Universidad Nacional Autónoma de Honduras

Facultad de Ciencias
Carrera de Física

Maestría en Física

Tópicos de Física Computacional II

Startup Company/Data Analysis Project

Ellis Moises Reyes
Miguel Angel Serrano

Tegucigalpa, Honduras 20 de Agosto, 2021

Índice

1. Introducción

En este documento se describe el proceso de análisis de datos, herramientas que pueden usarse en el desarrollo de un proyecto para mantener la reproducibilidad, un desglose de como usar los recursos para que el proyecto sea realizado en la nube y finalmente una comparación en los costos de estos servicios por parte de diferentes proveedores de PaaS.

2. Planteamiento del Problema

Actualmente se dice que los datos son el nuevo "petroleo.º .electricidad". Nos encontramos en la era de los datos, por lo tanto se estan usando en todos los campos, banca, universidades, marketing, economía, finanzas, política, ciencia, gestión de riesgos, logística y muchos más . Actualmente se espera que en los proximos años la demanda de DAAS o servicios de análisis de datos siga en aumento.

3. Justificación

Nuestra startup busca entrar en un mercado con mucha demanda, que seguira en aumento. Actualmente las empresas mas grandes en el rubro ocupan ayuda de terceros en la limpieza y anĺisis de los datos, ya que son las partes del proceso de análisis de datos mas demandantes en tiempo.

4. Objetivos

4.1. Objetivo General

Realizar la propuesta de un proyecto que consiste en una empreza que se encargara de todo el proceso de análisis de datos para cualquier usuarios con cualquier tipo de necesidades.

4.2. Objetivos específicos

- Como analistas de datos nuestro objetivo es seguir el proceso de análisis para recolectar, organizar y transformar datos. Para llegar a conclusiones basadas en datos que puedan ayudar a una organización, universidad, empresa, compañía o persona a tomar las mejores desiciones.

- Proveer a los usuarios una versión de sus datos fácil de interpretar para que una sencilla etapa de conclusiones/decisiones.
- En el futuro esperamos ofrecer servicios de ciencia de datos enfocados en la investigacion, donde esperamos usar datasets para hacer predicciones e inferencias estadísticas usando modelos de machine learning y deep learning.

5. Marco Teórico

Hay diversas formas para desarrollar un proyecto basado en análisis de datos. El procedimiento a seguir varia dependiendo de la organización. En nuestra Startup usamos el siguiente:

1. **Definir el problema**: Hacer preguntas para entender las expectativas y el problema que las partes interesadas (stakeholders) quieren resolver. Entender cual es el propósito.
2. **Recolectar y Almacenar los Datos**: Es necesario establecer como los datos serán recolectados, pero también definir como serán almacenados, si es necesario hacer backups, si los datos son confidenciales, como deben ser tratados y si es un requisito que al finalizar el proyecto estos datos deban destruirse. Cuando los datos ya están recolectados, establecer las fuentes, si fueron recolectados internamente por la organización o fueron recolectados externamente por terceros, esto es imprescindible ya a que de esta forma podemos definir la confiabilidad de los datos. (** Revisar Huawei Cloud services o cualquier empresa que ofrezca SAAS, IASS, PASS ...etc)
3. **Procesamiento**: En esta etapa, una vez se tienen los datos ya sea en formato CSV o en una base de datos es necesario revisarlos, encontrar errores, limpairlos datasets para eliminar valores atípicos que puedan sesgar el análisis. (Spreadsheets, SQL, BigQuery... etc)
4. **Análisis**: Usar herramientas para transformar y organizar la informacion, de tal manera que se pueda concluir o hacer predicciones basadas en datos. (Python, R, Jupyter).
5. **Visualización**: Presentar los datos de una manera facil de entender para las partes interesadas, usando gráficos, mapas, dashboards. (Tableau, PowerBI)
6. **Entregar** a las partes interesadas para que ellos puedan tomar las decisiones

5.1. Definir el Problema

En un proyecto real, primero se define el problema y se realizan preguntas especificas como:

- ¿Cual es el problema?
- ¿Puede ser resuelto con data? ¿Que tipo de data?
- ¿Donde esta esta data? ¿Existe, se debe de recoletar?
- ¿Esta data es publica o privada?
- ¿Quienes estan involucrados? ¿Quienes son los stakeholders?
- ¿Cuales son los limites del proyecto?

