Computação em Nuvem 24/25 Final Report

Grupo 10

Tiago Almeida – 58161

Manuel Campos – 58166

Guilherme Sousa – 58170

Gabriel Henriques - 58182

Motivation, dataset characterization, business capabilities and use cases

Motivation:

Escolhemos o tema "Stand de Carros Virtual" porque, dentro do nosso grupo, todos partilhamos um interesse por automóveis e temos planos de adquirir um no futuro. O facto de todos termos o gosto por carros em comum tornou o desenvolvimento do projeto muito mais interessante.

Além disso, considerámos o projeto desafiante do ponto de vista técnico, já que envolve o desenvolvimento de vários microsserviços, cada um com responsabilidades distintas.

Acreditamos que este projeto nos permite aplicar conhecimentos adquiridos ao longo do curso, como orquestração de serviços, tolerância a falhas, escalabilidade, e comunicação assíncrona, num contexto que é simultaneamente prático e alinhado com os nossos interesses pessoais.

Dataset Characterization:

O dataset escolhido contém praticamente todas as informações relevantes que o Craigslist fornece sobre vendas de carros, incluindo colunas como preço, estado (condição), fabricante, latitude/longitude e outras 18 categorias.

Para o nosso projeto decidimos retirar grande parte das categorias do dataset, isto porque achámos que não eram úteis para o nosso projeto e estariam só a ocupar espaço, um exemplo óbvio de uma categoria considerada "inútil" foi a descrição de cada anúncio, metade dos anúncios tinham todos a mesma descrição e ocupava muito espaço.

URL do Dataset:

https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data?resource=download

Business Capabilities:

- 1. Gestão de Anúncios de Carros
- 2. Venda de Carros e Transações
- 3. Agendamento de Reuniões
- 4. Manutenção de Carros
- 5. Inspeção Periódica de Carros
- 6. Anúncios em Destaque

Inicialmente eram estas as *business capabilities* que tínhamos em mente para aplicar no nosso projeto, conseguimos aplicar todas menos a última.

Use cases:

- 1. Comprar um carro
- 2. Alugar um carro
- 3. Devolver um carro alugado
- 4. Explorar os anúncios de carros
- 5. Marcar uma reunião com o staff
- 6. Ver detalhes dos carros

- 7. Adicionar um anúncio de um carro
- 8. Alterar um anúncio
- 9. Remover um anúncio
- 10. Adicionar um anúncio aos anúncios em destaque
- 11. Ver o histórico de anúncios
- 12. Ver o histórico de vendas e alugueres
- 13. Ver relatórios de inspeções e manutenções

Tal como as *business capabilities*, estes eram os *use cases* que tínhamos planeado para o nosso projeto, não foi possível fazer os últimos 4 pontos devido à complexidade dos mesmos e do tempo que tínhamos para o projeto.

API Specification



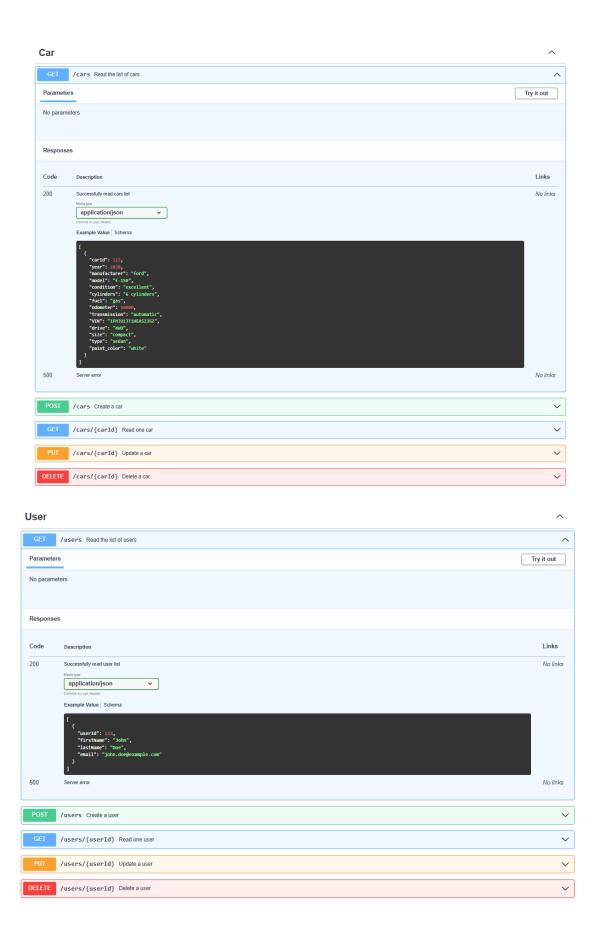
An API for managing a cloud-based car dealership service.

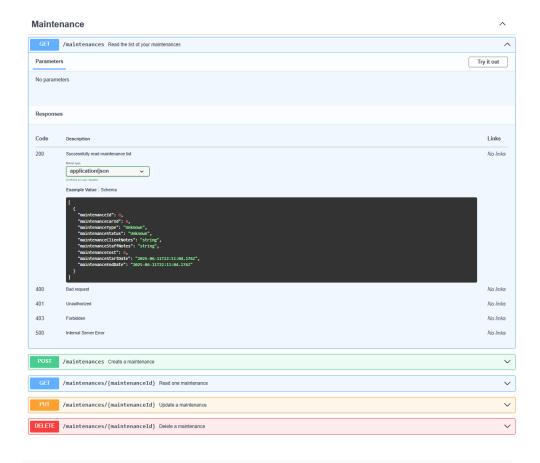
IMPORTANT: All endpoints in this API must be prefixed with '/api'. For example:

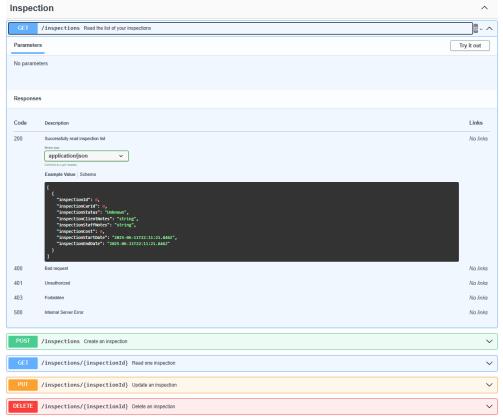
- · '/cars' should be accessed as '/api/cars'
- · '/users' should be accessed as '/api/users'
- · '/meetings' should be accessed as '/api/meetings' And so on for all endpoints.

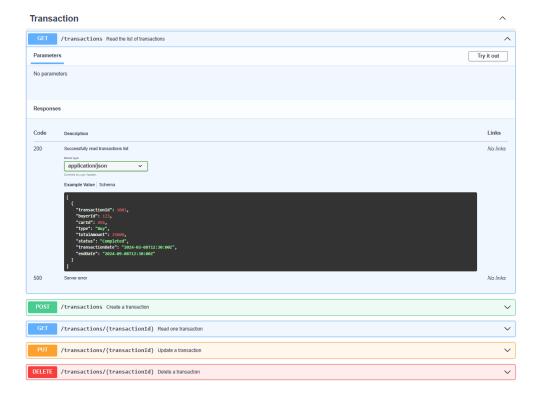
Servers

/api - Base URL for all endpoints. All paths in this document must be prefixed with this URL.

