6 000 000, se realizarmos pesquisas mais complexas que impliquem mais junções de informação, o tempo de resposta nas visualizações no cliente 1 é consideravelmente inferior ao cliente 2.

Por último, no acesso aos dados, o seu aumento não tem consequências diretas na forma como acedemos aos dados, sendo que neste patamar não é necessária qualquer intervenção quando o fluxo aumenta.

### **Métrica “Aumento da concorrência e taxas de interação”:**

Esta métrica apresenta consequências diretas em duas das quatro etapas em que definimos esta arquitetura, nomeadamente a visualização dos dados e o acesso aos dados, pois quando aumenta a concorrência significa que há um aumento de conexões e pedidos ao servidor. Nestes casos, a nível de escalonamento é necessário ter em atenção a latência gerada com o aumento da concorrência. As ações a tomar nestes casos de aumento das conexões, pedidos ao servidor, frequência de acesso e número de sessões iniciadas ao mesmo tempo é a capacidade de processamento de pedidos, a CPU e RAM da instância que está a ser utilizada. Este aumento implica alterar as configurações da máquina, a capacidade de processamento, quantos cores utiliza para evitar a latência, apesar de que a latência é até certo ponto inevitável, há medidas que podem ser tomadas para melhorar este parâmetro, tais como:

- Disponibilizar o sistema em mais do que uma zona da AWS, pois quanto maior a distância entre o servidor e o cliente, mais tempo demorará a conexão;
- Aumentar o número de pedidos que podem ser feitos ao servidor e melhorar a rede de acesso utilizada, neste caso a AD e VPN.

### 4.3. Plano de custos

Nesta secção serão apresentados os custos médios para a manutenção do sistema *As-Is*. De notar que os valores aqui apresentados não refletem os atuais números da empresa, por razões contratuais, os dados não podem ser revelados. Para auxiliar na elaboração do plano de custos foi utilizada a ferramenta *Elasticsearch*, a fim de estudar o número de pedidos por mês que o sistema recebe e que são enviadas para o *BackOffice*, conseqüentemente enviados para o sistema de *Big Data*, pois tal como foi indicado anteriormente, quando uma transação chega ao *Backoffice* significa que a transação também foi enviada para o sistema *As-Is* e o mesmo acontece quando a transação é atualizada.

De salientar que nestes cálculos, apenas foram cotados os serviços de uso exclusivo do sistema *As-Is*, ou seja, recursos que são partilhados por outros serviços da empresa não foram contabilizados, como é o caso dos valores da AD e das licenças do Office para uso do Microsoft Excel. No caso do Power BI, por decisão interna, a ferramenta utilizada é a versão desktop que é gratuita e assim não entrará nos cálculos.

Após esta breve introdução, com base na documentação da AWS, e também com recurso à consola da AWS para estudo dos recursos consumidos, os valores aqui apresentados representam os custos mensais para o sistema *As-Is* e anuais. Todos os cálculos dos serviços da AWS, tiveram o auxílio da calculadora fornecida pela entidade para calcular os valores dos seus produtos em que é possível definir todas as características dos serviços.

Tabela 222222: Custos do sistema As-Is

Tecnologias	Quantidade	Valor mensal	Valor anual
AWS Lambda Function	1	123,15 €	1 477,80 €
AWS SQS	2	13,34 €	160,08 €
SQL Server	2	307,80 €	3 693,57 €
EC2	1	331,05 €	3902,60 €
Total		861,45 €	10 337,37 €

É necessário ter em conta que os custos da instância EC2 não são estáveis e podem aumentar mais que uma vez ao ano, pois os cálculos realizados tiveram em conta as configurações da instância à data da realização da pesquisa.





Figura 111111111111: Utilização de CPU

Reverendo os custos mensais e anuais do sistema *As-Is*, o recurso mais dispendioso é a instância utilizada, devido ao tipo escolhido para a criação deste sistema, mas a qualquer momento, a instância EC2 poderá aumentar ainda mais os seus custos porque no futuro será necessário aumentar a capacidade de armazenamento/processamento da instância e quanto mais elevados forem estes recursos mais caro será o serviço. O aumento de preço entre o tipo de instância atual e o tipo seguinte é de cerca de 80,89%. Se for necessário aumentar apenas o volume da instância, de 500 GB para 1 TB os custos aumentam cerca de 19,11%.

Outra vantagem deste sistema a nível de utilização é as ferramentas que são utilizadas para a visualização e análise dos dados, porque são ferramentas muito conhecidas, utilizadas por grande parte dos utilizadores a nível mundial, não só para utilizarem os cubos, mas sim para o seu trabalho diário, noutras vertentes, nomeadamente o Microsoft Excel e o Microsoft Power BI.

No que diz respeito a possíveis melhorias do sistema atual, verificou-se que quanto à questão de escalonamento seria possível aplicar um *autoscaling* na instância EC2, permitindo que a equipa não tivesse de fazer uma monitorização manual. Esta melhoria não foi implementada porque um plano de escalonamento foi desenvolvido e um estudo de crescimento foi feito, prevenindo assim quando a máquina poderia atingir o seu máximo, mas não é 100% fiável, há sempre margem de erro, sendo que o *autoscaling* é 100% fiável.

Para finalizar, na secção 7, será feita a comparação dos dois sistemas, o atual contra o sistema desenvolvido no POC e será tirada a conclusão final e qual o melhor sistema a implementar.

