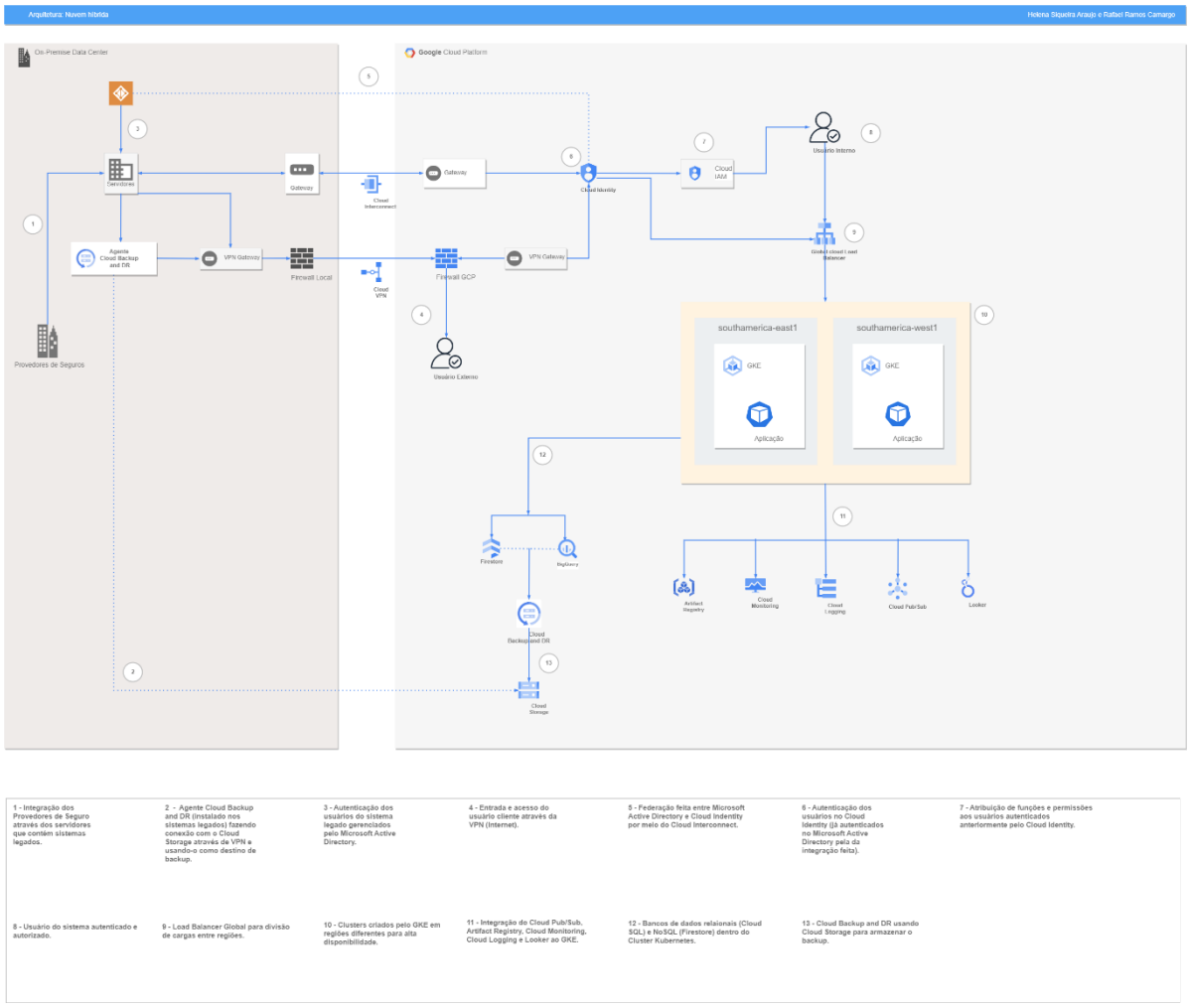
Documentação do projeto de migração da EHR Healthcare para nuvem híbrida

Arquitetura: Sistema de Híbrido



Aplicações Locais

- ✓ Os provedores de seguro (Operadoras de planos de saúde) são integrados com os sistemas legados, que se encontram nos servidores das instalações On-premise, onde uma equipe cuida do suporte dessa aplicação. Os sistemas legados serão mantidos no ambiente On-premise, enquanto os bancos de dados serão migrados para o GCP.
- ✓ O backup dos sistemas legados será feito através de um agente Cloud backup and DR que estará instalado em tais sistemas, fazendo conexão através da VPN com o Cloud Storage, para realizar o armazenamento do backup.

