*-PRÁCTICA BIG DATA ARCHITECTURE VERÓNICA BERMEJO GÓMEZ-*

**Diseño del DAaaS**

***1.Definición de la estrategia del DAaaS:***

Ofrezco un dashboard (tipo donut), provisto por kibana, de visualización del top10 de los destinos turísticos más demandados según el mes del año, según el tipo de turismo (familia, pareja, amigos) y según el país. Todo ello se puede elegir mediante filtros desplegables.

***2.Arquitectura DAaaS***

Para ello :

-Scrapeo la página de Booking, consultando su API, buscando los destinos más contratados.

-Luego con esta información hago stelling, volcando los datos en un bucket de google storage en una carpeta input.

-Posteriormente hago un cluster con dataproc, convirtiendo los datos en HDFS y envio un job de hadoop que devuelve la información requerida a una carpeta output de google storage nuevamente.

-Con estos datos voy a una máquina virtual en google compute engine y lanzo elastic search con kibana donde creo mi dashboard.

***3.DAaaS Operating Model Design and***

***Rollout***

Estos procesos se ordenarán de la siguiente manera:

1-Un operador levanta el scrapper todas las noches a cierta hora.

2-Sobreescribe el resultado del scrapper en el bucket de la carpeta input de google storage.

3-Este mismo operador, posteriormente, ejecuta dataproc y lanza el job diariamente, produciendo el archivo con la información necesaria, que se almacenará en la carpeta output de google storage.

4-Por último el operador lanza un script de Python que envía los últimos resultados del algoritmo de hadoop a un índice de Elasticsearch.

5-kibana lee toda la información del índice y actualiza su dashboard.

6-Delivery. Doy al cliente una clave de acceso a kibana para poder visualizar el dashboard.

***4.Desarrollo de la plataforma DAaaS.***

-EN EL SCRAPEO DE DATOS, hacemos un script de Python para conseguir sacar de la API de Booking los destinos más contratados.

-Vamos a utilizar google cloud platform, pues tenemos crédito gratis.

En primer lugar HAREMOS STELLING, volcando nuestros datos en un bucket de cloud storage en una carpeta input.

-Desde dataproc creamos un clúster y LANZAMOS UN JOB desde la consola de dataproc, que ejecuta un algoritmo con java, convirtiendo nuestros datos en HDFS, que volverán a una carpeta output de google storage.

-El último paso de búsqueda -visualización lo haremos con ELASTICSEARCH-KIBANA: Creamos una instancia en una máquina virtual de Compute Engine (con al menos 4 GB de memoria y 2vCPUs) y entramos en ella por SSH.

Primero instalamos la herramienta WGET, para descargar las últimas versiones de elastic y kibana y posteriormente las instalamos y las configuramos.

En VPC network creo un firewall rule para que acepte conexiones de las IP que le digamos.

Y por último arrancamos tanto elastic como kibana.

Cojo la external IP de elastic en su instancia en compute engine.

Ya tengo mi server de elasticsearch.

<HTTP://IPDELAINSTANCIA:9200>

Ya puedo cargar mis datos utilizando Python y el API REST de elasticsearch.

Abro kibana con el enlace: <HTTP://IPDELAINSTANCIA:5601> y creo mi dashboard.

**DIAGRAMA**