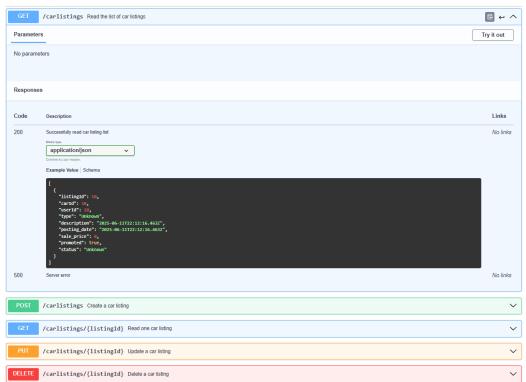


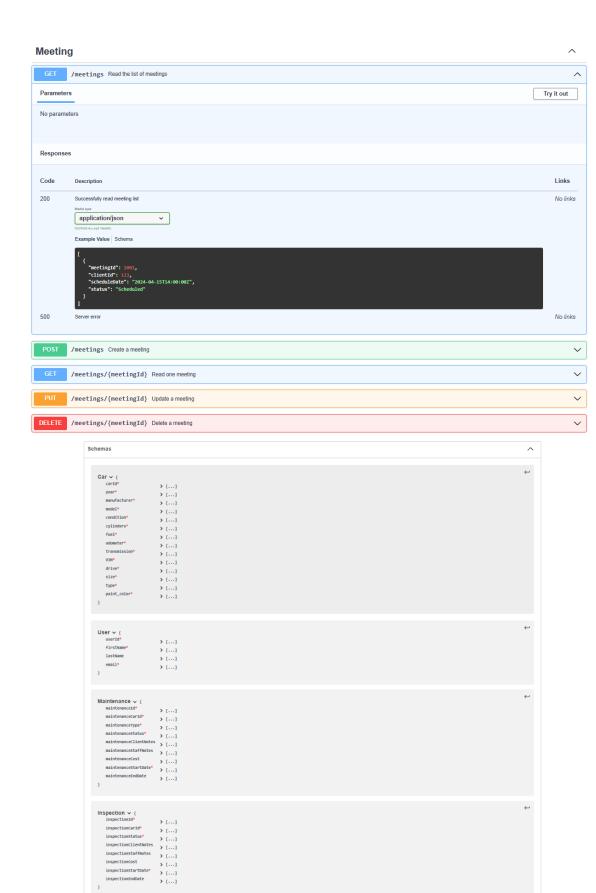






Car Listing





```
Transaction ✓ {
  transactionId*
                    > [...]
  buyerId*
                    > [...]
  carId*
                 > [...]
  type*
                    > [...]
  totalAmount*
                    > [...]
                    > [...]
  transactionDate*
                    > [...]
CarListing V {
                  > [...]
  carId*
                    > [...]
  userId*
  type*
                    > [...]
  description*
                    > [...]
  posting_date*
  sale_price*
 promoted*
                  > [...]
                    > [...]
                    > [...]
Meeting ✓ {
meetingId*
                    > [...]
 clientId*
                    > [...]
  scheduleDate*
                    > [...]
                    > [...]
```

Functional and Non-Functional Requirements

1. Functional Requirements:

1.1. Gestão de anúncios de carros:

1.1.1. Adicionar um anúncio de um carro

- Os utilizadores podem adicionar um anúncio de carro fornecendo os detalhes do carro (fabricante, modelo, ano, estado, etc.) e os detalhes do anúncio (preço de venda, descrição).
- O sistema deverá validar os campos obrigatórios e rejeitar submissões incompletas.
- O anúncio deverá ser atribuído um "listingId" único

1.1.2. Alterar um anúncio de um carro

 Os utilizadores podem alterar os seus anúncios de carro para modificar detalhes como o preço, a descrição ou o estado. O sistema deverá validar todas as atualizações antes de aplicar as alterações.

1.1.3. Remover um anúncio de um carro

- Os utilizadores podem remover os seus anúncios de carro quando desejarem.
- Os anúncios removidos deixaram de estar visíveis para os compradores.
- O sistema não deverá reter registos históricos dos anúncios removidos.

1.1.4. Ver anúncios de carros

- Os utilizadores podem navegar pelos carros disponíveis através de uma lista paginada e filtrável.
- O sistema deverá permitir a filtragem por detalhes do carro, como fabricante, modelo, ano, estado, etc.
- Cada item da lista deverá apresentar uma descrição, o preço de venda, a data de publicação e um link para a página completa do anúncio do carro.

1.1.5. Ver um anúncio de um carro específico

 Os utilizadores podem clicar num anúncio para ver todos os detalhes do carro, informações do vendedor (utilizador) e iniciar o processo de compra do carro.

1.2. Venda de Carros e Transações:

1.2.1. Comprar um carro

- Os utilizadores podem iniciar a compra de um carro ao selecionar um anúncio e proceder com a transação.
- O sistema deverá processar os pagamentos e atualizar o estado do anúncio para "vendido" após a conclusão.

1.2.2. Alugar um carro

- Os utilizadores podem alugar um carro ao selecionar um anúncio e definir um período de aluguer.
- O sistema deverá verificar a disponibilidade para aluguer antes de confirmar a transação.
- O sistema deverá processar os pagamentos e atualizar o estado do anúncio para "reservado" após a conclusão.

1.2.3. Devolver um carro alugado

- Os utilizadores que disponibilizarem carros para aluguer podem marcar um carro como devolvido no final do período de aluguer.
- O sistema deverá verificar o estado da devolução e atualizar o anúncio em conformidade.

1.2.4. Ver o histórico de transações

- Os utilizadores podem visualizar as suas transações passadas, incluindo detalhes de compras, alugueres e devoluções.
- O sistema deverá fornecer recibos e atualizações sobre o estado das transações.

1.3. Manuntenção de Carros:

1.3.1. Marcar uma manuntenção

- Os proprietários de carros podem agendar serviços de manutenção (básica, completa) e adicionar notas prévias ao agendamento.
- O sistema deverá armazenar os pedidos de manutenção juntamente com os detalhes do carro, definindo o estado da manutenção como "Em curso".

1.3.2. Ver o histórico de manuntenção

- Os utilizadores podem consultar os registos de manutenção anteriores de um carro.
- O sistema deverá armazenar detalhes como a data do serviço, tipo, custo e notas do mecânico.

1.3.3. Pós-manuntenção

- Quando a manutenção do carro termina, um membro do staff adiciona notas, o custo da manutenção e define o estado como "Concluída".
- O sistema deverá armazenar os detalhes adicionados pelo staff e registar a data de conclusão.

1.4. Inspeção Periódica de Carros:

1.4.1. Marcar uma inspeção

Os utilizadores podem agendar uma inspeção periódica para os seus carros.

1.4.2. Ver o relatório da inspeção

 Os utilizadores podem consultar os relatórios de inspeções anteriores de um carro, incluindo problemas e recomendações.

1.4.3. Pós-Inspeção

- Quando a inspeção do carro termina, um membro do staff adiciona notas, o custo da inspeção e define o estado como "Concluída"
- O sistema deverá armazenar os detalhes adicionados pelo staff e registar a data de conclusão.

1.5. Reunião com Staff:

1.5.1. Marcar uma reunião

- Os utilizadores podem agendar uma reunião com o proprietário do carro no respetivo anúncio.
- O sistema deverá permitir a escolha de um intervalo de tempo disponível.

1.5.2. Gerir reuniões

- Os utilizadores podem alterar a data ou cancelar reuniões.
- O sistema deverá registar o estado da reunião (agendada, concluída, cancelada).

2. Non-Functional Requirements:

2.1. Autenticação e Autorização:

2.1.1. Integração com o Auth0

- Objetivo: Fornece autenticação segura para os utilizadores e gestão da identidade dos mesmos.
- Implementação:
 - Utiliza o Auth0 como fornecedor de identidade

- Controla o registo de utilizadores, o início de sessão e a gestão de sessões
- Protege os endpoints da API através de autenticação baseada em tokens

• Configuração:

- o Necessita de variáveis de ambiente:
 - AUTH0 DOMAIN
 - AUTH0_CLIENT_ID
 - AUTHO CLIENT SECRET
 - AUTH0_AUDIENCE
 - AUTH0_CALLBACK_URL

2.1.2. Role-Based Access Control (RBAC)

- **Objetivo**: Controla o acesso a diferentes funcionalidades do sistema com base nas funções dos utilizadores.
- Implementação:
 - Define diferentes funções de utilizador (utilizadores e administradores)
 - o Restringe o acesso a operações sensíveis
 - Gere permissões para:
 - Gestão de anúncios de carros
 - Marcação de manutenções
 - Gestão de inspeções
 - Marcação de reuniões

2.2. Continuous Integration and Deployment (CI/CD):

2.2.1. GitHub Actions

- **Objetivo**: Automatiza os processos de teste e de implementação.
- Componentes:
 - o test.yml: Corre testes automáticos
 - o build-push.yml: Gere a criação e o envio das imagens Docker
- Funcionalidades:
 - o Testes automatizados
 - Construção de imagens Docker
 - o Envio das imagens para o repositório de containers

2.3. Monitorização e Observabilidade

2.3.1. Prometheus

- Objetivo: Recolha e armazenamento de métricas.
- Funcionalidades:
 - o Recolhe métricas de todos os microsserviços
 - o Armazena dados em séries temporais
 - o Disponibiliza capacidades de consulta
- Acesso: Disponível via port-fowarding (porta 9090)