Gestão de Identidade e Acesso

Usuários Internos

- Os usuários internos serão todos os colaboradores da empresa EHR Healthcare. Isso engloba desenvolvedores, administradores de sistemas e outros colaboradores envolvidos na gestão e operação dos sistemas.
- No ambiente local, há o Microsoft Active Directory, que gerencia os usuários que entram no sistema.
- O Microsoft Active Directory é federado com o Cloud Identity, permitindo que, com a autenticação bem-sucedida do usuário interno no ambiente local pelo AD, tais credenciais do usuário sejam estendidas ao GCP, fazendo com que o usuário faça login único, sem a necessidade de fornecer credenciais a mais (Single Sign-On).
- Para uma camada a mais de segurança, será usada a autenticação multifator (MFA) mesmo após o SSO (Single Sign-On) bem-sucedido.
- Após a autenticação, os usuários internos serão associados a papéis específicos no Cloud IAM, em seguida, terão acesso ao sistema como usuários internos autenticados e autorizados.

Usuários Externos

- Os Usuários Externos serão todos os clientes que acessarem o sistema de fora da empresa. Isso engloba médicos, pacientes e outras partes interessadas fora da empresa EHR Healthcare.
- Os usuários externos terão um software de cliente VPN instalado em seus dispositivos (computadores ou dispositivos móveis). Isso irá permitir que eles se conectem à rede privada da empresa.
- A empresa EHR Healthcare irá configurar um servidor VPN na infraestrutura do GCP ou em outro local designado. Este servidor VPN é configurado para aceitar conexões seguras dos clientes VPN (software ou aplicativo instalado nos dispositivos dos usuários).
- Os usuários externos irão fornecer suas credenciais de autenticação ao cliente VPN, que, por sua vez, os autentica no servidor VPN usando métodos como nome de usuário/senha, certificados digitais ou outras formas seguras de autenticação.
- Uma vez autenticados, é estabelecido um túnel VPN seguro entre o dispositivo do usuário externo e o servidor VPN da empresa. Esse túnel protege a comunicação entre o dispositivo do usuário e a rede privada da empresa.
- Com o túnel VPN estabelecido, os usuários externos poderão acessar recursos internos da empresa, como aplicativos, serviços e sistemas de TI, como se estivessem fisicamente dentro da rede privada.

Configuração de conexão híbrida

Será usada a abordagem de configuração híbrida, onde o sistema terá uma conexão dedicada como principal, para cargas de trabalho que exigem maior largura de banda e menor latência e uma conexão VPN, para que haja uma alta disponibilidade.

Conexão primária (Cloud Interconnect)

- O Cloud Interconnect oferece interconexões dedicadas que são conexões físicas de alta velocidade entre a infraestrutura local e o Google Cloud Platform. Elas oferecem baixa latência e alto desempenho, fazendo com que seja ideal para cargas de trabalho sensíveis à latência, como no caso da empresa EHR, que necessita de uma troca de dados rápida entre ambientes locais e recursos na nuvem.

Conexão secundária (Cloud VPN)

- O GCP oferece a opção de criar VPNs para estabelecer uma conexão segura entre o ambiente on-premise e os recursos na nuvem. No caso do sistema em questão, a VPN como conexão secundária atua como medida de contingência para garantir a continuidade dos serviços, mesmo em situações de falha.
- A VPN utiliza túneis que criptografam o tráfego entre os pontos finais, o que faz com que se tenha uma conexão protegida de ameaças externas entre a infraestrutura local e o ambiente GCP, garantindo que os dados não sejam interceptados por indivíduos não autorizados.
- Será feita a configuração do protocolo HTTPS para garantir a segurança da comunicação entre os usuários externos e os serviços GCP.
- O uso de gateways e firewalls no sistema será feito para garantir a segurança e o controle do tráfego de dados.