# 5

## 5. Prova de conceitos

Nesta secção será apresentada a prova de conceitos, que surge como um instrumento que visa testar uma determinada solução para auxiliar a empresa à tomada de decisão [7], também para dar a confiança à entidade antes de fazer a alteração de um sistema interno, verificar se a solução corresponde ao esperado e avaliar se o projeto a ser desenvolvido será um agente transformador perceptível para o negócio.

A necessidade de realizar uma prova de conceitos surge com o interesse de a empresa querer alterar o seu sistema de *Big Data* atual, e por esse motivo, decidiu-se fazer uma prova de conceitos para estudar uma alternativa que acrescentasse mais valor à empresa, que desse mais perspetivas de futuro e, igualmente importante, a possibilidade de utilizar os serviços do fornecedor de *Cloud* que oferece mais confiança e é considerada uma das melhores entidades de serviços *Cloud*. Esta prova de conceitos tem como objetivo avaliar uma solução e auxiliar na tomada de decisão de mudar a arquitetura e conseqüentemente o sistema atual.

A prova de conceitos foi desenvolvida por uma equipa parceira do nosso fornecedor de serviços de *Cloud*, na qual, a empresa é o cliente e a equipa o fornecedor de serviços. Nesta perspetiva, o papel desempenhado consistiu em definir os objetivos do POC, acompanhamento da equipa, auxiliar e providenciar qualquer informação que fosse necessária, este acompanhamento enquadra-se na metodologia utilizada internamente pela empresa, o método Agile, no qual se prioriza o contacto direto e continuo entre a equipa e o cliente. Após o acompanhamento no desenrolar do protótipo, seguiu-se a apresentação da possível solução e por fim, como cliente, a análise da proposta e tomada de decisão que será discutido na secção 6. Na tabela seguinte, é possível consultar as etapas da prova de conceitos e os intervenientes do projeto.

Tabela 333333: Etapas e intervenientes da prova de conceitos

<b>Etapas da prova de conceitos</b>	<b>Intervenientes</b>
Definição dos objetivos	Cliente
Desenvolvimento da prova de conceitos	Equipa externa, AWS, cliente
Demonstração	Equipa externa, AWS, cliente
Análise e tomada de decisão	Cliente

De seguida, será descrita cada etapa da prova de conceitos realizada, onde serão apresentados os objetivos, como se realizou a comunicação entre os intervenientes, a solução apresentada, a análise da mesma e a decisão tomada pela empresa, fundamentada.

### **Definição dos objetivos**

Numa primeira fase, antes de iniciar a prova de conceitos, foi necessário traçar os objetivos da solução e os recursos necessários para a prova de conceitos. Como principais objetivos, ficaram definidos os seguintes:

- Uma solução com uma arquitetura *serverless*;
- Utilizar serviços, se possível, apenas do fornecedor de serviços da empresa;
- Uma solução que permitisse uma análise e visualização de dados mais diversificada do que a solução utilizada atualmente;
- Uma solução que permitisse um crescimento exponencial e com visão futurista;

Os objetivos da solução pretendem colmatar todas as dificuldades encontradas no sistema de *Big Data* atual, nomeadamente, a monitorização do espaço da máquina utilizada, as licenças necessárias, a capacidade de processamento de uma quantidade de dados tão grande, a responsabilidade pela infraestrutura, estar a utilizar serviços da Microsoft e não da AWS, que é o principal fornecedor de serviços de *Cloud* da empresa, o que leva a dificuldades de manutenção de serviços que sejam da Microsoft pois não usufruem de suporte direto tal como usufruem da AWS por causa dos restantes serviços. Na parte da visualização de dados, a limitação à utilização de cubos OLAP e consequentemente visualização de dados utilizando apenas Excel e Power BI, o que não permite uma integração com os outros sistemas da empresa.

## **Desenvolvimento da prova de conceitos**

A segunda etapa da prova de conceitos integra todo o desenvolvimento e estudo de uma possível solução feita pela equipa especializada em soluções de BI parceira da AWS e o nosso papel nesta etapa passou por manter uma comunicação ativa com a equipa, esclarecer dúvidas relacionadas com os dados que lhes fornecêssemos, significados de métricas e a lógica de negócio. Este acompanhamento foi feito utilizando ferramentas de comunicação, como o Slack para uma comunicação mais fácil e resposta imediata a possíveis problemas e dúvidas e o Microsoft Teams e Zoom para reunir com a equipa a fim de esclarecer qualquer questão que surgisse. Também foi necessário facilitar o acesso aos nossos serviços de Big Data e utilizadores com acesso às nossas instâncias AWS para o desenvolvimento do protótipo, utilizando as contas da empresa.

## **Demonstração**

A terceira fase do POC consistiu em apresentar o resultado obtido, ou seja, a solução que cumprisse todos os objetivos definidos anteriormente e que solucionasse os problemas identificados no sistema atual. Esta demonstração consistiu na apresentação e descrição da arquitetura *serverless* a utilizar e também a demonstração de todas as fases necessárias, desde a demonstração de ETL dos dados até às ferramentas para a visualização de dados, que serão discutidas e apresentadas no presente relatório, no capítulo 6.

