Trabalho final

Data limite para entrega do trabalho no Moodle: **29 de maio de 2024 às 23:59**
Apresentações em sala de aula:

- Turma diurna LEIRT/LEIM última aula do semestre (**31 de maio de 2024**);
- Turma noturna LEIC/LEIM última aula do semestre (**3 de junho de 2024**);
- Turma diurna LEIC últimas aulas do semestre (**3 e 6 de de junho de 2024**);

Apoio ao desenvolvimento:
Para além do desenvolvimento que cada grupo realizará autonomamente, haverá aulas específicas, predominantemente nas aulas de 3 horas, para apoio à realização do trabalho.
Componentes a entregar:
1) Ficheiro ZIP com as componentes desenvolvidas, incluindo ficheiros README com informações sobre configurações, pressupostos de execução, teste ou outros
2) Documento em formato PDF com descrição da solução: Diagramas de arquitetura, contratos protobuf e pressupostos entre as partes envolvidas, formatos de dados e mensagens envolvidos nas interações, bem como os aspectos relevantes da implementação, eventuais pontos de falha e objetivos que foram e não foram atingidos

Objetivos: Saber planear e realizar um sistema para submissão e execução de tarefas de computação na nuvem, com requisitos de elasticidade, utilizando de forma integrada serviços da Google Cloud Platform para armazenamento, comunicação e computação, nomeadamente, Cloud Storage, Firestore, Pub/Sub, Compute Engine e Cloud Functions.

Descrição: Desenvolva um sistema, designado *CNV2024TF*, para detetar características (*labels*), como por exemplo, *tree*, *street*, *night*, *cat, fish*, em ficheiros de imagem (JPG, PNG, etc.) e traduzir essas características de inglês para português. O sistema deve ter a característica de elasticidade, aumentando ou diminuindo a capacidade de processamento de imagens. As funcionalidades do sistema estão disponíveis para aplicações cliente através de dois serviços gRPC, SF e SG, alojados no mesmo servidor, um para operações funcionais e outro para a gestão da elasticidade do sistema.

A. Operações funcionais (SF):

- Submissão de um ficheiro imagem para deteção de características. A operação recebe o conteúdo do ficheiro em *stream* de blocos, guardando o mesmo como um blob no serviço Cloud Storage. No final, a operação retorna um identificador do pedido (por exemplo, uma composição única entre o nome do bucket e do blob) que será usado posteriormente para obter as características detectadas na imagem e a respetiva tradução.
- A partir de um identificador retornado na chamada à operação anterior, obter a lista de todas as características encontradas na imagem e respetivas traduções, bem como a data em que a imagem foi processada;

- Obter todos os nomes de ficheiros armazenados no sistema entre duas datas e que contêm uma determinada característica (por exemplo imagens com gatos);
- [opcional] Outras operações que considere relevantes (por exemplo, *download* de uma imagem das que foram encontradas na operação anterior).

As operações de submissão são disponibilizadas através de um servidor gRPC, o qual funciona como a fachada do sistema, isto é, a aplicação cliente não conhece nada sobre a plataforma GCP. Para aumentar a disponibilidade e o balanceamento de carga do sistema, devem existir várias instâncias do servidor gRPC, cada uma a executar-se numa VM de um *instance group*.

B. Operações para gestão de elasticidade (SG):

- Aumento e diminuição do número de instâncias de servidor gRPC com operações funcionais;
- Aumento e diminuição do número de instâncias da aplicação de processamento de imagens.

A arquitetura do sistema *CNV2024TF* usa vários serviços GCP, sendo as diferentes interações, entre os seus vários componentes, apresentadas na Figura 1.