Cloud Global Load Balancer e distribuição de regiões

- O Global Load Balancer será usado para distribuir o tráfego globalmente entre as instâncias nas duas regiões escolhidas, oferecendo alta disponibilidade (as solicitações serão direcionadas para instâncias mais próximas e disponíveis) e escalabilidade global, fazendo com que os usuários tenham uma melhor experiência.

Regiões

Configura-se duas regiões para ter uma maior disponibilidade e recuperação de desastres na região southamerica-east1 e na região southamerica-west1 recebendo carga de trabalho do Global Load Balancer.

A região southamerica-east1 localizada em São Paulo (Brasil) foi escolhida como região principal visando a maior proximidade possível para uma menor latência. Apesar do preço elevado, deve-se levar em conta regulamentações de privacidade de dados (LGPD) e a menor latência possível nessa região. Também visando manter a conformidade regulatória, essa também será a região escolhida para armazenar backups.

A região southamerica-west1 localizada em Santiago (Chile) foi escolhida como região para alta disponibilidade. Utilizando essas duas regiões, a latência do sistema será a menor possível.

Imagem retirada do site oficial da google: https://cloud.google.com/about/locations?hl=pt-br#lightbox-regions-map

Legenda:

- ❑ Região principal
- ❑ Região secundária

- A região southamerica-east1 será configurada como região principal para o tráfego normal e recuperação de desastres.
- A região southamerica-west1 será usada para alta disponibilidade.
- O Global Load Balancer distribuirá o tráfego entre essas duas regiões se tudo estiver em seu funcionamento correto. Em caso de falhas na região southamerica-east1, o Global Load Balancer estará configurado para redirecionar o tráfego automaticamente para a southamerica-west1.
- Para redundância geográfica os Clusters GKE serão implantados em cada região, de forma que serão independentes.

Zonas

- Para que haja alta resiliência a falhas dentro das regiões, o Cluster será configurado como multi-zonal em cada uma das regiões.
- Cada região terá três zonas de disponibilidade.
  - → Para a região southamerica-east1: southamerica-east1-a, southamerica-east1-b, southamerica-east1-c;
  - → Para a região southamerica-west1: southamerica-west1-a, southamerica-west1-b e southamerica-west1-c.
- Uma zona em cada região será escolhida como região principal, onde ela receberá o Cluster mestre. O cluster mestre distribuirá seus nós para as outras duas regiões. As zonas southamerica-east-1-a e southamerica-west1-a (da região principal e da região secundária respectivamente) serão escolhidas como zonas principais para a implantação do cluster mestre, e as demais receberão os nós.

Google Kubernetes Engine (GKE)

- Ao criar Clusters pelo GKE, será especificado o número de nós, o tipo de máquina virtual, zonas do GCP além de outras configurações relevantes para a implantação do Cluster. O GKE cuida da criação e configuração dos nós automaticamente, já que o mesmo é um serviço de orquestração de contêineres projetado para automatizar o deployment.
- Com a gestão do GKE, o sistema se beneficiará da automação em termos de escalabilidade, atualizações, configurações e manutenções.
- No GKE estarão configurados os serviços do Artifact Registry, Cloud Monitoring, Cloud Logging, Cloud Pub/Sub, Looker.
- Dentro dos clusters serão usados os bancos de dados BigQuery e Cloud Firestore.
- Será utilizado o serviço Cloud Backup e DR, onde o mesmo usará o Cloud Storage para armazenar os dados.

Serviços integrados ao GKE

Artifact Registry

Utilizado para armazenar artefatos de software no GCP, neste sistema o Artifact Registry será usado para o armazenamento de imagens de contêiner, integrando seus fluxos de trabalho ao GKE.