## **Análise e tomada de decisão**

Neste ponto, após ser apresentada a possível decisão, surge a discussão entre o cliente e o parceiro para avaliar se a solução cumpre os requisitos propostos, foi feita também a análise interna da solução comparando com o sistema atual, os custos, utilizando os mesmos valores, avaliou-se o risco, as possibilidades de crescimento e a posterior tomada de decisão. No capítulo 6 será apresentada a solução que resultou da prova de conceitos, o seu plano de escalonamento, plano de custos e a análise geral da solução resultante da prova de conceitos.

# 6

## 6. Sistema *To-Be*

Neste capítulo será apresentada a solução que nasceu da prova de conceitos, será descrita a arquitetura para a solução, os custos desta solução caso seja implementada e os serviços utilizados, bem como a função que desempenham.

O objetivo desta solução é dar não só aos clientes uma plataforma que lhes permita analisar os seus dados, mas também elevar o nível de análise, aplicando KPIs, análises interativas e conexão entre o software da empresa e o *Big Data*. Permitir também a utilização interna, desenvolvendo métricas que permitam alertar e estudar possíveis problemas do sistema. A solução visa dar aos clientes uma solução mais inovadora e mais versátil, com uma capacidade de processamento superior.

### 6.1. Arquitetura

Neste ponto da secção, será descrita a arquitetura da solução resultante da prova de conceitos, uma arquitetura *serverless*, na qual, não há a preocupação com a infraestrutura, ou seja, o principal requisito desta arquitetura é que a empresa se foque apenas no sistema/aplicação de software sem ter de se preocupar com a infraestrutura, os recursos em si, tendo a possibilidade de estudar e automatizar todos os processos de software.

A arquitetura apresentada na Figura 12121212Figura 1212, é constituída por 4 etapas fulcrais, sendo que um dos pontos é semelhante à arquitetura do sistema *As-Is* e alguns recursos já utilizados na arquitetura anterior são também reutilizados nesta solução, pensado para permitir uma alteração mais segura e confiante de um sistema para o outro.

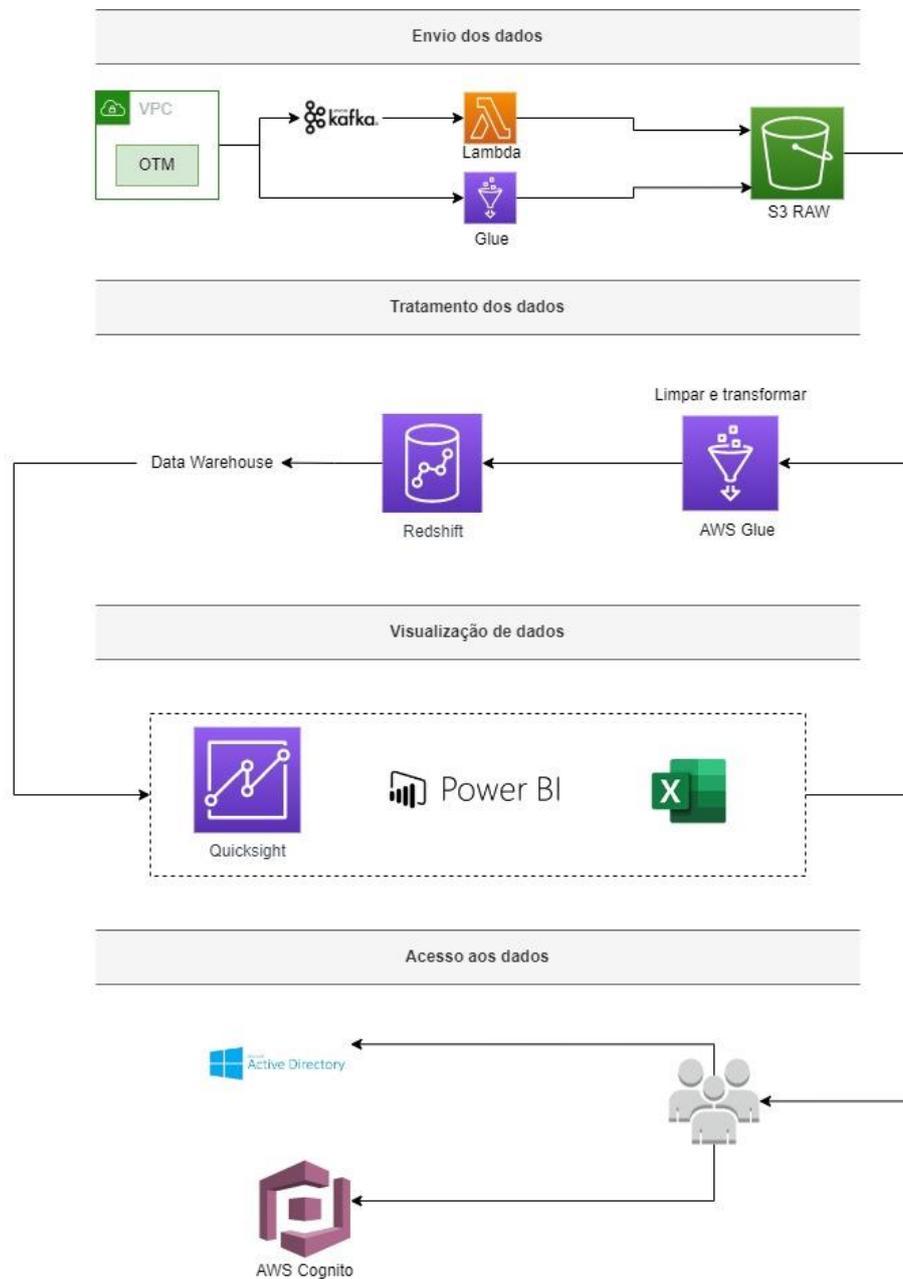


Figura 121212121212: Arquitetura do sistema *To-Be*

### Envio de dados

A primeira etapa a apresentar foca-se no objetivo de como os dados utilizados pelo sistema serão extraídos e como são enviados até ao destino de armazenamento.