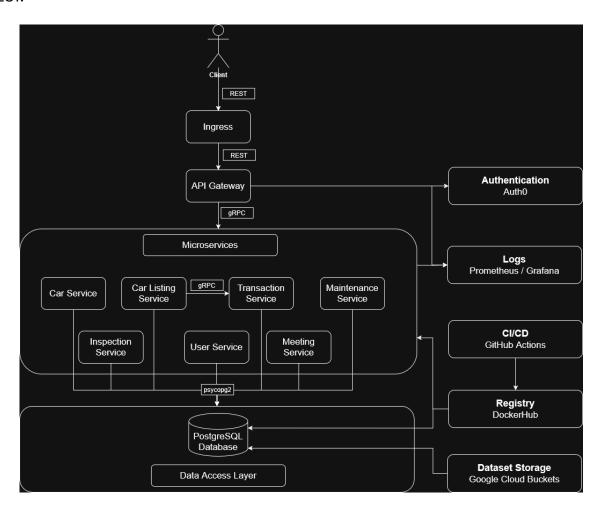
2.3.2. Grafana

- Objetivo: Visualização de métricas e criação de dashboards
- Funcionalidades:
 - o Dashboards personalizadas para diferentes métricas
 - o Monitorização em tempo real
 - o Configuração de alertas
- **Acesso:** Disponível via port-fowarding (porta 3000)

Architecture (Application and Technical)

Descrição da Arquitetura:

O sistema segue uma arquitetura de microsserviços, sendo que cada serviço gere um domínio específico. Os microsserviços comunicam entre si utilizando o gRPC, enquanto os clientes externos interagem com o sistema através de um API Gateway baseado em REST.



Gateway API:

- Atua como o único ponto de entrada para todos os pedidos externos, expondo uma API REST.
- Recebe pedidos HTTP REST e encaminha-os para os microsserviços apropriados via gRPC.

- Lida com a autenticação e a validação de pedidos utilizando o Auth0.
- Inclui a recolha de métricas do Prometheus para monitorização.
- Fornece uma interface web com templates para autenticação de utilizadores e acesso ao painel.

User Service:

- Gere as informações de Users e armazena os dados na base de dados PostgreSQL partilhada.
- Fornece operações CRUD para entidades Users.

Car Service:

- Gere as informações de Cars e armazena os dados na base de dados PostgreSQL partilhada.
- Fornece operações CRUD para entidades Car.

Car Listing Service:

- Gere informações de Car Listings e armazena os dados na base de dados PostgreSQL partilhada.
- Fornece operações CRUD para entidades de Car Listing.
- Interage com o Transaction Service para atualizações de estado após a conclusão da venda/arrendamento.

Transaction Service:

- Gere Transactions de compra e aluguer e armazena os dados na base de dados PostgreSQL partilhada.
- Fornece operações CRUD para entidades de Transaction.

Maintenance Service:

- Gere agendamentos de Maintenances e armazena os dados na base de dados PostgreSQL partilhada.
- Fornece operações CRUD às entidades de Maintenance.

Inspection Service:

- Gere Inspections periódicas de automóveis e armazena os dados na base de dados PostgreSQL partilhada.
- Fornece operações CRUD para entidades de Inspection.

Meeting Service:

- Gere Meetings entre utilizadores e funcionários do stand e armazena os dados na base de dados PostgreSQL partilhada.
- Fornece operações CRUD para entidades de Meeting.

Camada de Base de Dados:

- Base de dados PostgreSQL única partilhada entre todos os microsserviços.
- Contém tabelas para cars, users, car_listings, transactions, maintenance, inspection, and meetings.
- Garante a integridade referencial através de restrições de foreign key.
- Preenchido com dados de amostra para testes e desenvolvimento.
- Os detalhes de ligação à base de dados são configurados através de variáveis de ambiente.

Monitorização e Observabilidade:

- Todos os serviços incluem a recolha de métricas do Prometheus para monitorização.
- Os dashboards do Grafana estão disponíveis para visualização de métricas.
- As métricas incluem contagens de pedidos, histogramas de latência, pedidos ativos e tempo de operação da base de dados.

Autenticação e Autorização:

- Integração com Auth0 para autenticação de utilizadores.
- Controlo de acessos baseado em funções com permissões para diferentes operações.
- Tokens JWT utilizados para autenticação de API.
- Gestão de sessão para interface web.

CI/CD e Delivery:

- Os workflows do GitHub Actions automatizam a pipeline de integração e delivery contínua.
- Compilações, testes e entregas automatizadas são acionadas por commits na main branch
- As Docker images s\u00e3o criadas para cada microsservi\u00fco e pushed para o DockerHub.

Implementation (High Level Description)

O sistema foi implementado como uma cloud native application com arquitetura de microsserviços, incluindo uma infraestrutura completa de monitorização, segurança e deployment.

1. Infraestrutura e Deployment

1.1. Containers (Kubernetes Deployment):

- Cada serviço deployed num pod isolado.
- Service Mesh para comunicação entre serviços.
- Persistent Volume Claims para armazenamento de dados.

1.2. Infraestrutura da Base de Dados:

- PostgreSQL Container.
- Docker image dedicada.
- Persistent Volume para armazenamento de dados.
- Environment variables para configuração.
- Serviço dedicado para acesso.

1.3. Load Balancing e Escalabilidade:

NGINX Ingress Controller:

- Uso de Cookies para gestão de sessões.
- Health check path com /health.
- Suporte de protocolo HTTP/1.1.

Horizontal Pod Autoscaling:

- o Scaling dinâmico com 2-3 réplicas por serviço.
- o CPU e Memory-Based scaling (80% de threshold).
- o Rolling update strategy.
- Limites de Recursos e Requests.

Load Balancing de serviços:

- o Serviços ClusterIP para comunicação interna.
- o Round-robin load balancing.

2. Monitorização

2.1. Prometheus Setup:

- ConfigMap para configuração do Prometheus.
- Armazenamento persistente para metrics.
- Recolha de metrics de todos os microsserviços.

2.2. Grafana Integration:

- Uso de secrets para environment variables.
- Configurações de Segurança CSRF.
- Dashboards próprias para monitoring.

2.3. Health Monitoring e Logging:

• Health Checks:

- o Liveness probes.
- o Readiness checks.
- Automatic recovery.

Monitorização de Serviços:

- Health checks automáticos.
- o Recovery automática.
- o Alertas relacionados com health checks.

3. Security Implementation

3.1. Auth0 Integration:

- OAuth 2.0 implementation.
- JWT token validation.
- Session management.
- Callback URL handling.

3.2. Role-Based Access Control:

- Uso de JWT para verificação de permissões.
- Nomes customizados e adequados para permissões.
- Validação de permissões.
- Error handling para acessos não autorizados.

4. Service Architecture

4.1. Core Services com ports próprias:

API Gateway:

Implementação da REST API:

- Uso de *Flask* para endpoints *REST*.
- CRUD operations para todos os serviços.
- Uso de *RBAC*.
- Gestão de requests / responses.

o Endpoints de Serviços:

- Cars API (/api/cars).
- Users API (/api/users).
- Maintenance API (/api/maintenances).
- Inspection API (/api/inspections).
- Transaction API (/api/transactions).
- Car Listing API (/api/carlistings).
- Meeting API (/api/meetings).

Endpoints de Autenticação:

- Login (/login).
- Callback (/callback).
- Dashboard (/dashboard).
- Logout (/logout).
- Health Check (/health).

Microsserviços:

- o Car Service.
- o Car Listing Service.
- o User Service.
- Transaction Service.
- Maintenance Service.
- o Inspection Service.
- Meeting Service.

4.2. Comunicação entre serviços:

- REST API para comunicação externa.
- gRPC para comunicação entre microsserviços.
- Protocol Buffers para serialização.

5. Gestão de Dados

5.1. Database Schema:

- car: Informação de veículos.
- users: Gestão de utilizadores.
- meeting: Marcação de reuniões.
- maintenance: Gestão de manutenções.
- inspection: Registos de inspeções.
- car_listing: Gestão de listings.
- transaction: Histórico de transações.

5.2. Persistência de Dados:

- Persistent Volume Claims.
- Estratégias de Backup da base de dados.

6. Desenvolvimento e Operações

6.1. Container Management (Docker):

- Imagens Docker.
- Build e Push de imagens.
- Variáveis de Ambiente para configuração.
- Volume management.

6.2. Configurações Kubernetes:

- Deployment Strategy:
 - Rolling updates.

Service Configuration:

- o Comunicação interna entre serviços.
- o Comunicação externa pelo Ingress.