- Haverá um rastreamento detalhado sobre qualquer mudança feita nos artefatos.
- O Artifact Registry pode ser usado para recuperar os artefatos essenciais como parte do plano de backup e recuperação.
- Haverá integrações com outros serviços GCP que são componentes do sistema construído:
  - → Integração com GKE: Visando simplificar o processo de atualização de software, os clusters GKE serão configurados para que quando novas versões das aplicações forem implantadas, durante a implantação os clusters sejam capazes de puxar imagens de contêineres diretamente do Artifact Registry.
  - → Integração com Looker: Através do Looker será possível realizar uma integração por intermédio de APIs com o Artifact Registry para que possam ser analisados os dados sobre versões, distribuições e mudanças nos artefatos.
  - → Integração com Cloud Pub/Sub: Qualquer evento sobre alteração nos artefatos será notificado.

Cloud Monitoring

O Google Cloud Monitoring será responsável por monitorar e alertar sobre o desempenho, a disponibilidade e a integridade dos aplicativos e recursos no ambiente GCP. Ele ajudará a identificar problemas rapidamente para que a confiabilidade contínua seja garantida.

- O Cloud Monitoring terá alertas configuráveis para notificação de eventos importantes.
- Haverá monitoramento de métricas em tempo real.
- O Cloud Monitoring também tem o recurso de painéis personalizáveis para visualização de dados de monitoramento.

Cloud Logging

Pelo Cloud Logging será possível coletar, pesquisar e analisar registros gerados por aplicativos e serviços GCP. Desta forma, ele desempenha um papel fundamental para o sistema em questão para realizar análises de segurança e atender a requisitos de conformidade.

- O Cloud Logging será integrado com o Cloud Monitoring a fim de fornecer alertas baseados em condições específicas nos logs.
- Os alertas serão configurados para avisar sobre eventos críticos de forma automática.
- Permitirá a coleta e análise de registros específicos de aplicações dos clusters implantados.
- O Cloud Logging coleta dados de serviços distintos, podendo ser contêineres, aplicativos, serviços em execução ou até mesmo máquinas virtuais.

Google Cloud Pub/Sub

O Google Cloud Pub/Sub será usado para notificação por mensagens em tempo real, fazendo com que o sistema seja escalável, flexível e resiliente.

- Os consumidores e produtores de mensagens não precisam estar ativos ao mesmo tempo para utilizar do serviço, pois o Cloud Pub/Sub é assíncrono, o que aumenta a escalabilidade do sistema.
- O consumidor terá a opção de receber e processar mensagens através de subscrições, que são pontos de extremidades onde são recebidas as mensagens de um tópico.
- Consumidores recebem mensagens através dos tópicos, que são a fonte de comunicação do sistema, por onde os produtores publicam as mensagens. Tais mensagens são dados enviados através do serviço em questão.
- O Cloud Pub/Sub possui controle de acesso baseado em função, autenticação de clientes e criptografia de dados.

Looker

O Looker é uma plataforma de Business Intelligence que permitirá aos usuários internos do sistema comunicar insights de maneira eficaz, bem como explorar, analisar, visualizar dados de forma interativa e também impactará na tomada de decisões baseadas em dados.

- Haverá criação de dashboards, gráficos de tipos variados e opções para criação de visualizações personalizadas.
- Facilitará colaboração entre equipes, pois permitirá o compartilhamento de painéis, descobertas e relatórios.
- Haverá integração entre Looker e BigQuery ML (usuários terão a opção de acessar esses recursos).
- Relatórios e dashboards serão atualizados de forma automática.
- Dado o nível de sensibilidade dos dados que aqui se encontram, o Looker oferece trilhas de auditoria, auditoria de dados e acesso baseado em função.

Bancos de dados

BigQuery (para bancos de dados relacionais)

O BigQuery é altamente eficiente para bancos de dados relacionais e suporta dados estruturados, semi-estruturados e não estruturados, o que faz com que seja possível a extração de insights de estruturas que não seguem o modelo de tabela.