Todo o processo tem início quando os dados chegam ao sistema principal da empresa, o OTM, tal como no sistema *As-Is*, mas neste caso, a comunicação e envio dos dados para o destino é feito com recurso a Apache Kafka e Lambda Function, recursos já utilizados

internamente, ou seja, não serão recursos exclusivos e novos, o que é uma vantagem na migração desde ponto. Quanto aos dados, serão enviados para um Bucket S3, neste ponto sem qualquer tratamento, informação “crua”.

Existe também um processo a parte, responsável por todas as atualizações de transações realizadas no OTM, o Aws Glue vai estar a escutar o serviço interno do OTM e atualizar o Bucket S3 sempre que haja uma atualização numa transação, como por exemplo, a alteração de um determinado estado. A VPC não sofre alterações em relação ao sistema *As-Is*.

#### **Descrição dos serviços utilizados no primeiro nível:**

- Apache Kafka(Narkhede et al., 2017)– este recurso é utilizado para streaming dos dados em tempo real, que visa fornecer uma plataforma de alto rendimento e baixa latência para o tratamento de dados. No caso do sistema *To-Be*, terá precisamente a função de *streaming* de dados para o S3.
- Lambda Function(Poccia, 2017)– a Lambda Function, neste sistema, tem o mesmo objetivo que no sistema *As-Is* e poderá ser reutilizada a mesma função lambda do sistema atual da empresa neste novo sistema.
- AWS Glue (Abbasi, 2020) – este serviço fornece diversos componentes, entre os quais, a extração de dados de diversas fontes e este será o seu papel neste ponto. Quando uma atualização de dados for feita, o AWS Glue irá buscar esses novos dados atualizados e irá introduzir os mesmos no S3.

#### **Tratamento de dados**

Neste ponto, serão tratados os dados, através da outra vertente do serviço *AWS Glue*, onde os dados serão limpos, tratados e carregados num Amazon Redshift, que será o Data Warehouse de todo o sistema.

O Amazon Redshift tem uma performance elevadíssima, com uma capacidade excelente de fazer diversos *joins* à informação de forma muito mais rápida do que a forma tradicional sem que se perca capacidade de processamento, o que no geral, causa perdas de performance e os sistemas ficam muito demorados na parte do tratamento dos dados, como é o caso dos Cubos OLAP, que dada a grande quantidade de dados, demoram mais de 1h a executar e atualizar a informação a cada dia que passa no sistema atual.

Após o tratamento dos dados e de os dados estarem no Amazon Redshift, os dados estarão prontos para serem visualizados, ou seja, prontos para serem utilizados nas ferramentas de visualização de dados desta solução.

### **Descrição dos serviços utilizados no primeiro nível:**

- AWS Glue(Abbasi, 2020)– este serviço é um serviço de integração de dados sem servidor que facilita descobrir, preparar e combinar dados para realizar análises. O AWS Glue oferece todos os recursos necessários para que a integração dos dados e os dados estejam disponíveis em minutos e não meses. Através deste serviço é possível implementar processos de ETL, executar e monitorizar visualmente os fluxos de trabalho ETL utilizando a sua interface visual, o AWS Glue Studio. Neste ponto, o AWS Glue será responsável pelos processos ETL e enviar a informação para o destino final, neste caso o Amazon Redshift (Amazon Web Services, Inc, s.d.).
- AWS Redshift (Abbasi, 2020) – o recurso é o Data Warehouse mais utilizado na *Cloud* que utiliza SQL para analisar os dados estruturados. É também um serviço sem servidor e será utilizado para armazenar os dados da nova solução. Dados esses que depois serão utilizados pelo Quicksight para a construção das análises, mas uma grande vantagem é que permite conectar ao Power BI e realizar as análises e construção de *dashboards* nesta ferramenta.

### **Visualização de dados**

Nesta penúltima etapa, estão as tecnologias que serão utilizadas para visualizar os dados, nomeadamente o Amazon Quicksight, inovador e com um poder de processamento elevadíssimo, capaz de criar *dashboards* interativos, nos quais cada cliente poderá trabalhar, permite também ter um sistema de *Reporting* e conectar os *dashboards* à plataforma já existente de cada cliente, entre outras funcionalidades. Tem um motor que incorpora com Machine Learning, que permite aumentar o nível das pesquisas, oferecendo uma opção bastante inovadora que consiste em realizar pesquisas avançadas com insights.

Como a informação após o ETL é guardada no Amazon Redshift, permite conectar também ao Power BI e também ao Microsoft Excel, utilizando uma conexão ODBC, permitindo que os clientes tenham uma solução mais inovadora, mas que nunca percam a solução que já tinham anteriormente, dando-lhes confiança na mudança de serviços, e também possibilidade à empresa de inovar e oferecer melhores produtos e acrescentar mais valor à sua plataforma.