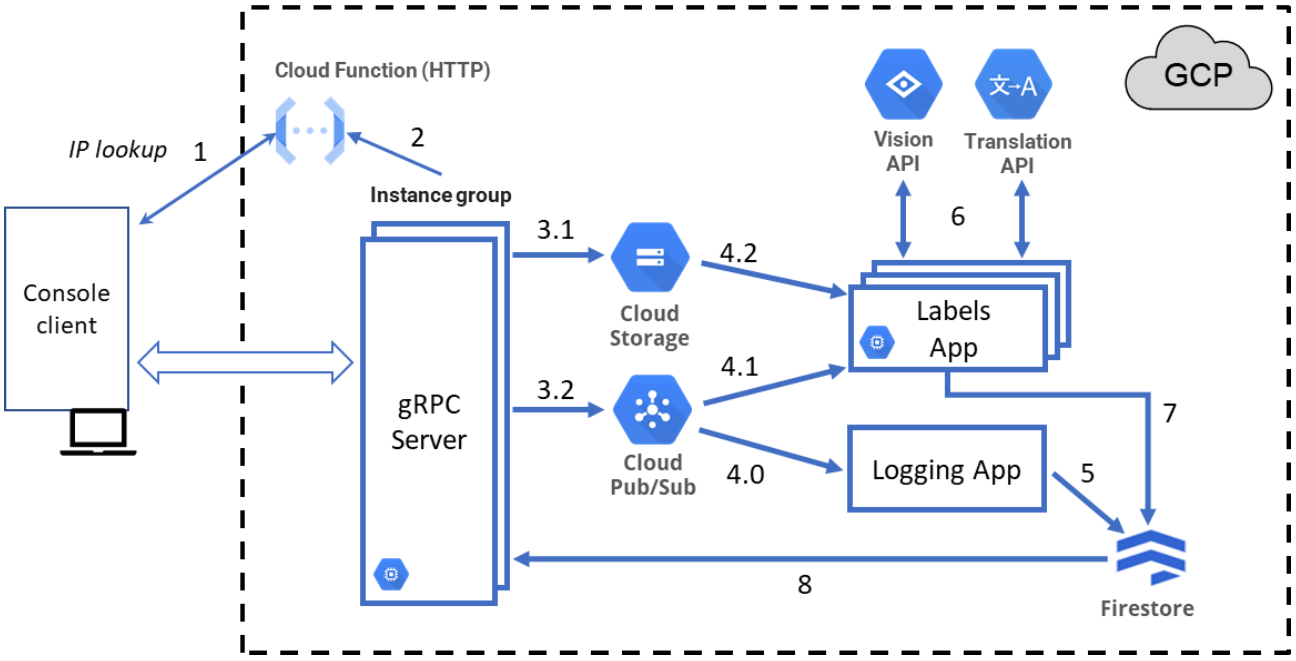

Figura 1: Componentes do *CNV2024TF* e principais interações funcionais

- O serviço Cloud Storage armazena as imagens a processar;
- O serviço Firestore guarda a informação relevante sobre processamento de um ficheiro, nomeadamente o identificador do pedido, data do processamento, as características detetadas nas imagens e traduções, ou outros que achar convenientes;

- O serviço Pub/Sub é usado para a troca desacoplada de mensagens entre os servidores de gRPC e as aplicações de processamento de imagens (*Labels App*);
- O serviço Compute Engine aloja máquinas virtuais e *instance groups* onde se executam os servidores de gRPC e as aplicações (*Labels App*) de deteção de características de imagens;
- O serviço Cloud Functions é usado para implementar uma *cloud function* para *lookup* dos endereços do grupo de servidores gRPC (*trigger* HTTP).

Fluxo de operações: Tendo em conta os números de sequência de ações, apresentados na Figura 1, a lista seguinte resume cada uma das funcionalidades:

- O serviço *Lookup Function,* usado pela aplicação cliente (1) para obtenção dos endereços IP dos servidores gRPC, deve ser desenvolvido como uma *Cloud Function* que obtém (2) os endereços IP das VM que fazem parte do *instance group*. A aplicação cliente permite ao utilizador escolher um IP da lista retornada pela função. Em caso de falha de ligação ao servidor gRPC através do IP escolhido, o utilizador tenta outro IP ou repete o processo de *lookup* para atualizar a lista dos IP até estabelecer uma ligação;
- Após a submissão de uma imagem, a mesma é guardada no Cloud Storage (3.1) e é retornado ao cliente gRPC um identificador único para posteriormente ser possível realizar as interrogações. De seguida, é enviado para um tópico Pub/Sub (3.2) uma mensagem que contém o identificador do pedido, o nome do *bucket* e do *blob,* onde ficou a imagem. A mensagem será posteriormente processada pela aplicação de deteção de labels;
- Com o objetivo de registar (*logging*) no Firestore todas as solicitações de processamento, existe uma subscrição no tópico para onde são enviados os pedidos, a qual é subscrita apenas pela aplicação Logging App (*fan-out pattern*). A *Logging App* recebe assim as mesmas informações enviadas para as aplicações de processamento (4.0). No Firestore são guardadas (5) estas informações numa coleção de nome *Logs* dedicada para o efeito;
- Associado ao tópico referido anteriormente existe uma outra subscrição partilhada por vários *workers* (*work-queue pattern*). Um *worker* (aplicação *Labels App*) de análise de imagem recebe, em cada mensagem, o nome do *bucket* e do *blob* da imagem a processar (4.1) que permite obter uma referência global (URI `gs://`) do Cloud Storage (4.2), interagindo depois com o serviço Vision API (6) para identificação de *labels* e com o serviço de tradução (6) para traduzir cada uma das *labels* detetadas de inglês para português ou outra língua;
- Após o processamento da imagem, são guardadas no Firestore (7) as informações relevantes do pedido e do resultado da análise;

- A aplicação cliente, a qualquer momento, usando o identificador do pedido, pede ao servidor gRPC informações sobre as imagens submetidas. Para retornar essa informação o servidor gRPC consulta o Firestore (8).

Aspetos de implementação:

- A API de visão faz deteção de características sobre imagens (jpg, png, etc.), retornando também para cada característica uma pontuação de confiança (entre 0 e 1). (https://cloud.google.com/vision/docs/labels) (https://cloud.google.com/vision/docs/labels#label_detection_requests)
- A API de tradução recebe, na forma de *strings*, um texto para traduzir, a língua desse texto (podendo ser também detetada automaticamente), e a língua do texto de destino. As línguas têm siglas pré-definidas, por exemplo, “pt”, “en”, “es” ou “it”. https://github.com/googleapis/google-cloud-java/blob/main/google-cloud-examples/src/main/java/com/google/cloud/examples/translate/snippets/DetectLanguageAndTranslate.java

Critérios de avaliação do trabalho:

- 30% - qualidade do relatório, que permita a um leitor entender claramente a arquitetura e as decisões de interação entre as partes, evitando apresentar código, exceto se o mesmo ajudar a explicar detalhes relevantes. O relatório deve indicar os pressupostos assumidos, indicando eventuais comparações com outras decisões possíveis. Deve constar no relatório qual a(s) parte(s) onde cada elemento do grupo teve mais ou menos responsabilidade.

- 60% - Operacionalidade, simplicidade e flexibilidade das soluções, nomeadamente na configuração e utilização da solução;
  - Nesta avaliação será ponderado o resultado da apresentação da funcionalidade da solução a toda a turma nas aulas da última semana de aulas. Para tal, será posteriormente estabelecido para cada grupo um calendário de apresentação, bem como um guião dos aspetos principais a demonstrar.

- 10% - participação individual de cada elemento do grupo durante as aulas afetas à realização do trabalho, bem como na apresentação do trabalho à turma.

José Simão

Luís Assunção

Fernanda Passos