Esta implementação resulta num sistema robusto com:

- Monitorização com *Prometheus* e *Grafana*.
- Autenticação Segura com Auth0.
- Deployment escalável com containers.
- Armazenamento de dados persistente.
- Pipeline Automatizada de Deployment.
- Role-based access control.
- Load Balancing e Gestão de Sessões.
- Scaling automático com base na utilização de recursos.

O sistema foi implementado para ser:

- Altamente disponível (*Kubernetes* deployment).
- Seguro (*Auth0* integration).
- Observável (Prometheus / Grafana).
- Escalável (Containerized microservices).
- Fácil de Manter (Separação de responsabilidades clara).

Deployment Scripts

O processo de deployment é baseado em imagens Docker publicadas no Docker Hub, geradas pela pipeline CI/CD. Estas imagens são utilizadas nos manifestos de Kubernetes para deployment dos microsserviços e da base de dados. No entanto, não temos scripts de deployment totalmente automáticos. Em vez disso, utilizamos um script chamado `apply_kubernetes_manifests.sh`, que, ao ser executado, realiza o deploy para o Google Kubernetes Engine (GKE).

Devido à necessidade de obter o IP do Ingress para configuração no Auth0, optámos por um processo de deployment manual utilizando este script.

CI/CD Pipeline

Para automatizar o ciclo de desenvolvimento contínuo do projeto, foram criados dois workflows no GitHub Actions, localizados na pasta `.github/workflows`:

1. Build and Push Docker Images (`build-push.yml`):

Este workflow é ativado automaticamente em commits no branch `main` quando há alterações nas pastas `databases/` ou `microservices/`. O pipeline executa os seguintes passos:

- Faz checkout do repositório.
- Configura o Docker Buildx.
- Faz autenticação no Docker Hub utilizando `secrets` configurados no GitHub (`DOCKER_USERNAME`, `DOCKER_TOKEN`).
- Constrói e faz push da imagem Docker do serviço de base de dados.
- Identifica dinamicamente os microsserviços definidos no `docker-compose.yml`, constrói as respetivas imagens e faz push no Docker Hub com a tag `latest`.

Com isto, todas as alterações feitas aos principais componentes (bases de dados e microsserviços) são automaticamente integradas e disponibilizadas como imagens no Docker Hub.

2. Test Workflow (`test.yml`):

Este workflow é ativado em commits e pull requests para a branch `main`. Executa testes automáticos para garantir que o código continua funcional após alterações:

- Faz checkout do código.
- Instala todas as dependências necessárias.
- Gera automaticamente o código gRPC a partir dos `.proto`.
- Executa os testes com pytest, localizados na pasta `tests/`.

Este pipeline permite manter a integridade funcional do sistema ao longo do tempo e facilita a deteção de erros.

Test and Evaluation Techniques (To Group 13)

Avaliação da documentação:

A documentação fornecida pelo outro grupo estava clara, bem estruturada e precisa.

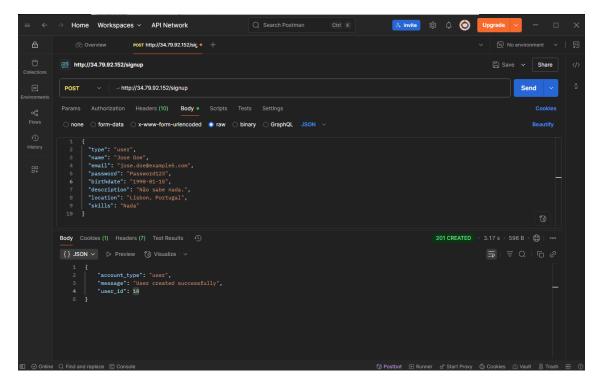
Incluía uma lista abrangente de todos os endpoints disponíveis, cada um com explicações detalhadas sobre a sua funcionalidade.

A inclusão do URL base e de uma especificação OpenAPI tornou fácil compreender como interagir com a API e testar cada endpoint de forma eficiente.

No geral, a documentação facilitou os testes e demonstrou uma forte atenção ao detalhe.

Smoke Tests:

Inicialmente, como pedido pelo grupo 13, foi necessário dar *Sign Up* para utilizar as outras funcionalidades do projeto, para isso, efetuámos um POST com os dados na imagem abaixo:



O pedido funcionou como esperado e o utilizador ficou logado.

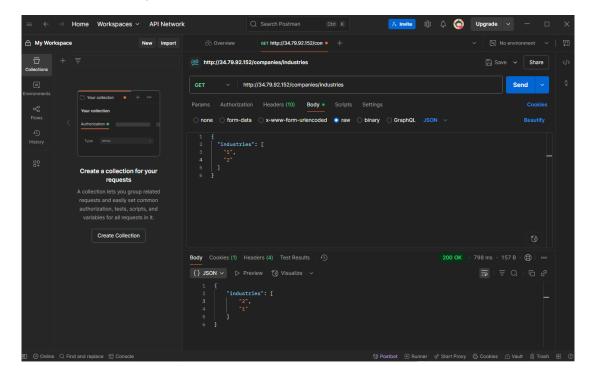
Endpoint: /companies/industries:

• **GET** /companies/industries:

o Resultado: Retornou a lista de industries com sucesso.

Status Code: 200 OK.

o **Observações:** A estrutura da response coincidiu com a documentação.

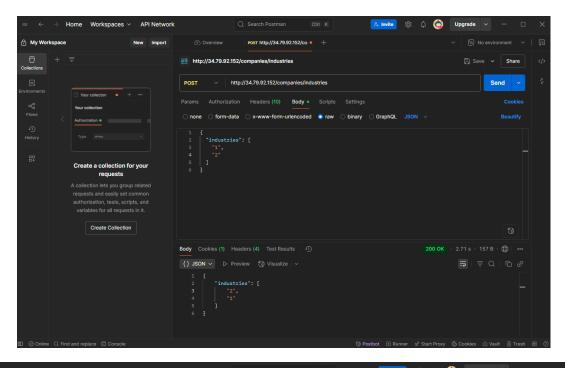


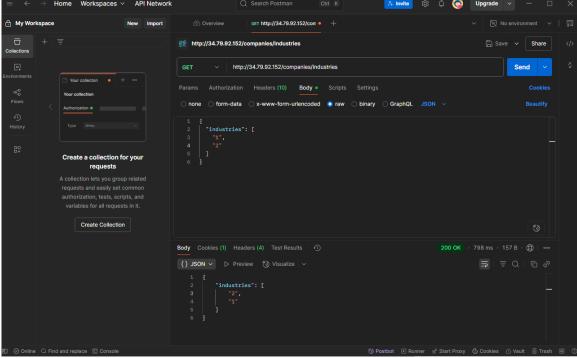
POST /companies/industries:

o Resultado: Foram criadas novas industries com sucesso.

o Status Code: 200 OK.

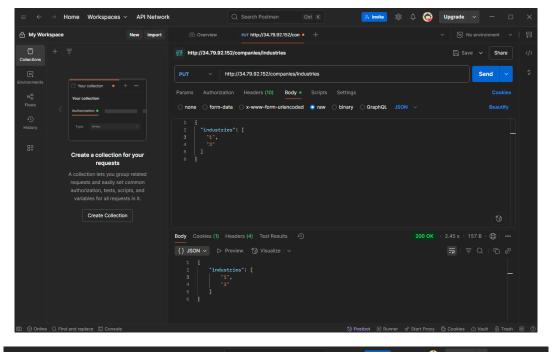
Observações: O corpo do request foi validado corretamente e os elementos adicionados apareceram corretamente em pedidos GET subsequentes.

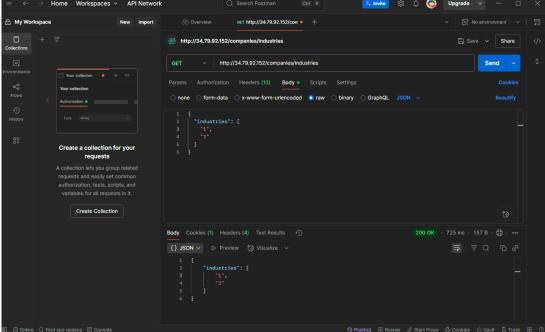




• PUT /companies/industries:

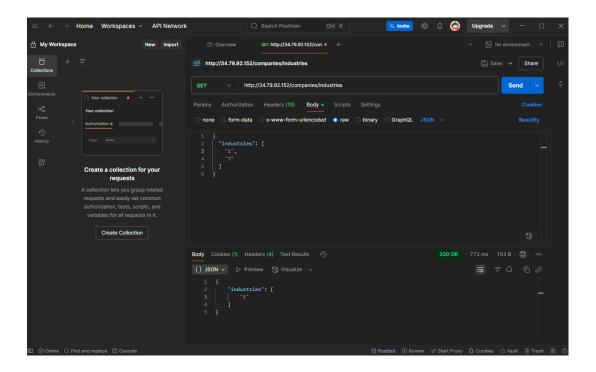
- o **Resultado:** As industries da company foram modificadas com sucesso.
- Status Code: 200 OK.
- Observações: As mudanças foram alteradas e presistidas com sucesso.