### **Descrição dos serviços utilizados no primeiro nível:**

- AWS Quicksight (Abbasi, 2020) – é um serviço de *Business Intelligence*, com diversas funcionalidades, que permitem uma análise de dados detalhadas. Alguns recursos deste

serviço são conectar e escalar todos os dados, construir painéis personalizados, aproveitar as integrações de ML para insights e ainda permite integrar com outros serviços da AWS, o que será uma grande vantagem pois permitirá integrar diretamente com a plataforma principal da empresa.

- Microsoft Power BI – é uma ferramenta interativa de visualização de dados, já utilizada no sistema *As-Is* e com esta nova solução há a possibilidade de os clientes continuarem a utilizar um serviço com o qual já estão familiarizados.

### **Acesso aos dados**

Neste último ponto, tal como foi dito anteriormente, como há a possibilidade de os utilizadores se conectarem ao Excel e ao Power BI, uma parte do acesso aos dados é feito tal como no sistema *As-Is*. Uma vez que a outra ferramenta pertence a AWS, é possível reaproveitar o serviço utilizado para a conexão ao sistema OTM e utilizá-lo no Amazon Quicksight e outra vantagem é também, como será possível conectar o painel do Quicksight ao sistema OTM, os clientes poderão utilizar diretamente no sistema OTM sem terem de aceder a uma plataforma distinta e o login poderá ser reaproveitado.

### **Descrição dos serviços utilizados no primeiro nível:**

- AWS Cognito (Poccia, 2017) – este recurso permite controlar os acessos às aplicações com rapidez e facilidade, e uma vez que já é utilizado para controlar o acesso às aplicações das empresas, será utilizado nesta solução uma vez que haverá integração direta com o Quicksight e a aplicação, a forma de aceder será o login utilizado já na aplicação, ou seja, com recurso ao serviço do Cognito da AWS.
- *Active Directory* – o serviço da *Active Directory* é reaproveitado do sistema *As-Is*, ou seja, o seu funcionamento não altera em nada em relação ao já utilizado no sistema atual.

## **6.2. Plano de escalonamento**

Nesta secção será discutido o tema do escalonamento da nova solução. Quando a empresa decidiu realizar uma prova de conceitos em busca de uma nova solução que contemplasse todas as necessidades da empresa e acrescentasse valor ao já existente, um dos objetivos era que o sistema novo teria como base uma arquitetura *serverless*, isto é, é uma arquitetura sem servidor, uma arquitetura orientada a eventos que permite que as entidades se

foquem no software e no produto que querem desenvolver e entregar, melhorando continuamente sem terem de se preocupar com a infraestrutura onde as aplicações se localizam. A responsabilidade da infraestrutura neste tipo de arquiteturas fica a cargo da entidade que fornece os serviços, neste caso a AWS.

Este requisito é uma mais-valia na questão do escalonamento, pois não é necessária qualquer intervenção por parte da empresa e deixa de ser uma preocupação, sendo que o tipo de arquitetura escolhida já pressupõe que a entidade não terá qualquer responsabilidade sobre a infraestrutura.

Em suma, não há a necessidade de desenvolver um plano de escalonamento devido ao tipo de arquitetura utilizada na solução do sistema *To-Be*. Apesar de não ser necessário, serão descritas as intervenções necessárias nos mesmos casos de uso utilizados para a elaboração do plano do sistema *As-Is*. Assim sendo, através da Tabela 444444Tabela 4444Tabela 44Tabela 4, é possível verificar em que etapas será necessário intervir, segundo as métricas definidas:

Tabela 444444: Plano de escalonamento do sistema *To-Be*

<b>Métricas</b>	<b>Envio dos dados</b>	<b>Tratamento dos dados</b>	<b>Visualização dos dados</b>	<b>Acesso aos dados</b>
<b>Aumento do número de dados</b>	A carga sobre o apache Kafka aumentará, mas não é necessária intervenção. No caso do S3, será necessário aumentar a sua capacidade de armazenamento	Será necessário escalar o AWS Redshift. Quanto ao AWS Glue, não é necessária qualquer interação.	Necessário aumentar o espaço de uma funcionalidade do Quicksight.	Não há necessidade de intervenção
<b>Aumento da concorrência e taxas de interações</b>	Não há necessidade de intervenção	Não há necessidade de intervenção no AWS Glue, pode alterar-se o tipo de AWS Redshift escolhido.	Não há necessidade de intervenção, mas os custos irão aumentar, inevitavelmente	Não há necessidade de intervenção

### **Métrica “Aumento do número de dados”:**

No caso do sistema *To-Be*, quando há um aumento significativo da quantidade de dados, esta situação irá afetar diretamente o serviço responsável pelo armazenamento dos dados, o S3, onde é armazenada a informação num primeiro ponto e o Redshift onde é guardado o resultado, mas também é necessário aumentar o espaço SPICE do Quicksight pois também são armazenados dados nesta funcionalidade do serviço.

Quanto aos restantes serviços, não há a necessidade de qualquer intervenção, pois mesmo que aumente a quantidade de dados, não afeta os serviços diretamente, poderá influenciar a performance, mas quanto aos dados e ao número dos dados não há configurações presentes que possam sofrer alterações.

