Greenfield cloud deployment resulting in 20% cost saving

Customer: One of the top biopharmaceutical companies

Summary

The customer is an international clinical-stage biopharmaceutical company focusing on cellular immunotherapy treatments for cancer is looking at adopting cloud services for the very first time. They plan to structure their database on Google cloud platform. The intention is to enhance performance and have efficient research outputs from their applications especially since they handle large volumes of data. They were also looking at the ability to scale at any point of time during peak loads along with complete automation of continuous integration and continuous deployment (CI/CD) process for easier deployments and better auditing, monitoring and log management.

About Customer

The customer is a clinical-stage biopharmaceutical organization with the scientific vision of revolutionizing the treatment of cancer. They specialize in the research, clinical development and commercialization of cancer immunotherapy treatments. The combination of technologies from its academic, clinical and commercial research partners have enabled the company to create a fully integrated approach to the treatment of cancer with immunotherapy. They plan to work with Powerup to use Google Cloud Platform (GCP) as its cloud platform for their Cancer Research program.

Problem Statement

The customer plans to use Google Cloud Platform (GCP) as its cloud platform for their Cancer Research program. Data scientists will be using a Secure File Transfer Protocol (SFTP) server to upload data on an average of one to two times a month with an estimated data volume of 2-6 TB per month.

The data transferred to GCP has to undergo a two-step cleansing process before uploading it on a database. The first step is to do a checksum to match the data schema against the sample database. The second step is transcoding and transformation of data after which the data is stored on a raw database.

Proposed Solution

Greenfield setup on GCP

Understanding customer needs while also understanding the current python models and workflows to be created were the first steps in initiating this project. Post these preliminary studies and sign-off, a detailed plan and solution architecture document formed a part of the greenfield project deliverables.

The set up included shared services, logging UAT and production accounts. The Cloud Deployment Manager (CDM) was configured to manage their servers, networks, infrastructure and web applications. Cloud Identity and Access Management (IAM) roles were created to access different GCP services as per customer specification, which helped in securely accessing another service.

On-premise connectivity is established via VPN tunnels.

The data scientists team have built nearly 50+ python/R models that help in the data processing. All the models are stored in GitHub currently. Python model will meet performance expectations when deployed and CI/CD pipelines to be created for 48 python models.

Once the data arrives on the database, the customer wants the python code to process the data and store the results on an intermediate database.

Multiple folders were created to deploy production, UAT and management applications. Cloud NAT was set up to enable internet access, Virtual Private Cloud (VPC) peering done for inter-connectivity of required VPCs and SFTP server was deployed on Google Compute Engine.

Once data gets uploaded on the raw GCS, checksum function will be triggered to initiate data cleansing. In the first phase, the data schema will be verified against a sample database after which the data will be pushed to transcoding and transformation jobs. Processed data will be stored to GCS.

All the python/R models will be deployed as a docker image on a Kubernetes cluster that is managed by Google ensuring that GCP is taking care of high availability and scaling.

The customer will have multiple workflows created to process data that in turn would be able to define all the workflows for python model executions.

The customer team will view the current data through a web application.

The processed data also has to be synced back to the on-premise server. An opensource antivirus tool is used to scan and verify data before migrating to Google Cloud Storage (GCS).

Monitoring and Logging

Monitoring tools such as stackdriver for infrastructure and application monitoring as well as log analytics was used as it supported features like tracing, debugging and profiling to monitor the overall performance of the application.

Additional tools such as Sensu to monitor infrastructure, Cloud Audit logging that checks Application Program Interface (API) activities, VPC flow logs to capture network logs and FluxDB as well as Grafana to store data on the database and visualize and create dashboards respectively were utilised.

Stackdriver logging module ensures centralized logging and monitoring of the entire system.

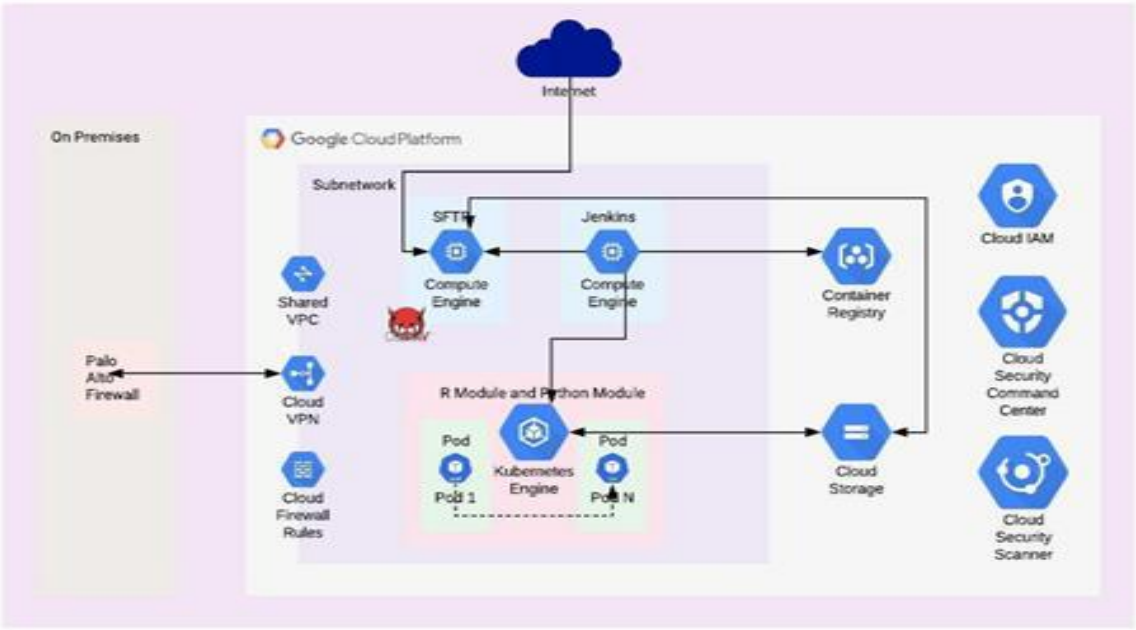
Security and Compliance

IAM with least permissible access and Multi-Factor Authentication (MFA) be enabled as an additional layer of security for account access. The databases won't have direct access to critical servers like database and app servers. Firewall rules will be configured at the virtual networking level for effective protection and traffic control regardless of the operating system used. Only the required ports will be opened to give access to the necessary IP addresses.

Both data in transit and at rest are by default encrypted in GCP along with provisions for static code analysis and container image-level scanning.

CI/CD pipeline

Setup CI/CD pipeline using Jenkins which is an open-source tool that facilitates modern DevOps environment. It bridges the gap between development and operations by automating building, testing and deployment of applications.



Benefits

After the successful deployment of code, code integration and log auditing got simpler. The customer was able to handle large blocks of data efficiently and auto-scaling at any point of time during new product launches and marketing events became effortless. This improved their performance as well.

The customer was also able to scale up without worrying about storage and compute requirements. They could move into an Opex model on the cloud by paying as per usage.

Moving to GCP enabled the customer to save 20% of their total costs as they could adopt various pricing models and intelligent data tiering.

Cloud platform

GCP.

Technologies used

Shared VPC, Cloud VPN, Compute Engine, Kubernetes Engine, Cloud Storage, Cloud Security Scanner, Cloud IAM, Cloud Security Command Center, Cloud Registry.