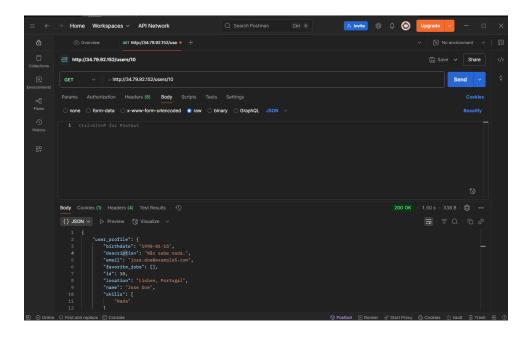
- **DELETE** /companies/industries/{name}:
 - o **Resultado:** A industry selecionada foi apagada com sucesso.
 - o Status Code: 200 OK.

 Observações: O recurso deixou de estar visível depois de apagado, como esperado.

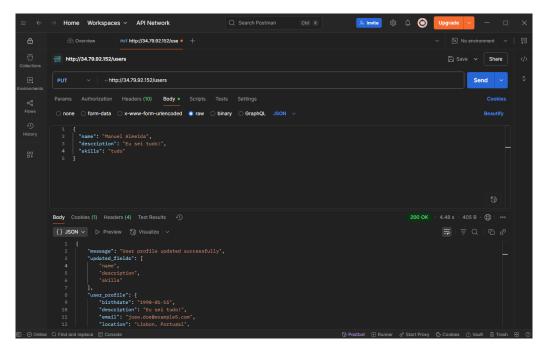


Endpoint: /users:

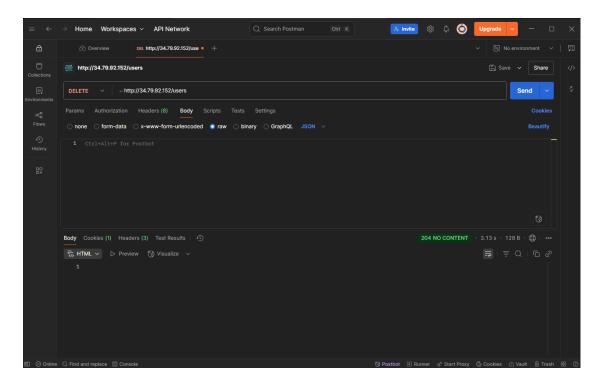
• **GET** /users/{id}: Foi obtido o utilizador que foi criado no sign up acima, funcionou como esperado.



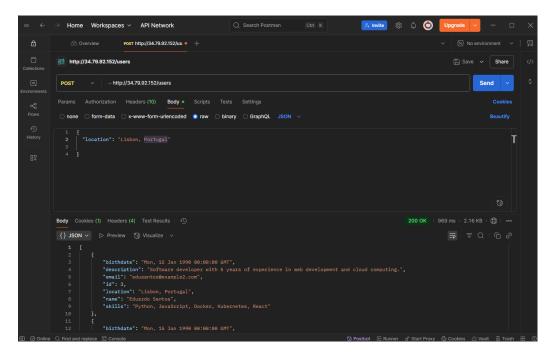
• **PUT /users:** Após tentar alterar o nome, descrição e skills, podemos ver que funcionou e as informações foram atualizadas.



 DELETE /users: Após efetuada a operação foi enviado 1 como resposta, indicando que o utilizador foi corretamente apagado.



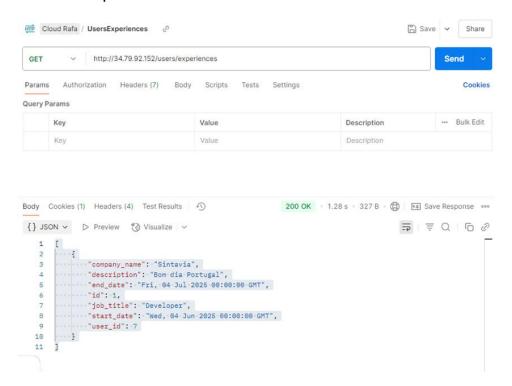
• POST /users: Foi testado com a localização e também funcionou corretamente.



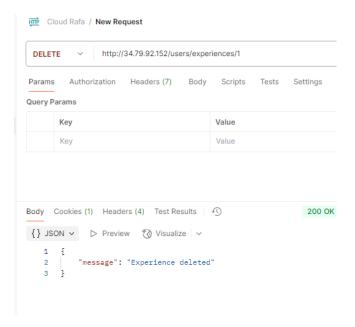
Endpoint: /users/experiences:

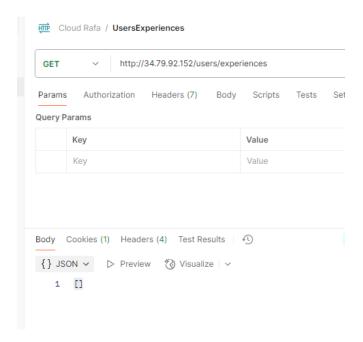
• POST /users/experiences:

• **GET** /users/experiences:



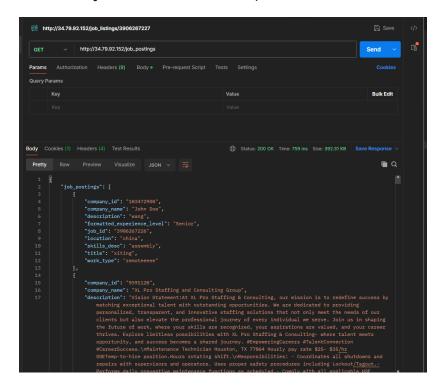
• **DELETE** /users/experiences/{id}:



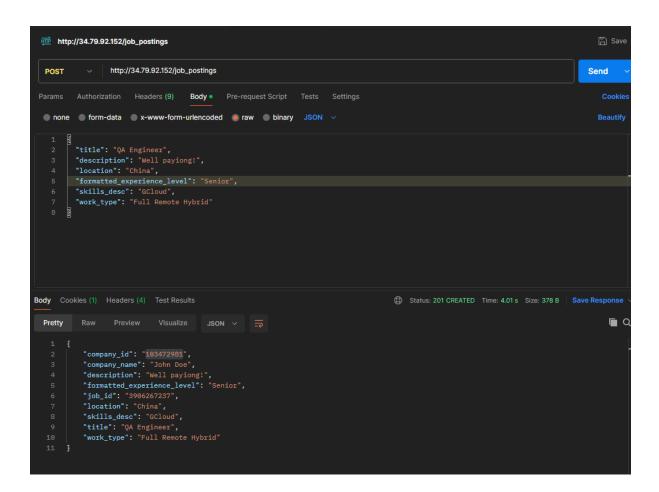


Endpoint: /job_postings:

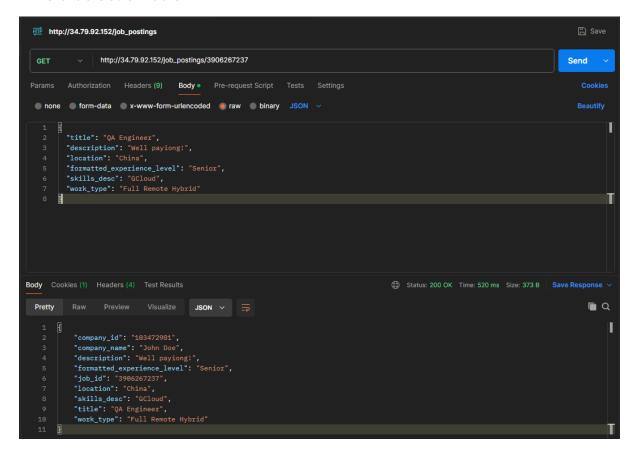
- **GET**/job_postings:
 - o **Resultado:** Retornou a lista de job postings com sucesso.
 - Status Code: 200 OK.
 - o **Observações:** A estrutura da response coincidiu com a documentação.