### **Métrica “Aumento da concorrência e taxas de interações”:**

Quanto a esta métrica, no sistema *To-Be* não é requerida qualquer alteração caso o número de interações ou a concorrência aumente no que diz respeito à maioria dos serviços do sistema, mas caso aumentem as chamadas ao Redshift poderá ser necessário alterar e aumentar o tipo de instância. É necessário ter em conta que o Quicksight taxa o número de pessoas que utiliza o serviço e como a empresa pretende conectar este serviço à plataforma atual, o número de interações será 200% superior ao atual do sistema de *Big Data* atual, e se todos os utilizadores utilizarem ao mesmo tempo o Quicksight, o valor taxado será superior.

## **6.3. Plano de custos**

Nesta secção serão apresentados os custos desta nova solução, a nível mensal e anual. Todos os valores aqui apresentados dos novos serviços contemplados na nova solução foram calculados consoante os valores tabelados pela AWS, os valores dos serviços/tecnologias que já eram utilizadas e taxadas pela empresa foram alterados, a fim de não partilhar dados internos da empresa.

De notar que apenas os recursos exclusivos para criação do sistema *To-be* foram cotados, uma vez que determinados serviços são utilizados por outros produtos internos e complementam o sistema *To-Be*, tal como é o caso do Apache Kafka e AWS Cognito. Quanto ao Power BI, não é contabilizado pois, por decisão interna, é utilizado apenas a versão Desktop que é gratuita.

Após esta breve introdução e com o auxílio da calculadora de valores da AWS, os valores mensais e anuais da solução para o sistema *To-Be* podem ser consultados na seguinte tabela:

Tabela 555555: Custos do sistema *To-Be*

Tecnologias	Quantidade	Valor mensal	Valor anual
AWS Lambda Function	1	123,15 €	1 477,80 €
AWS Glue	1	0 €	0 €
AWS Redshift	1	417,43 €	5009,12 €
AWS S3	1	26,45 €	317,4 €
AWS Quicksight	1	613,64 €	7363,68 €
Total		1180,62 €	14167,44 €

Todos os valores aqui apresentados foram a colmatação dos requisitos e o estudo das características que a empresa gostaria de ter, tais características não podem ser partilhadas por questões internas e os valores apresentados foram alterados seguindo a mesma escala utilizada no sistema *As-Is*.

Consultando a tabela é possível verificar que o serviço mais caro é também o mais inovador e mais recente, o AWS Quicksight e segundo a utilização e as características que o serviço AWS Glue teria no sistema, é possível utilizar os recursos da *AWS Free Tier*, sendo que os custos seriam nulos. Quanto à AWS Lambda, os custos são os mesmos que os do sistema atual da empresa pois seria utilizada o mesmo lambda.

#### 6.4. Análise do sistema *To-Be*

Nesta secção será feita a análise do sistema *To-Be*, os seus pontos fortes e os pontos fracos, bem como a análise se a solução apresentada cumpre todos os requisitos definidos na prova de conceitos.

Iniciando pelos pontos fortes da solução apresentada, o sistema combina soluções inovadores (AWS Redshift e AWS Quicksight) com ferramentas já conhecidas por grande parte dos utilizadores a nível interno e externo (Microsoft Excel e Microsoft Power BI), o que oferece maior confiança aos utilizadores caso ocorra a mudança, mas surge também uma contradição, pois estes dois serviços não são da AWS. Apesar de não serem da AWS (requisito inicial), no caso do Microsoft Excel, uma vez que a empresa já tem as licenças Office, terá sempre acesso a este recurso, portanto é uma vantagem ter uma solução que o permite utilizar. Quanto ao

Microsoft Power BI, apenas é utilizada a versão gratuita, então não acarreta custos para a empresa e a ferramenta permite a ligação direta ao AWS Redshift, o que também é uma mais-valia para a entidade, pois assim consegue oferecer mais variedade de ferramentas e formas de visualizações.

Outra importante vantagem é a arquitetura *serverless* e a abstração de responsabilidades relativa a infraestruturas e manutenção de serviços, essa responsabilidade passa toda para o fornecedor de serviços e a empresa consegue focar-se nos processos de ETL, na melhor forma de os otimizar e focar também na visualização de dados.

A desvantagem mais clara são os custos que esta solução apresenta, é uma solução muito mais dispendiosa do que o sistema atual da empresa. E outro ponto a ter em conta é o facto que estar apenas a escolher serviços de um fornecedor de serviços que requer uma confiança muito elevada e pode acontecer no futuro que a empresa queira trocar alguns recursos e esse processo pode ser mais complicado e lento, uma vez que nem todos os recursos são iguais e os *workflows* podem ser distintos.

Reverendo os requisitos definidos na prova de conceitos:

1. Uma solução com uma arquitetura *serverless*;
2. A utilizar serviços, se possível, apenas do fornecedor de serviços da empresa;
3. Uma solução que permitisse uma análise e visualização de dados mais diversificada do que a utilizada atualmente;
4. Uma solução que permitisse um crescimento exponencial e com visão futurista;

Podemos concluir que os requisitos 1,2 e 3 foram cumpridos com sucesso e o requisito 4 é o mais generalista mas analisando ao pormenor a ferramenta AWS Quicksight, esta incorpora Machine Learning e oferece ao utilizador funcionalidades de pesquisa utilizando ML incorporado que garante uma qualidade superior a qualquer outro sistema e já com uma visão no futuro, o serviço AWS Quicksight permite uma integração fácil e simples do serviço AWS SageMaker(Abbasi, 2020), que é um serviço que permite aos utilizadores criar e treinar modelos de ML para qualquer caso de uso.