El objetivo de las preguntas es obtener información que puede ser usada para obtener conocimientos y con ellos resolver problemas. Es recomendable realizar preguntas especificas y que sus respuestas sean medibles y cuantificadas. Ya que la información obtenida sera cuantitativa o cualitativa.

5.2. Recolectando Datos

Si los datos no serán recolectados internamente, es necesario comprobar la validez y confiabilidad de ellos. Preguntarse ¿Quien? ¿Donde? ¿Cuando? fueron recolectados.

Si los datos serán recolectados internamente hay que definir un intervalo de tiempo, que tipos de datos son necesarios y la cantidad a recolectar, el grupo de estudio del cual se recolectaran los datos y demás parametros del origen de los datos que sean pertinentes.

En general, se debe asegurar que los datos presentan la menor cantidad de bias (tendencia/prejuicio) posible, que sean originales, actualizados, citados y seguros. Es muy importante tener en cuenta la ética cuando se maneja información de terceros, ya que algunos datos son propiedad del cliente, no de la empresa. Los recursos finales deben enviarse entregarse a los clientes o como es debido en la etapa final del ciclo de vida de los datos, destruirse.

5.3. Procesando Datos

Cuando se procesan datos, lo más importante es mantener la integridad de los mismos. Es necesario que los datos sean correctos, los datasets completos, consistentes y confiables. Normalmente la limpieza de los datos se da en spreadsheets o en una base de datos, hay que tener en cuenta que en la limpieza los datos serán replicados, transferidos y manipulados en el proceso. También, es necesario tomar en cuenta la población, si se usa una muestra, que la muestra tenga un nivel de confianza mayor al 95 % y un margen de error pequeño de tal modo que los datos tengan una significancia estadística alta. Una vez que se comprueba la significancia estadística, se procede a limpiar. Para ello se usan diversas herramientas en spreadsheets y funciones en SQL.

5.4. Analizando Datos

Una vez los datos están limpios y en un formato útil estos pueden ser organizados (en tablas y bases de datos que permiten manipulación, filtrado y clasificación). Datos organizados son información que puede ser analizada por medio de cálculos y modelamiento para encontrar tendencias, relaciones, patrones y correlaciones entre los datos.

Una herramienta importante en esta etapa es la adición de datos, esto permite juntar datos de múltiples fuentes para poder combinarlos en una sola colección sumarizada.

5.5. Visualizando Datos

En esta etapa los datos se muestran a las partes interesadas en graficos, dashboards o mapas. Estas visualizaciones deben ser efectivas, convincentes y faciles de entender.

Los elementos mas importantes se resumen en el Cuadro 1:

6. Metodología

Mencionaremos recursos útiles en el procesamiento de datos en herramientas gratuitas y herramientas de pago:

Item	Herramienta	Observación
Adquisición de Datos	▪ Data co-op ▪ Web scraping ▪ Compras a terceros ▪ Encuestas ▪ Public Data ▪ Open Data	Los datos pueden ser obtenidos de los mismos clientes, se pueden usar servicios para web scraping o también comprar los datos a terceros
Transformación de Datos	▪ Herramientas para agregacion de datos ▪ VLOOKUP en spreadsheets ▪ Join() en SQL	Es uno de los elementos mas importantes ya que los datos pueden provenir de muchas fuentes por lo tanto es nececsario juntar los datos para poder encontrar correlaciones
Entrega de Datos	Bases de datos	Como el cliente tiene acceso a los datos

Cuadro 1: esumen sobre elementos mas importantes de marco teórico.