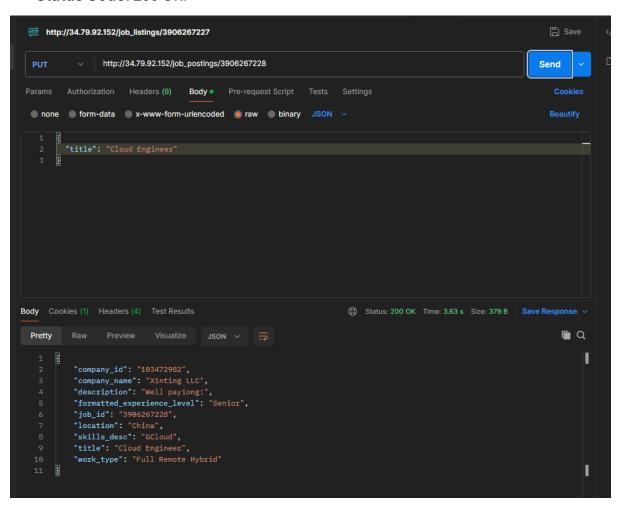
- **POST**/job_postings:
 - Resultado: Criou e retornou o job posting com a informação dada corretamente.
 - o Status Code: 201 CREATE.



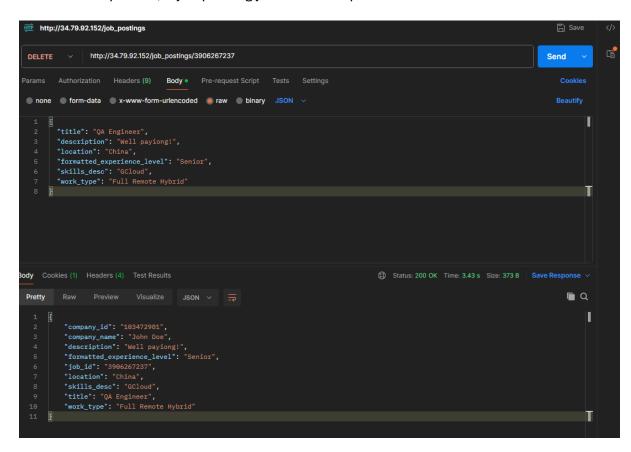
- **GET** /job_postings/{job_id}:
 - Resultado: Retornou a informação do job posting com o id correspondente.
 - Status Code: 200 OK.



- PUT /job_postings/{job_id}:
 - Resultado: Atualizou e devolveu a informação o job posting com o id correspondente.
 - Status Code: 200 OK.



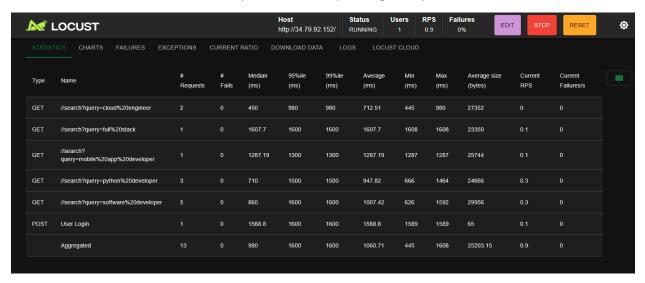
- DELETE /job_postings/{job_id}:
 - o **Resultado:** Apagou e devolveu o job posting com o id correspondente.
 - o Status Code: 200 OK.
 - Observações: Após o DELETE, voltou-se a testar os GETs e, tal como esperado, o job posting já não estava presente.



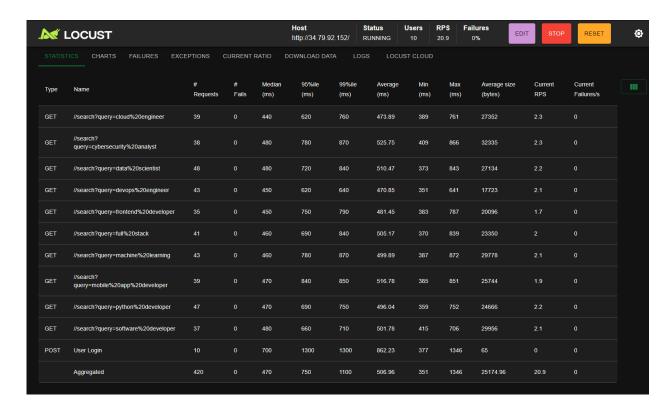
Stress Tests:

Endpoint: /search:

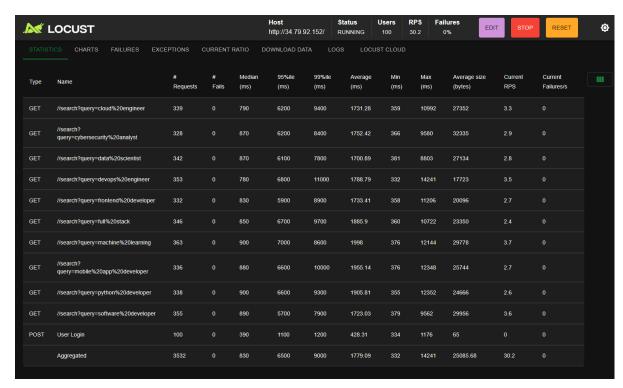
• 1 user: Máximo de 0.9 RPS (1 user criado por segundo).



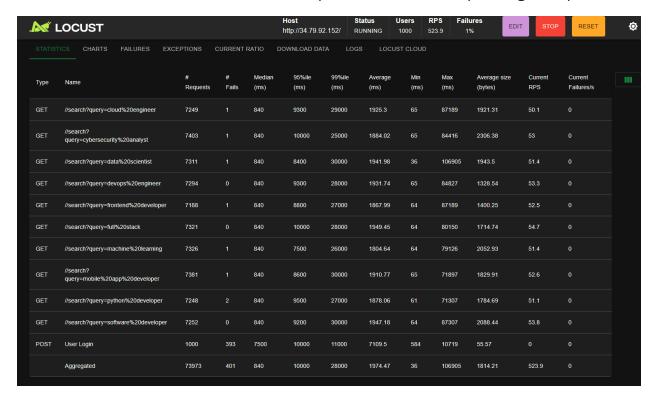
• 10 users: Máximo de 20.9 RPS (1 user criado por segundo).



100 users: Máximo de 30.2 RPS (1 user criado por segundo).



• 1000 users: Máximo de 523.9 RPS (100 users criados por segundo).



• Erros encontrados: Todos relacionados com o serviço de autenticação.

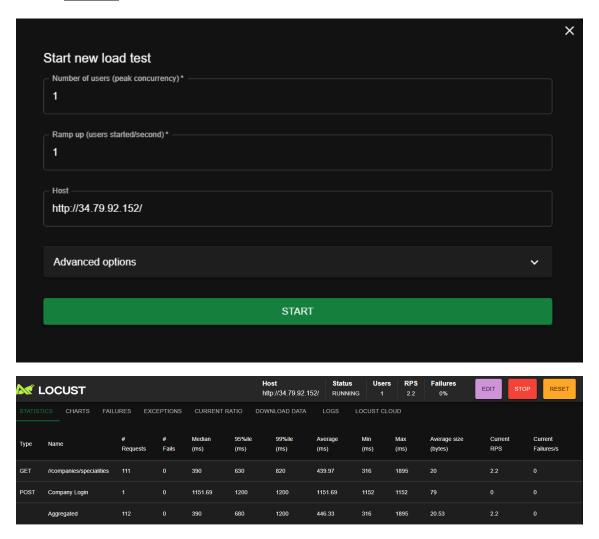
```
Failed to login: {"error":"Authentication service error"}

Failed to login: {"error":"Authentication service error"}
```