Em suma, podemos concluir que relativamente ao ponto 4, foi tida em conta a possibilidade de crescer e formas de melhorar o sistema continuamente com o leque de serviços oferecidos.

## 7. Conclusões

Nesta secção serão comparados ambos os sistemas resultantes deste trabalho, bem como a decisão tomada com base na análise realizada e também serão apresentados os próximos passos a realizar. Através da tabela 4, é possível verificar a análise e comparação dos dois sistemas:

Tabela 666666: Dados de comparação dos dois sistemas

	<b>Sistema <i>As-Is</i></b>	<b>Sistema <i>To-Be</i></b>
<b>Custos</b>	Custos mais baixos	Custos mais elevados
<b>Escalonamento</b>	Manual	Automático
<b>Reutilização de recursos internos</b>	Sim	Sim
<b>Recursos de visualização de dados</b>	2	3
<b>Suporte aos serviços</b>	Não	Sim
<b>Permite integração com outros serviços</b>	Não	Sim

Consultando a tabela, é possível concluir que os pontos positivos da nova solução são superiores ao do sistema atual em utilização, 5 em 6 pontos são a favor do novo sistema e apenas 1 em 6 são a favor do sistema atual. No que diz respeito a cada parâmetro, apesar dos custos do sistema novo serem mais caros, este permite uma integração tanto com a plataforma interna da empresa, utilizada por todos os clientes e permitindo assim incorporar uma nova funcionalidade

na plataforma, o que acrescenta valor ao software e é oferecido também suporte a todos os serviços da AWS, o que não acontece com o SQL Server e o SSAS utilizado atualmente, pois cada vez que há um problema, tem de ser a empresa a resolver, ou procurar em fóruns e contactar o suporte geral da Microsoft que não é exclusivo para estes serviços enquanto que o da AWS sim é.

Concluindo o tema, a entidade optará por implementar o novo sistema de BI na empresa, pois do ponto de vista interno e com base na análise realizada, os custos compensam o valor obtido com a solução. Os próximos passos a realizar são: implementação da solução, adicionar a solução na plataforma interna, desenvolver todas as visualizações e estudo de KPIs para utilização interna a fim de detetar possíveis problemas e melhorar processos.

Assim, como fruto do trabalho realizado no decorrer do estágio, foi realizada a análise de uma nova implementação de uma solução inovadora que permitirá acrescentar valor ao produto final da empresa e permitirá que a empresa cresça na área de BI, integrando serviços recentes e oferecendo aos seus clientes múltiplas soluções para trabalharem os seus dados. Permitirá ainda colmatar outros problemas internos, tais como a implementação e entrega de relatórios aos seus clientes, sendo este um ponto extra do novo sistema. No decorrer no projeto foi ainda possível desenvolver e implementar diversos relatórios e visualizações utilizando o sistema atual, tanto para a realização do presente trabalho mas também a pedido de colaboradores para auxílio das suas atividades diárias.

Quanto à experiência profissional e pessoal adquirida no decorrer do estágio, a experiência foi muito gratificante, permitiu ganhar conhecimento e utilizar múltiplas ferramentas que são bastante utilizadas no mercado de trabalho, tais como Power BI, SQL Server, AWS Quicksight. A oportunidade de interagir com profissionais especializados da área de BI, permitiu também ganhar mais conhecimento e perspectivas de como um projeto de BI se realiza e todo o processo necessário. Estar presente em todas as fases foi uma grande vantagem pois permitiu não só aprender sobre a gestão do projeto, bem como compreender o mundo das autoestradas e todo o negócio envolvente.

## Referências Bibliográficas

- Abbasi, A. (2020). *Aws certified big data study guide: <br>* (1.<sup>a</sup> ed.). John Wiley and Sons.
- Abhishek, M. K., & Rajeswara Rao, D. (2022). A Scalable Framework for High-Performance Computing with Cloud. Em M. Tuba, S. Akashe, & A. Joshi (Eds.), *ICT Systems and Sustainability* (Vol. 321, pp. 225–236). Springer Nature Singapore.  
[https://doi.org/10.1007/978-981-16-5987-4\\_24](https://doi.org/10.1007/978-981-16-5987-4_24)
- Abiodun, M. K., Awotunde, J. B., Ogundokun, R. O., Misra, S., Adeniyi, E. A., Arowolo, M. O., & Jaglan, V. (2021). Cloud and Big Data: A Mutual Benefit for Organization Development. *Journal of Physics: Conference Series*, 1767(1), 012020.  
<https://doi.org/10.1088/1742-6596/1767/1/012020>
- Avram, M. G. (2014). Advantages and Challenges of Adopting Cloud Computing from an Enterprise Perspective. *Procedia Technology*, 12, 529–534.  
<https://doi.org/10.1016/j.protcy.2013.12.525>
- Baldini, I., Castro, P., Chang, K., Cheng, P., Fink, S., Ishakian, V., Mitchell, N., Muthusamy, V., Rabbah, R., Slominski, A., & Suter, P. (2017). Serverless Computing: Current Trends and Open Problems. Em S. Chaudhary, G. Somani, & R. Buyya (Eds.), *Research Advances in Cloud Computing* (pp. 1–20). Springer Singapore.  
[https://doi.org/10.1007/978-981-10-5026-8\\_1](https://doi.org/10.1007/978-981-10-5026-8_1)
- Bello, S. A., Oyedele, L. O., Akinade, O. O., Bilal, M., Davila Delgado, J. M., Akanbi, L. A., Ajayi, A. O., & Owolabi, H. A. (2021). Cloud computing in construction industry: Use