6.1. Open Source o de Acceso Gratuito

Recursos muy útiles y bastante utilizados en el proceso, ya sea para tener control de versiones, administrar un repositorio, obtener datos y usar computo para el análisis de estos datos. En la Startup se busca usar la mayor cantidad de herramientas open source posible debido a la flexibilidad y alcance ya que generalmente en estos proyectos la comunidad es amplia y por lo tanto el soporte es global, tambien, buscamos mantener reproducibilidad en los proyectos y las herramientas open source permiten lograr tal objetivo.

6.1.1. Notebooks y PaaS

- Ambiente de desarrollo para ciencias de datos anaconda navigator.
- Para escribir notebooks y realizar analisis en la nube project jupyter, tambien en google colab se puede usar computo (CPUs, GPUs) para ejecutar notebooks usando jupyter.
- En una imagen de una maquina virtual se puede compactar específicamente los recursos necesarios para el análisis de datos, se puede crear la imagen solo con herramientas de acceso gratuito(GNU/LINUX, Jupyter ...etc).
- Docker permite ejecutar en contenedores jupyter con diversas librerías de tal manera que no es necesario una instalación completa de ambientes como ser anaconda.
- Binder permite convertir un repositorio de git en notebooks ejecutables ofreciendo recursos de computo al usuario.

6.1.2. Datasets

- En cuentas oficiales de gobierno se puede encontrar open data, esto significa que ya esta estructurada y bien mantenida. También, puede usarse el repositorio de opendatanetwork.
- Alrededor de la web también puede encontrarse public data, que seria toda data que se encuentra en el dominio publico, normalmente no esta estructurada. Hay diversas plataformas de datos públicos en las cuales se puede obtener data, como ser: kaggle, kdnuggets, google cloud public datasets, google datasearch.
- En papers with code puede encontrarse artículos de investigación en el campo de ciencias de datos con los datasets usados en el paper.

6.1.3. Repositorios

Git es una herramienta usada para control de versiones, actualmente es un estándar en DevOps. Algunas de las web mas usadas para mantener repositorios y proyectos de desarrollo son: GitHub, GitLab, Bitbucket.

6.2. Herramientas de Pago

6.2.1. Visualización de Datos

La forma de presentar los datos es muy importante, para ello se debe tomar en cuenta a quien va dirigido el análisis o presentación.

- El método de McCandless para visualización de datos, divide el proceso en diferentes elementos que juntos generan una buena visualización.
- También, puede usarse un repositorio de visualizaciones para investigar que tipo seria el adecuado para el dataset que se este analizando (repositorio de visualizaciones)
- Algunas de las herramientas mas usadas son Tableau y Power BI, ambas son de pago por suscripción

Herramienta	**Descripción**	**Precio/mes ($)**
Tableu	Herramienta para visualización de datos desplegada en la nube	42 por usuario
Power BI	Herramienta para visualización de datos	20 por usuario

Cuadro 2: Comparación de precios por subscripción e herramientas de visualización de datos.

6.2.2. Limpieza y Análisis

La herramienta a usar para la limpieza de los datos depende del tamaño de los datos:

- Para BigData es necesario usar bases de datos para mantener los datasets, en análisis de datos normalmente se usan bases de datos relacionales ya que los datos deben organizarse de forma estructural por lo tanto pueden usarse diversos proveedores de Bases de Datos en la nube, estos servicios son de pago.