уре	Name		# reqs		# fails			Min	Max	Med	req/s	failur			
т		query=cloud%20engineer	11245		2(0.02%)	18		64	96135	- 820	47.33		0.01		
r		query=cybersecurity%20analyst	11374		(0.02%)	18	60	64	89003	810	47.88		0.01		
		guery=data%20scientist	11192		(0.02%)	18			106984	820	47.11		0.01		
		query=devops%20engineer	11251		(0.00%)	18		64	85648	820	47.36		0.00		
r		query=frontend%20developer	11163		(0.01%)	18	73	63	144611	820	46.99		0.00		
т		guery=full%20stack	11417		(0.01%)	19	29	63	149992	820	48.06		0.00		
	//search?	query=machine%20learning	11433	1	(0.01%)	18	12	63	88435	820 İ	48.13		0.00		
		guery=mobile%20app%20developer	11428		(0.01%)	19	97	64	87077	820	48.10		0.00		
	//search?	query=python%20developer	11403	3	(0.03%)	18	84	61	86669	820 İ	48.00		0.01		
		guery=software%20developer	11209		(0.00%)	18	98	63	90448	820	47.18		0.00		
ST	User Logi	n '	1000	393(39.30%)	j 71	89	584	10719	7500 İ	4.21		1.65		
				J											
	Aggregate		114115	406	(0.36%)	19	12		149992	820	480.35		1.71		
		entiles (approximated)													
уре	Name			50% 	66%	75% 	80%	98	% 95% -	98%		99.9% 9			# reqs
	//search?	query=cloud%20engineer		820	940	1000	1100	130		20000		74000			11245
	//search?	query-cybersecurity%20analyst		810	930	1000	1100	130	0 7000	19000	26000	79000	88000	89888	11374
	//search?	query=data%20scientist		820	940	1000	1100	130	0 5800	19000	29000	70000 1	02000	107000	11192
		query-devops%20engineer		820	940	1000	1100	130	0 6200	19000	27000	61000	85000	86000	11251
	//search?	query=frontend%20developer		820	940	1000	1100	130	0 5900	19000	27000	74000 1	27000	145000	1116
ET	//search?	query-full%20stack		820	940	1000	1100	130	0 8500	20000	28000	76000 1	46000	150000	11417
ET	//search?	query=machine%20learning		820	940	1000	1100	130	0 4900	19000	26000	73000	85000	88000	1143
ET	//search?	query-mobile%20app%20developer		820	940	1000	1100	130	0 5600	21000	30000	72000	84000	87000	11428
ET	//search?	query=python%20developer		820	930	1000	1100	130	0 6200	19000	26000	64000	83000	87000	11403
	//search?	query=software%20developer		820	940	1000	1100	130	0 6100	20000	29000	78000	87000	90000	11209
OST	User Logi			7500	9800	10000	10000	1000	0 10000	10000	11000		11000		1000
	Aggregate			820	940	1000	1100	140				74000 1			
rror re	port														
occurr	ences	Error													
		POST User Login: HTTPError('500 Server Error: INTERNAL SERV													
		GET //search?query=cloud%20engineer: HTTPError('502 Server													
		GET //search?query=cybersecurity%20analyst: HTTPError('502													
		GET //search?query=machine%20learning: HTTPError('500 Serve										ing')			
		GET //search?query=python%20developer: HTTPError('502 Serve													
		GET //search?query=data%20scientist: HTTPError('500 Server													
	GET //search?query=mobile%20app%20developer: HTTPError('500 Server Error: Internal Server Error for url: //search?query=mobile%20app%20deve							velope	r')						
	GET //search?query=python%20developer: HTTPError('500 Server Error: Internal Server Error for unl: //search?query-python%20developer')														
	GET //search?query-frontend%20developer: HTTPError('502 Server Error: Bad Gateway for url: //search?query-frontend%20developer')														
	GET //search?query=full%20stack: HTTPError('502 Server Error: Bad Gateway for url: //search?query=full%20stack')														
		GET //search?query=data%20scientist: HTTPError('502 Server	Error: Bad Gatev	way fo	or url: /	//searc	h?auerv	=data	%20scien	ist')					

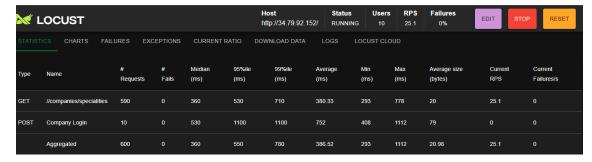
Endpoint: /companies/specialities:

• <u>1 user:</u>



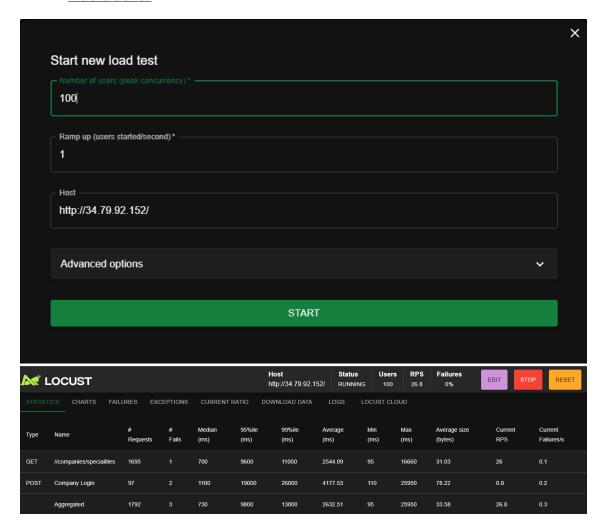
Não houve erros e havia cerca de 2 pedidos por segundo. Tudo a funcionar normalmente.

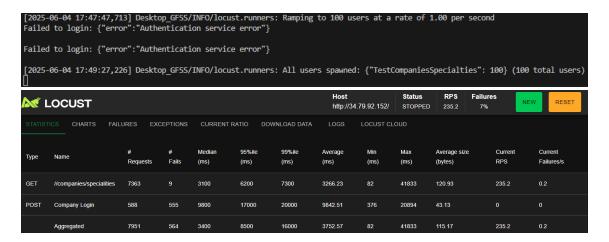
• 10 users:



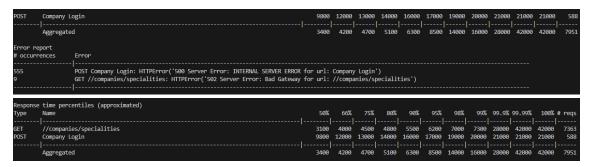
Também não existiram erros e os pedidos por segundo multiplicaram por 10 como era de esperar. Tudo a funcionar normalmente também.

• 100 users:





Com 100 utilizadores houve uma pequena percentagem de erros. Mais especificamente ao dar login, os pedidos multiplicaram-se por 10 novamente. Os erros ocorridos estão na imagem acima e nas imagens abaixo.



Endpoint: /companies:

• 1 User: Máximo de 1.5 RPS.



• 10 Users: Máximo de 19.8 RPS.



• 100 Users: Máximo de 37.5 RPS.



• 1000 Users: Máximo de 131 RPS.



Erros ao usar 1000 Users:

```
# occurrences Error

133 POST Company Login: HTTPError('500 Server Error: INTERNAL SERVER ERROR for url: Company Login')

134 GET //companies: HTTPError('502 Server Error: Bad Gateway for url: //companies')

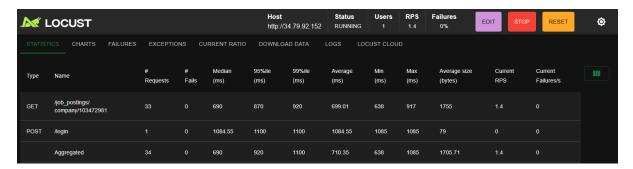
142 POST Company Login: HTTPError('502 Server Error: Bad Gateway for url: Company Login')

190 POST Company Login: HTTPError('504 Server Error: Gateway Time-out for url: Company Login')
```

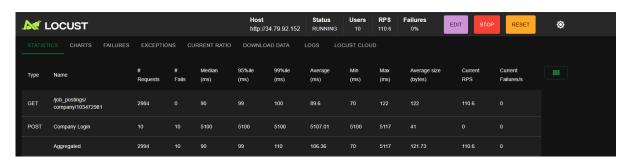
Endpoint: job_postings/company:

O endpoint em si aguentou com a carga de stress causada, no entanto, a autenticação começou a falhar com o aumento dos users, que por sua vez, fazia com que não fosse possível testar o endpoint que se queria realmente testar.