- cases, benefits and challenges. *Automation in Construction*, 122, 103441. <https://doi.org/10.1016/j.autcon.2020.103441>
- Berisha, B., Mëziu, E., & Shabani, I. (2022). Big data analytics in Cloud computing: An overview. *Journal of Cloud Computing*, 11(1), 24. <https://doi.org/10.1186/s13677-022-00301-w>
- Buddha, J. P., & Beesetty, R. (2019). *The definitive guide to AWS application integration: With Amazon SQD, SNS, SWF and step functions*. Apress.
- Castro, P., Isahagian, V., Muthusamy, V., & Slominski, A. (2022). *Hybrid Serverless Computing: Opportunities and Challenges*. <https://doi.org/10.48550/ARXIV.2208.04213>
- Côrte-Real, N. (2022). *Big Data & Analytics*. INFLUÊNCIA.
- Ferreira, A. M. (sem data). *Introdução ao Cloud Computing* (2015.<sup>a</sup> ed.). FCA.
- González Moyano, C., Pufahl, L., Weber, I., & Mendling, J. (2022). Uses of business process modeling in agile software development projects. *Information and Software Technology*, 152, 107028. <https://doi.org/10.1016/j.infsof.2022.107028>
- Howes, T., Smith, M., & Good, G. S. (2003). *Understanding and deploying LDAP directory services* (2nd ed). Addison-Wesley.
- Ifrah, S. (2019). *Deploy containers on AWS: With EC2, ECS, and EKS*. Apress.
- Jiang, L., Pei, Y., & Zhao, J. (2020). Overview Of Serverless Architecture Research. *Journal of Physics: Conference Series*, 1453(1), 012119. <https://doi.org/10.1088/1742-6596/1453/1/012119>
- Kimball, R., & Caserta, J. (2004). *The data warehouse ETL toolkit: Practical techniques for extracting, cleaning, conforming, and delivering data*. Wiley.
- Marr, B. (2015). *Big data: Using smart big data, analytics and metrics to make better decisions and improve performance*. Wiley.

- McHaney, R. (2021). *Cloud technologies: An overview of cloud computing technologies for managers* (First edition). Wiley.
- Narkhede, N., Shapira, G., & Palino, T. (2017). *Kafka: The definitive guide: real-time data and stream processing at scale* (First edition). O'Reilly Media.
- Oussous, A., Benjelloun, F.-Z., Ait Lahcen, A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*, 30(4), 431–448. <https://doi.org/10.1016/j.jksuci.2017.06.001>
- Poccia, D. (2017). *AWS Lambda in action: Event-driven serverless applications*. Manning Publications.
- Rahman, L. A., & Rana, M. E. (2021). The Convergence Between Big Data and the Cloud: A Review. *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, 592–598. <https://doi.org/10.1109/ICDABI53623.2021.9655772>
- Rajan, R. A. P. (2018). Serverless Architecture—A Revolution in Cloud Computing. *2018 Tenth International Conference on Advanced Computing (ICoAC)*, 88–93. <https://doi.org/10.1109/ICoAC44903.2018.8939081>
- Rosini Filho, A. M., Rosini, A. M., & Palmisano, A. (2020). A ERA DO BIG DATA: PRINCIPAIS IMPLICAÇÕES SOBRE SEGURANÇA E PRIVACIDADE E AS NOVAS TECNOLOGIAS CAPAZES DE AUXILIAR PROCESSOS INVESTIGATIVOS E DETECÇÃO DE FRAUDES EM TEMPO REAL. *Journal on Innovation and Sustainability RISUS*, 11(3), 13–34. <https://doi.org/10.23925/2179-3565.2020v11i3p13-34>
- Rossi, R., & Hirma, K. (2022). *Characterizing Big Data Management*. <https://doi.org/10.48550/ARXIV.2201.05929>

- Rumetna, M. S., Lina, T. N., Rajagukguk, I. S., Pormes, F. S., & Santoso, A. B. (2022). Payroll Information System Design Using Waterfall Method. *International Journal of Advances in Data and Information Systems*, 3(1). <https://doi.org/10.25008/ijadis.v3i1.1227>
- Sbarski, P., Cui, Y., & Nair, A. (2022). *Serverless architectures on AWS* (Second edition). Manning Publications Co.
- Sidi, E., El, M., & Amin, E. (2016). Star Schema Advantages on Data Warehouse: Using Bitmap Index and Partitioned Fact Tables. *International Journal of Computer Applications*, 134(13), 11–13. <https://doi.org/10.5120/ijca2016908108>
- Tardio, R., Mate, A., & Trujillo, J. (2020). An Iterative Methodology for Defining Big Data Analytics Architectures. *IEEE Access*, 8, 210597–210616. <https://doi.org/10.1109/ACCESS.2020.3039455>
- Trovati, M., Hill, R., Anjum, A., Zhu, S. Y., & Liu, L. (Eds.). (2015). *Big-Data Analytics and Cloud Computing*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-25313-8>