6.2.3. Servicios en la Nube

En general estos servidores en la nube proporcionan recursos informáticos escalables y bajo demanda para aplicaciones seguras, flexibles y eficientes

- Actualmente diversas compañías ofrecen IaaS, PaaS y SaaS como servicios por demanda en la nube. Amazon Web Services, Microsoft Azure, Huawei Cloud Services y Google Cloud son algunos de los proveedores de estos servicios. En Amazon los servidores se llaman EC2 (Amazon Elastic Compute Cloud), en Huawei se llaman ECS (Elastic Cloud Servers), en Azure se llaman Azure Vms (Azure Virtual Machines). Las tres también ofrecen servicios para instancias de Bases de Datos con diversos dialectos de SQL. Google Cloud por su parte ofrece un servicio especial para análisis de datos en su consola, llamado BigQuery. BigQuery es una plataforma para análisis de datos sin servidor ya que no es necesario crear instancias individuales o máquinas virtuales para usar BigQuery. En su lugar, BigQuery asigna recursos informáticos automáticamente cuando sean necesarios, por lo tanto es un servicio mucho mas simple de utilizar y el tiempo usado en la configuración, adaptación y migración es mínimo comparado con el tiempo dedicado en crear las instancias de servidores en las otras compañías ya que se requiere mas entrenamiento para la configuración y uso correcto de las consolas en AWS, AZURE Y HUAWEI.

OS	Tipo de Servicio	Descripción	Precio/Hora ($)	Compañía
Linux	General Purpose	4CPUs, 16 GB RAM	0.1856	AWS
Linux	General Purpose	4CPUs, 16 GB RAM	0.28	HUAWEI
Linux	General Purpose	4CPUs, 16 GB RAM	0.1670	
Linux	Memory Optimized	4CPUs, 16 GB RAM	0.2660	AWS
Linux	Memory Optimized	4CPUs, 16 GB RAM	0.3900	HUAWEI
Linux	Memory Optimized	4CPUs, 16 GB RAM	0.2660	AZURE
Linux	GPU Accelerated	4CPUs, 16 GB Ram, 64vCPUs	3.68	AWS
Linux	GPU Accelerated	1 NVIDIA V100-16Q/16GB, 32vCPUs	5.89	HUAWEI
Windows	General Purpose	4CPUs, 16 GB RAM	0.8560	AWS
Windows	General Purpose	4CPUs, 16 GB RAM	0.43	HUAWEI
Windows	General Purpose	4CPUs, 16 GB RAM	0.5970	AZURE
Windows	Memory Optimized	4CPUs, 16 GB RAM	0.9520	AWS
Windows	Memory Optimized	4CPUs, 16 GB RAM	0.57	HUAWEI
Windows	Memory Optimized	4CPUs, 16 GB RAM	0.57	AZURE
Windows	GPU Accelerated	4 GPUs, 16GB RAM. 48vCPUs	6.12	AWS
Windows	GPU Accelerated	1 NVIDIA V100-16Q/16GB, 32vCPUs	7.40	HUAWEI

Cuadro 3: Tabla comparativa de costos de diversas instancias de servidores de computo en AWS, AZURE, HUAWEI

Deployment Option	Tier	Compute	Storage	Backup	Price/Month ($)	Compañía
Single Server	General Purpose	4vCPUs, 16 RAM	1024 GB	1024 GB	658.07	AWS
Single Server	General Purpose	4vCPUs, 16 RAM	1024 GB	1024 GB	426.95	HUAWEI
Single Server	General Purpose	Gen 5 4vCore	1024 GB	1024 GB	475.95	AZURE

Cuadro 4: Tabla comparativa de costos de diversas instancias de servidores de base de datos MySQL en AWS, AZURE y HUAWEI

Nota: Tomar en cuenta que otros costos deben ser agregados como ser dirección IP, ancho de banda, arquitectura de los CPUs, tipo de servidor y otros elementos en la configuración de las instancias. También, es necesario un entrenamiento para poder configurar y usar correctamente los diversos servicios ofrecidos por AWS, AZURE y HUAWEI.

Por tal razón para este proyecto se decidió que usar el servicio de BigQuery para administrar y mantener los datos estructurados en RDB seria lo mas eficiente a nivel de costo/beneficio y facilidad de uso ya que solo es necesario subir los datasets a la nube y rápidamente en la misma consola se hacen los queries necesarios sin configurar nada mas. Analizaremos los precios de BigQuery en su propia sección.

