REPORTE DE ANÁLISIS DE COSTOS

Empresa Biotech Explorer Corp.

El objetivo de este reporte es ofrecer un análisis sobre el costo de disponibilizar una solución para el análisis de imágenes de secuecniación genética en la nuble de Amazon (AWS). Este análisis se realiza de manera específica para cada componente de la arquietctura, y al final de manera general para entender costos totales. Adicionalemente se podrán modificar los requerimientos técnicos para realizar una proyección de los costos según modificaciones de parámetros de números de instancias y capacidad de almacenamiento.

- Los componentes de la arquietctura son los siguientes:
  - Servidor EC2 de modelos (engine) para el procesamiento de secuencias de ADN
  - Servidor de bases de datos Postgresql RDS
  - Repositorio S3 de imágenes de secuenciación genética

Contexto

Biotech Explorer Corp. es una pequeña empresa (StartUp) del sector farmacuéutico que realiza análsis de imágenes de secuenciación genética en búsqueda de patrones genéticos que permitan construir vacunas contra enfermedades. Para esto desarrolla modelos de Machine Learning (computer vision) que se encargan de procesar un gran banco de imágenes de secuenciación genética para extraer datos e información que almacena en un servidor para posteriormente realizar análsisi y otros procesos para extraer patrones genpeticos valiosos para la construcción de vacunas. Estos procesos se realizan sistemáticamente y con cargas constantes durante todo el día, dado que la empresa dispone de una infraestructura dedicada solo a estos procesos y que no intervienen ni afectan los procesos de ventas, usuarios y otros procesos sustantivos como nómina, contabilidad, entre otros. La empresa tiene **ingresos mensuales** operacionales. por US$5MM, con un margen operacional del 10 500K mensuales, los que se utilizan para el pago de nóminas para los 100 empleados , gastos administrativos y operacionales, disponiendo aproximadamente el 1% de este presupuesto mensual para el pago de infraestructura tecnológica destinada a la investigación de nuevos patrones genéticos para el desarrollo de vacunas; es decir el presupuesto mensual para costos de operación tecnológica es de máximo US$ 5K.

In [ ]:
```
import pandas as pd
```

1. Análisis de costos servidor EC2

Teniendo en cuenta que se realizarán procesos sistemáticos, con cargas de trabajo constantes, y que los procesos de análisis de las imágenes de secuenciación de ADN se realizará en memoria, por la naturaleza de este proceso se disponen instancias **dedicadas** para estos modelos, con **16vCPUs** para potenciar el uso de paralelismo, y con una memoria RAM de **32 GB**. A pesar que no se contemplan picos programados, sino que la carga esperada (en un percentil 95) será constante. Adicionalmente esta instancia **no requerirá** de almacenamiento local ya que las imágenes y la información extraída se persisten en servidores (storages) dedicados para esto.

Teniendo en cuentas las características definidas y los precios estimados según la caulculadora de precios de AWS, la tabla de precios **de la propuesta** es la siguiente:

In [ ]:
```
ec2_princing= 0.408
ec2_instance = 1
ec2 = {'Recurso IT': 'Servidor de modelos Python',
                'Servicio Cloud': 'EC2',
                'Region': 'Norte de Virginia',
                'Especificación' : 'a1.4xlarge\n(32 GiB memory\t16 vCPUs)',
                'Instancias dedicadas': 'Sí',
                'Instancias' : ec2_instance,
                'Precio por hora on-demand': ec2_princing,
                'Cant. horas al mes': 730,
                'Fee mensual (USD)': 1460,
                'Costo mes (USD)':1757.84}
df_ec2 = pd.DataFrame([ec2])
df_ec2
```

Out[ ]:

	Recurso IT	Servicio Cloud	Region	Especificación	Instancias dedicadas	Instancias	Precio por hora on-demand	Cant. horas al mes
0	Servidor de modelos Python	EC2	Norte de Virginia	a1.4xlarge\n(32 GiB memory\t16 vCPUs)	Sí	1	0.408	730

1.1 Proyección de aumento de costos dado aumento de instancias

Proyectando un aumento en número de instancias hasta 10, el costo mensual estimado (proyectado) sería el siguiente.