- Oferece suporte à linguagem SQL para consultas tradicionais em tabelas.
- Visando uma otimização de custo, as práticas recomendadas serão adotadas: compromisso de compra para que haja descontos, otimização de clustering, uso de particionamento e monitoramento e ajuste das configurações a medida que a empresa EHR expande para além do nível estadual e os requisitos da carga de trabalho aumentam.
- Possui escalabilidade automática para lidar com grande volume de dados devido ao fato de ser totalmente hospedado na nuvem.
- As consultas serão rápidas e eficientes mesmo que os dados sejam grandes: O BigQuery faz uso de um modelo de processamento distribuído, onde conjuntos maiores de dados são divididos em menores partes.
- O BigQuery será integrado Cloud Storage.
- Será utilizado o BigQuery ML integrado ao Looker para que sejam treinados os modelos de machine learning dos dados que estão armazenados no BigQuery de forma direta.
- Com o uso de machine learning, é possível fazer análises preditivas:
  - → Auxílio nas detecções precoces de doenças baseadas em exames dos pacientes.
  - → Previsão de risco de doenças pelo histórico dos pacientes.
  - → Insights valiosos para uma reposta eficiente a crises epidemiológicas através de padrões durante pandemias.

Cloud Firestore (NoSQL)

Devido às aplicações voltadas para o cliente serem baseadas na web, a escolha do Firestore como banco de dados NoSQL se tornou mais adequada. Isso se deve pelo fato de que o Firestore possui suporte para aplicações web e desse modo, poderá consultar dados, sincronizá-los, e armazená-los nos aplicativos web e móveis que o usuário cliente estará acessando nesse sistema.

- Ideal para aplicações de saúde onde os registros médicos precisam estar atualizados, a sincronização desses dados será feita em tempo real.
- O Firestore terá integrações com outros serviços que fazem parte do sistema:
  - → Integração com Cloud Storage: Recuperação de documentos e imagens diretamente do Firestore.
  - → Integração com Cloud Logging: Opção de depuração, auditoria e rastreio de atividades através da análise de registros do Firestore.
  - → Integração com Cloud Monitoring: Monitoramento de desempenho do Firestore.
  - → Integração com Cloud Pub/Sub: Alterações no Firestore ativarão eventos no Cloud Pub/Sub.

→ Integração com Cloud IAM: Controle de quem modifica, acessa e visualiza dados do Firestore.

- Ideal para armazenar dados complexos, o Firestore tem suporte para o armazenamento de registros médicos incluindo imagens e textos.
- Devido a sua estrutura flexível para dados complexos, O Firestore poderá armazenar registros médicos com informações variadas, como textos e imagens.

Recuperação de Desastres

Backups

Backups serão essenciais para segurança e recuperação de desastres tanto do ambiente on-premise quanto do ambiente GCP. Desta forma, há a proteção de dados críticos e a garantia da continuidade dos negócios.

- O backup completo, onde é feita a cópia de todos os dados, será realizado aos domingos, pelo menor número funcionários fazendo alterações no sistema.
- O backup incremental, onde será feita a cópia apenas das alterações que foram feitas desde o último backup, é feito após horário do expediente, todos os dias da semana.
- Backups mensais também serão realizados.
- Backups anuais serão guardados por até no máximo 5 anos por questões legais de compliance.
- O armazenamento dos backups será feito no Cloud Storage através do serviço Cloud Backup e DR.
- Devido à alta sensibilidade dos dados envolvidos, os testes de backup serão feitos mensalmente, para que haja certeza de que o conteúdo do backup guardado esteja funcionando corretamente.
- Serão realizados snapshots para que sejam monitoradas alterações nas aplicações em tempo real.

Compromisso com a conformidade regulatória

Devido a sensibilidade dos dados em questão e o compromisso com a Lei Geral de Proteção de Dados (LGPD), foi decidido que a região principal e de backup para recuperação de desastres seria localizada em São Paulo no Brasil. Embora o custo seja mais elevado, a latência será incrivelmente baixa e também é essencial que o sistema esteja em conformidade com a LGPD.

- O titular será informado de forma transparente sobre a coleta e uso de seus dados antes que seja obtido seu consentimento para tal.
- As políticas de privacidade serão mostradas em forma de aviso de forma clara e compreensível, descrevendo detalhadamente o propósito para a coleta de dados.
- Serão realizadas feitas Avaliações de Impacto à Proteção de Dados sempre que se provar necessário.
- O titular terá fácil acesso em questão da modificação dos seus dados, exclusão e alteração de tais informações, bem como tudo o que estiver no seu direito.