6.3. Precios de BigQuery en Google Cloud

En BigQuery cada proyecto está vinculado a una cuenta de facturación. Todos los cargos que se aplican a BigQuery por las tareas que se ejecutan en el proyecto se facturan en dicha cuenta. Se puede usar el servicio bajo demanda que en este modelo de precios, se debe pagar específicamente por la cantidad de bytes procesados por cada consulta tanto si los datos se almacenan en BigQuery como si se guardan en una fuente de datos externa, ya que se basa únicamente en el uso, el primer TB de datos de consultas procesado del mes es gratuito. También, hay planes mensuales y anuales, donde el precio es más bajo.

Los precios de BigQuery tienen dos componentes principales:

- El precio de análisis: es el coste de procesar consultas, como las consultas SQL, las funciones definidas por el usuario, las secuencias de comandos y determinadas instrucciones de manipulación de datasets.

Opción	**Operación**	**Detalles**	**Price ($)**
Bajo demanda	Consultas, Queries	El primer TB del mes es gratis	5 por TB
Tarifa fija a corto plazo	Consultas, Queries	Se pueden cancelar las ranuras flexibles y solo se paga por los segundos que haya estado desplegado el compromiso	4/hora por 100 ranuras

Cuadro 5: Costo de la realización de consultas en BigQuery

- El precio de almacenamiento: es el coste de almacenar los datos en BigQuery. Se paga por el almacenamiento activo y a largo plazo.
    - El almacenamiento activo incluye todas las particiones de tablas y tablas que se hayan modificado en los últimos 90 días.
    - El almacenamiento a largo plazo incluye cualquier tabla o partición de tabla que no se haya modificado durante 90 días consecutivos. El precio de almacenamiento de esa tabla se reduce automáticamente en alrededor de un 50 %. No hay ninguna diferencia en el rendimiento, la durabilidad o la disponibilidad entre el almacenamiento activo y a largo plazo.

Operación	Detalles	Price ($)
Almacenamiento activo	Los primeros 10 GB del mes son gratuitos	0.020 por GB
Almacenamiento a largo plazo	Los primeros 10 GB del mes son gratuitos	0.010 por GB

Cuadro 6: Costo de almacenamiento BigQuery, el almacenamiento de 1TB durante un mes seria equivalente a 20.48 $

6.3.1. VPN en Google Cloud

Por seguridad podria usarse un tunel VPN con IPsec para realizar las consultas y envio de los datos, el costo seria de 184$ por mes tomando en cuenta el pago por la puerta de enlace y 1TB de trafico de salida que pasara por el tunel.

7. Asignación de Recursos

El siguiente análisis es basado en la siguiente idea: Si la startup tiene un presupuesto de 100 tokens, como se usarían eficientemente esos tokens. En este proyecto se busca usar la mayor cantidad de herramientas open source, por lo tanto aquí solo se desglosaran herramientas de pago. También, se supone que los integrantes del proyecto ya cuentan con una computadora la cual servirá de puerta de enlace a todas estas herramientas.

Operación	Detalles	Price ($)
Almacenamiento activo	Los primeros 10 GB del mes son gratuitos	0.020 por GB
Almacenamiento a largo plazo	Los primeros 10 GB del mes son gratuitos	0.010 por GB

Cuadro 7: Asignación de Recursos, los valores de porcentaje equivalen a cantidad de tokens y los valores mensual y anual son aproximados en dolares.

- En servicios en la nube se agregan los costos por almacenamiento y consultas en BigQuery, estamos suponiendo que por cada TB de almacenamiento se usa el triple en consultas y el primer TB en consultas es gratuito (en BigQuery). También, se agrega el costo por mantener una VPN donde los datos y consultas se movilizaran encriptados en la red, esto para mejorar la seguridad de los datos. Usando los precios investigados se llego a esos valores. Usar servicios en la nube bajo demanda nos permite usar ciertos recursos solo cuando sea necesario para así ahorrar costos.