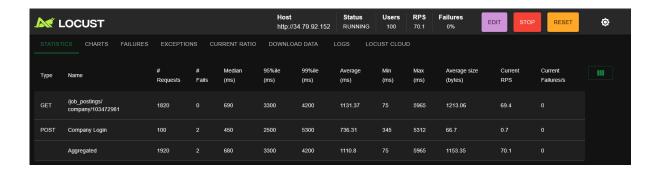
• 1 User:



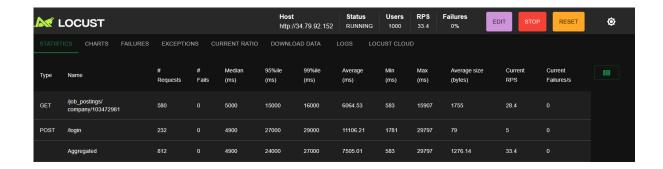
10 Users:



• 100 Users:



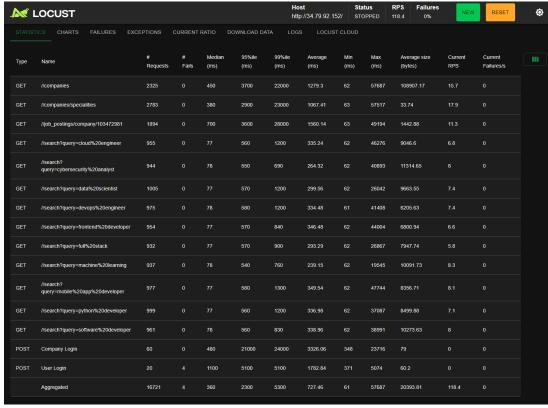
• 1000 Users (início):



• 1000 Users (começou a falhar):



<u>Todos os endpoints combinados:</u> Testámos todos os endpoints em conjunto, com **100** users totais, **1** user criado por segundo.





Feedback Final:

O sistema está funcional e bem documentado, demonstrando um bom nível de maturidade e organização no seu desenvolvimento. A documentação é clara, o que facilita a compreensão da arquitetura e a utilização dos diferentes serviços disponíveis. No entanto, existe uma área com potencial para melhoria: o serviço de autenticação.

Atualmente, o serviço de autenticação funciona corretamente, mas poderia ser otimizado para garantir um desempenho mais eficiente, especialmente em cenários de carga elevada ou picos de utilização.

Investir nesta melhoria não só aumentaria a robustez do sistema como também garantiria uma melhor experiência para os utilizadores em contextos de elevada concorrência, contribuindo para a estabilidade geral da aplicação.

Cost Analysis (FinOps)

Utilização e Custos Reais:

Durante a fase de desenvolvimento e testes do projeto, o custo total na Google Cloud Platform (GCP) foi de apenas \$10,30 nos meses de abril e maio de 2025:

• Abril 2025: \$4,14

Service description	SKU description	Usage amount	Usage unit	Cost (\$)
				4.14
Kubernetes Engine	Autopilot Pod mCPU Requests (europe-west4)	26,868.042	hour	1.32
Kubernetes Engine	Autopilot Kubernetes Clusters	11.918	hour	1.19
Cloud Monitoring	Prometheus Samples Ingested	7,761,562	Samples	0.47
Compute Engine	Balanced PD Capacity in Netherlands	2.25	gibibyte month	0.25
Networking	Cloud Load Balancer Forwarding Rule Minimum for Netherlands (europe-west4)	8.848	hour	0.24
Kubernetes Engine	Autopilot Pod Memory Requests (europe-west4)	37.334	gibibyte hour	0.20
Networking	Networking Private Service Connect Partner Select End Point	11.359	hour	0.11
Networking	Network Intelligence Center Network Analyzer Resource Hours	86	count	0.09
Networking	Network Intelligence Center Topology and Google Cloud Performance Resource Hours	86	count	0.09
Networking	Network Intelligence Center Internet to Google Cloud Performance Resource Hours	86	count	0.07
Networking	Cloud Load Balancer Forwarding Rule Minimum Global	1.895	hour	0.05
Kubernetes Engine	Autopilot SSD Pod Ephemeral Storage Requests (europe-west4)	150.14	gibibyte hour	0.02
Compute Engine	Network Internet Data Transfer Out from Netherlands to EMEA	0.161	gibibyte	0.02

• Maio 2025: \$6,16

Service description	SKU description	Usage amount	Usage unit	▼ Cost (\$)
				6.16
Kubernetes Engine	Autopilot Pod mCPU Requests (europe-west4)	42,661.097	hour	2.09
Kubernetes Engine	Autopilot Kubernetes Clusters	19.189	hour	1.92
Networking	Cloud Load Balancer Forwarding Rule Minimum for Netherlands (europe-west4)	16.095	hour	0.44
Compute Engine	Balanced PD Capacity in Netherlands	4.014	gibibyte month	0.44
Cloud Monitoring	Prometheus Samples Ingested	6,387,996	Samples	0.38
Kubernetes Engine	Autopilot Pod Memory Requests (europe-west4)	70.077	gibibyte hour	0.38
Networking	Networking Private Service Connect Partner Select End Point	18.364	hour	0.18
Networking	Network Intelligence Center Network Analyzer Resource Hours	87	count	0.10
Networking	Network Intelligence Center Topology and Google Cloud Performance Resource Hours	87	count	0.10
Networking	Network Intelligence Center Internet to Google Cloud Performance Resource Hours	87	count	0.07
Kubernetes Engine	Autopilot SSD Pod Ephemeral Storage Requests (europe-west4)	184.741	gibibyte hour	0.03
Compute Engine	Network Internet Data Transfer Out from Netherlands to EMEA	0.1	gibibyte	0.01

Este custo reduzido deveu-se não só à natureza leve do ambiente de desenvolvimento, mas também a uma estratégia consciente de contenção de custos, em que se privilegiou o desenvolvimento local em Minikube, utilizando o GKE apenas para testes pontuais e validações finais:

- GKE Autopilot (CPU, memória, horas de cluster): ~70% do custo total
- Monitorização e logging
- Redes e Armazenamento Básico (Load Balancer, Persistent Disk)

Custo Estimado em Operação Total:

Foi realizada uma estimativa de custos com base na configuração de recursos utilizada durante o desenvolvimento e nos resultados de um teste de stress com 200 utilizadores simultâneos, número esse que coincide com a quantidade de utilizadores que esperamos atingir numa fase inicial após o lançamento público do sistema.

Durante este teste, foram identificadas limitações em dois microsserviços, causadas por falta de recursos. Como medida preventiva, decidiu-se aumentar a alocação atual de recursos na estimativa para ambiente de produção, permitindo uma distribuição mais adequada entre os microsserviços com maior exigência.

O custo mensal estimado para suportar essa carga é de aproximadamente 1627,07€, com os seguintes destaques:

- CPU, RAM em instâncias do GKE: ~1590 €
- Serviços de Rede e Load Balancing: ~20 €
- Armazenamento Persistente e Hyperdisk: ~6 €

service_display_name	name	quantity	region -	total_price, EUR 🔻
GKE (Kubernetes Engine)	Zonal Kubernetes Clusters	0.0	global	0
GKE (Kubernetes Engine)	Regional Kubernetes Clusters	0.0	global	0
GKE (Kubernetes Engine)	Balanced PD Capacity in Netherlands	50.0	europe-west4	4.86282
GKE (Kubernetes Engine)	Custom Instance Ram running in Netherlands	39420.0	europe-west4	170.68304
GKE (Kubernetes Engine)	Custom Instance Core running in Netherlands	43800.0	europe-west4	1415.12093
Instances (Compute Engine)	Spot Preemptible N4 Instance Core running in Netherlands	1460.0	europe-west4	10.61086
Instances (Compute Engine)	Spot Preemptible N4 Instance Ram running in Netherlands	5840.0	europe-west4	4.82265
Instances (Compute Engine)	Hyperdisk Balanced Capacity in Netherlands	10.0	europe-west4	0.74269
Cloud Load Balancing (Networking)	Regional External Application Load Balancer Inbound Data	F 100.0		0.79574
Cloud Load Balancing (Networking)	Regional External Application Load Balancer Outbound Da	ta 100.0		0.79574
Cloud Load Balancing (Networking)	Regional External Proxy Network Load Balancer Forwardin	g 1.0		17.74931
Cloud Logging (Cloud Operations)	Log Storage cost	50.0	global	0
Cloud Logging (Cloud Operations)	Log Retention cost	100.0	global	0.88415
				1627.06792

Conclusions

Conseguimos atingir os objetivos principais que tínhamos planeado para este projeto. Escolhemos um tema que gostamos e conseguimos contribuir para a cloud, aplicando conceitos e práticas que fomos aprendendo ao longo do semestre. Durante a segunda fase do projeto, identificámos áreas que poderiam ter sido aperfeiçoadas, mas devido à gestão de múltiplos projetos de várias cadeiras em simultâneo, não conseguimos dedicar o tempo necessário para as melhorias desejadas.

Apesar disso, estamos satisfeitos com os resultados alcançados e com o conhecimento técnico adquirido sobre a cloud e os seus componentes relacionados. Este projeto foi uma ótima oportunidade para crescer as nossas competências e explorar uma "nova" tecnologia.