> Nota: si se desea se puede modificar este número

In [ ]:
```
ec2_instance_forecasting = 10

def ec2_fee_mensual(df):
        horas_mes = 730
        return horas_mes*2
```

```
def ec2_month_pricing(df):
    return (df['Instancias'] * ec2_princing * 730) + (df['Fee mensual (USD)'


def row_repeat(df, repeat_num, index_column_increment_1, index_column_increm
    fila_lista = df.iloc[0].values.tolist()
    for i in range(1, repeat_num+1):
        df.loc[len(df)] = fila_lista
        fila_lista[index_column_increment_1] = fila_lista[index_column_incre
        if index_column_increment_2 is not None:
            fila_lista[index_column_increment_2] = fila_lista[index_column_i
    df = df.drop(index=0).reset_index(drop=True)
    return df
```

In [ ]:
```
df_ec2_forecasting = df_ec2.copy()
df_ec2_forecasting = row_repeat(df_ec2_forecasting, ec2_instance_forecasting
df_ec2_forecasting['Fee mensual (USD)'] = df_ec2_forecasting.apply(ec2_fee_m
df_ec2_forecasting['Costo mes (USD)'] = df_ec2_forecasting.apply(ec2_month_p

df_ec2_forecasting
```

Out[ ]:

	Recurso IT	Servicio Cloud	Region	Especificación	Instancias dedicadas	Instancias	Precio por hora on-demand	Cant. horas al mes
0	Servidor de modelos Python	EC2	Norte de Virginia	a1.4xlarge\n(32 GiB memory\t16 vCPUs)	Sí	1	0.408	730
1	Servidor de modelos Python	EC2	Norte de Virginia	a1.4xlarge\n(32 GiB memory\t16 vCPUs)	Sí	2	0.408	730
2	Servidor de modelos Python	EC2	Norte de Virginia	a1.4xlarge\n(32 GiB memory\t16 vCPUs)	Sí	3	0.408	730
3	Servidor de modelos Python	EC2	Norte de Virginia	a1.4xlarge\n(32 GiB memory\t16 vCPUs)	Sí	4	0.408	730
4	Servidor de modelos Python	EC2	Norte de Virginia	a1.4xlarge\n(32 GiB memory\t16 vCPUs)	Sí	5	0.408	730
5	Servidor de modelos Python	EC2	Norte de Virginia	a1.4xlarge\n(32 GiB memory\t16 vCPUs)	Sí	6	0.408	730
6	Servidor de modelos Python	EC2	Norte de Virginia	a1.4xlarge\n(32 GiB memory\t16 vCPUs)	Sí	7	0.408	730
7	Servidor de modelos Python	EC2	Norte de Virginia	a1.4xlarge\n(32 GiB memory\t16 vCPUs)	Sí	8	0.408	730
8	Servidor de modelos Python	EC2	Norte de Virginia	a1.4xlarge\n(32 GiB memory\t16 vCPUs)	Sí	9	0.408	730
9	Servidor de modelos Python	EC2	Norte de Virginia	a1.4xlarge\n(32 GiB memory\t16 vCPUs)	Sí	10	0.408	730

Desde la óptica de costos mensuales, los modelos ocuparían el 35% del presupuesto mensual, con **US$ 1757.84**. El análisis de la proyección indica un aumento por encima del presupuesto mensual, lo cual sería en principis inviable y habría que analizar con el

departamento financiero si son costos que se pudisen asumir. De todas maneras escalar la solución a 2 instancias llevaría los costos de EC2 a **US$ 2055.68** lo cual estaría dentro de lo presupuestado, en caso de una contigencia que implicara escalar la solución a dos instancias.