- Serão feitas atualizações em políticas de privacidade e revisões contínuas das práticas de privacidade.
- Será mantido o registro detalhado de todas as atividades relacionadas aos dados.
- Contratos com terceiros que venham a manipular os dados terão clausulas claras e específicas sobre proteção dos dados.

Infraestrutura como código (IaC)

A infraestrutura como código permitirá o gerenciamento da infraestrutura de TI fazendo uso de códigos. O uso de IaC é suportado pelos seguintes serviços adotados para a construção desse sistema:

- GKE: O GKE possui suporte a IaC de forma que é feita a descrição e configuração do Kubernetes por meio de arquivos YAML.
- Cloud IAM: Políticas de IAM para recursos GCP serão configuradas para controle de acesso e permissões de usuários internos utilizando o Deployment Manager.
- Cloud Monitoring e Cloud Logging: Será feita a configuração do Deployment Manager para que seja configurada a gestão e monitoramento de logs.
- Artifact Registry: Serão realizadas configurações de políticas de acesso aos artefatos, bem como o gerenciamento no repositório do Artifact Registry fazendo uso do Deployment Manager.
- Cloud BigQuery e Cloud Firestore: O Deployment Manager será utilizado para definir a configuração inicial do Firestore e o BigQuery para ser parte de uma abordagem IaC.
- Cloud Pub/Sub: Tópicos e assinaturas serão configurados no Cloud Pub/Sub através do uso do Deployment Manager.
- Cloud Backup e DR: Políticas e agendamentos para backup e recuperação serão configurados usando IaC de seus próprios serviços.
- Cloud Storage: Será usado o Deployment Manager para gerenciar e definir os buckets dentro do Cloud Storage, bem como definir e gerenciar políticas de acesso.

Os seis pilares do GCP

Imagem retirada do site oficial da google:
https://cloud.google.com/architecture/framework?hl=pt-br

O sistema da EHR Healthcare foi feito se baseando nos pilares GCP:

- ✓ Design do sistema: A arquitetura do sistema foi desenhada de forma clara e detalhada, destacando cada parte do sistema com legendas para um melhor entendimento.
- ✓ Excelência Operacional: Devido a muitos serviços GCP suportarem integrações com outros serviços, o sistema foi projetado para que os serviços trabalhem juntos e com eficiência máxima.
- ✓ Segurança, privacidade e conformidade: O sistema foi projetado para ter o maior nível de segurança possível tanto na parte de acesso dos usuários quanto em relação a componentes de dentro do sistema, bem como questões sobre privacidade e conformidade foram frisadas e estarão em dentro da conformidade regulatória.
- ✓ Confiabilidade: O sistema será altamente disponível e resiliente, isso se deve principalmente à parte sobre as regiões escolhidas e escolha de bancos de dados utilizados.
- ✓ Otimização de custos: Todos os serviços do sistema serão aproveitados da melhor forma possível para que a empresa EHR Healthcare seja substancialmente beneficiada com o investimento feito na GCP.
- ✓ Otimização de desempenho: Os recursos do sistema foram feitos para que sejam escaláveis e configurados para que haja equilíbrio para alcançar o desempenho ideal, sem que haja gastos extras sem necessidade.

BÔNUS

Considerações a se fazer

O FinOps é uma prática valiosa que garante que a empresa aproveite o máximo do que foi investido em nuvem, fazendo com que o valor do negócio seja impulsionado, realizando o controle de custos e promovendo a eficiência operacional. Tal prática possui três princípios: A colaboração, a responsabilidade e a otimização. Seria interessante considerar, com base no teto

de gastos da empresa EHR Healthcare (no momento, não há dados específicos como o teto de gastos da empresa em questão para tirar o embasamento), a implantação desta prática no sistema para um monitoramento eficiente dos custos/uso no ambiente de produção:

- ★ Pode-se configurar a exportação dos dados de Billing (custos dos serviços) do projeto para o BigQuery, em seguida realizar a criação de relatórios no Looker.

Criação do projeto: Helena Siqueira Araujo e Rafael Ramos Camargo.