2. Análisis de costos storage S3 de impágenes de secuenciación de ADN

> Assumptions: a) se asumirán imágenes fotográficas de alta resolución de secuenciación de ADN con un peso de hasta **15 MB**, las que se producidas por equipos tecnológicos especializados capaz de procesar muestras de ADN animal y vegetal; b) los estudios genéticos se realizan sobre poblaciones de cientos de miles de individuos, nunca mayor a 5 millones; c) una vez que se ha terminado con el estudio de una población, la imágenes de secuenciación de ADN dejan de ser relevantes para la empresa y pasan a almacenarse en un storage "en frío" (backup) que está a cargo de otra área tercerizada (especializada en custodiar material e información genética); y d) las imágenes ya están almacenadas en instancias S3 y son proveedidas por una empresa aliada (esto reduce costos de carga hacia S3)

Teniendo en cuenta el peso de las imágenes de secuenciación genética, y estimando para poblaciones muy grande (en el orden de los 12 millones de individuos a estudiar), se estima una capacidad de almacenamiento máxima de aproximadamente **72TB**

Segpun el esquema de precios para una instancia **S3 estándar** se distribuyen los precios de la siguiente manera:

- los primero 50TB a US$ 0.023 x TB
- los próximos 450 TB a US$ 0.022 x TB
- los siguientes TB a US$ 0.021 x TB Y así va reduciéndose el precio

Para este caso en particular, solo interesan los dos primeros niveles de pricing:

- los primero 50TB a US$ 0.023 x TB
- los próximos 22 TB a US$ 0.022 x TB

El análisis de costos para este componente de infraestructura, que es considerado muy crítico ya que es el que almacena las imágenes a analizar, quedaría de la siguiente manera.

In [ ]:
```
s3_l1_princing= 0.023
s3_l2_princing= 0.022
storage_tb = 72
storage_gb = storage_tb  * 1024
```

```
s3 = {'Recurso IT': 'Repositorio de imágenes',
                'Servicio Cloud': 'S3',
                'Region': 'Norte de Virginia',
                'Especificación' : 'S3 Standard',
                'Instancias dedicadas': 'Sí',
                'Almacenamiento (TB)' : storage_tb,
                'Almacenamiento (GB)' : storage_gb,
                'Precio primeros 50TB': s3_l1_princing * (50 * 1024),
                'Precio resto TB': s3_l2_princing * (storage_tb-50) * 1024,
                'Costo mes (USD)': 1673.22}
df_s3 = pd.DataFrame([s3])
df_s3
```

Out[ ]:

	Recurso IT	Servicio Cloud	Region	Especificación	Instancias dedicadas	Almacenamiento (TB)	Almacen
0	Repositorio de imágenes	S3	Norte de Virginia	S3 Standard	Sí	72	

Para una capacidad de almacenamiento de 72TB, lo que supone una capacidad para almacenar unos 5 millones de impagenes de secuecniación de cadenas de ADN, el costo mensual es de aproximadamente **US$ 1673**, lo que sumado al costo de EC2 (US$ 1757) representa casi el 69% del presupuesto mensual destinado.

Realizando una proyección de costos aumentando la capacidad de almacenamiento a casi el doble de lo previsto, quedaría de la siguiente manera.

In [ ]:

```
def s3_diff_tb_pricing(df):
    diff = df['Almacenamiento (TB)'] - 50
    if diff > 0 and diff < 460:
        return diff * 1024 * s3_l2_princing
    else:
        0



def s3_month_pricing(df):
    return df['Precio primeros 50TB'] + df['Precio resto TB']


s3_tb_forecasting = 10
df_s3_forecasting = df_s3.copy()

df_s3_forecasting = row_repeat(df_s3_forecasting, s3_tb_forecasting, 5,incre
df_s3_forecasting['Precio resto TB'] = df_s3_forecasting.apply(s3_diff_tb_pr
df_s3_forecasting['Costo mes (USD)'] = df_s3_forecasting.apply(s3_month_pric


df_s3_forecasting
```

Out[ ]:

	Recurso IT	Servicio Cloud	Region	Especificación	Instancias dedicadas	Almacenamiento (TB)	Almacen
0	Repositorio de imágenes	S3	Norte de Virginia	S3 Standard	Sí	72	
1	Repositorio de imágenes	S3	Norte de Virginia	S3 Standard	Sí	77	
2	Repositorio de imágenes	S3	Norte de Virginia	S3 Standard	Sí	82	
3	Repositorio de imágenes	S3	Norte de Virginia	S3 Standard	Sí	87	
4	Repositorio de imágenes	S3	Norte de Virginia	S3 Standard	Sí	92	
5	Repositorio de imágenes	S3	Norte de Virginia	S3 Standard	Sí	97	
6	Repositorio de imágenes	S3	Norte de Virginia	S3 Standard	Sí	102	
7	Repositorio de imágenes	S3	Norte de Virginia	S3 Standard	Sí	107	
8	Repositorio de imágenes	S3	Norte de Virginia	S3 Standard	Sí	112	
9	Repositorio de imágenes	S3	Norte de Virginia	S3 Standard	Sí	117	

Probablemente no sea necesario escalar a más de 100TB la capacidad de almacenaimento. No obstante pudiera ser necesario hacerlo hasta 80 TB o 90TB y aún así mantenernos dentro del presupuesto definido. No obstante, lo más seguro es que nos mantengamos por debajo de los 72TB y de esa manera por debajo del 68% del presupuesto mensual definido.

3. Análisis de costos storage RDS

El servidor de bases de datos PostgreSQL almacenará toda la información que se pueda extraer del análisis de las imágenes. Cada imagen genera aproximadamente 900Kb de información, por lo que para una población de análisi grande (se estima grande 5MM de

imágenes) se necesitarían aproximadamente unos 5 GB de información para almacenar una población de análisis.

Es importante señalar, que a diferencia del storage de impagenes (S3), la base de datos **sí tiene que almacenar el hitórico de datos**. Por esta razón, se estima una capacidad de almacenamiento inicial para el análisis de al menos unas 200 poblaciones grandes, lo que sería aproximadamente 100GB de capacidad de almacenamiento. Teniendo en cuenta que en dicha base de datos podrían almacenarse otro tipo de información, como configuraciones, clasificadores y nomencladores, diccionarios, etc., se recomienda reservar 1TB de capacidad en la base de datos.

Dado que sobre el servidor de bases de datos no se realizarán procesos complejos de cómputo, ni de carga en memoria de muchos datos, se define un servidor de 15gb de memoria, con 4 vCPUs, con servicio habilitado de copias de seguridad de 100GB, y una capacidad de almacenamiento de 1TB en SSD (con 100GB utilizados al mes bajo demanda)

El esquema de precio para esta configuración quedaría de la siguiente manera

In [ ]:
```
rds_hrs_princing= 0.98
rds_tb_securitycopy_pricing = 0.23
rds_instancias = 1
rds_storage_tb = 1
rds_storage_gb = rds_storage_tb * 1024


rds = {'Recurso IT': 'Servidor de bases de datos PostgredSQL',
                'Servicio Cloud': 'RDS',
                'Region': 'Norte de Virginia',
                'Especificación' : 'db.m1.xlarge, 4vCPUs, Memory 15GiB',
                'Instancias dedicadas': 'Sí',
                'Instancias': rds_instancias,
                'Almacenamiento (TB)' : rds_storage_tb,
                'Almacenamiento (GB)' : rds_storage_gb,
                'Precio ejecución instancias (USD)': 715.40,
                'Precio almacenamiento (USD)': 235.52,
                'Precio copias de seguridad (USD)': 9.50,
                'Costo mes (USD)': 960.42}
df_rds = pd.DataFrame([rds])
df_rds
```

Out[ ]:

	Recurso IT	Servicio Cloud	Region	Especificación	Instancias dedicadas	Instancias	Almacenamie (
0	Servidor de bases de datos PostgredSQL	RDS	Norte de Virginia	db.m1.xlarge, 4vCPUs, Memory 15GiB	Sí	1	

Hasta este momento se tiene que:

- Costo mensual EC2: US$ 1757.84
- Costo mensual S3: US$ 1673.22
- Costo mensual RDS: US$ 960.42

Esto para un costo total mensual de **US$ 4391.48** estando por debajo aproximadamente **US$ 600** del presupuesto previsto de **US$ 5000**, lo cual se ajusta para operar durante al menos 1 año sin necesidad de escalar los costos; tiempo suficiente para vovler sobre un nuevo análisis y re-evaluar el costo de operación tecnológica y redefinir nuevos alcances y requerimientos.