- En Web se agregan los costos que conllevan la administración de una pagina, como ser el hosting, el nombre de dominio, estadísticas...etc. Estamos suponiendo una pagina pequeña para publicidad y contacto pero que este certificada con SSL.
- En software se agrega el costo por membresía en el uso de Power BI, estamos suponiendo en la startup habrá un encargado en visualización de datos que usara tal membresía.
- En respaldo se agrega un porcentaje en caso en un determinado mes, se ocupen realizar mas consultas de las definidas en el recurso dado a cloud services, se ocupe alguna membresía en Power BI o Tableau, se ocupe contratar por algunas horas computo en AWS, AZURE O HUAWEI. En general, para tener una holgura en los recursos usados en la nube.
- En gastos adicionales se agrega on porcentaje de recursos por cualquier situación adicional, en la cual los recursos establecidos no son suficientes incluso usando los de respaldo. Podría ser una situación de mantenimiento de hardware o compra de datos a terceros.

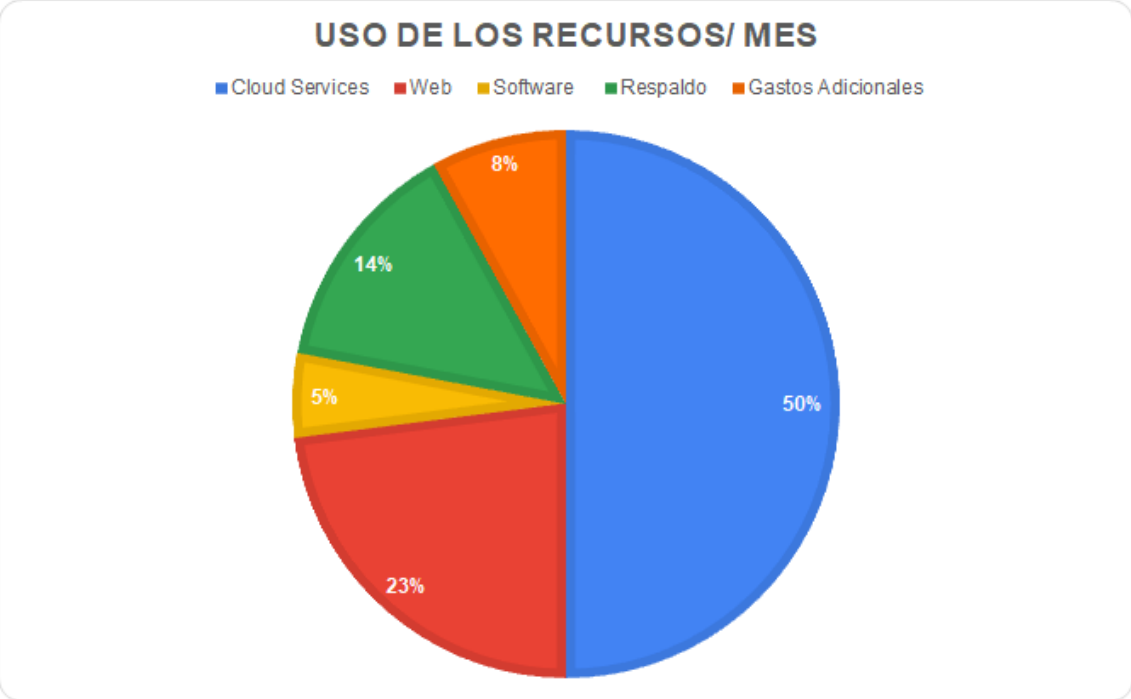

Figura 1: Pie Chart sobre la Asignación de Recursos

Dentro de los recursos que se asignan a los cloud services estos se distribuyen de la siguiente forma:

Recurso	Porcentaje
Almacenamiento en BigQuery	10
Consultas en BigQuery	5
Seguridad	85

Cuadro 8: Desglose sobre Cloud Services

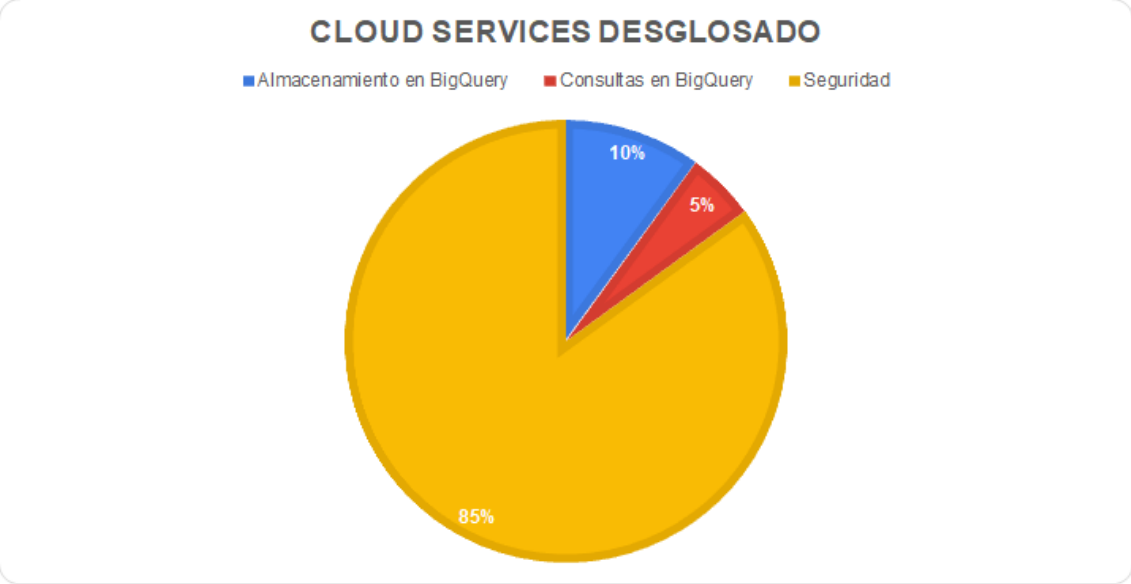

Figura 2: Pie Chart sobre el Desglose de Cloud Services

8. Anexos

Paginas web sobre las diferentes herramientas descritas en el perfil de proyecto, como ser: paginas de documentación, informacion general y calculadoras para costos de servicios en la nube:

- https://grow.google/dataanalytics/#?modal_active=none
- https://cloud.google.com/
- https://www.webfx.com/website-maintenance-pricing.html
- https://cloud.google.com/bigquery/docs/slots

- https://cloud.google.com/bigquery/docs/reservations-intro
- https://cloud.google.com/network-connectivity/docs/vpn/pricing?hl=es-419
- https://cloud.google.com/bigquery/pricing
- https://cloud.google.com/skus/?filter=bigquery&currency=USD
- https://www.huaweicloud.com/intl/en-us/pricing/#/cbr
- https://azure.microsoft.com/en-us/pricing/calculator/
- https://cloud.google.com/sql/pricing#mysql-pg-pricing
- https://azure.microsoft.com/en-us/pricing/details/azure-sql-database/single/
- https://aws.amazon.com/s3/pricing/
- https://calculator.aws/#/createCalculator/RDSMySQL
- https://docs.docker.com/engine/reference/commandline/docker/
- https://hub.docker.com/r/jupyter/base-notebook
- https://www.avenga.com/our-expertise/data-services/
- https://www.talend.com/resources/what-is-data-as-a-service/
- https://help.tableau.com/current/pro/desktop/en-us/gettingstarted_overview.htm
- https://docs.microsoft.com/en-us/power-bi/
- https://docs.docker.com/
- https://jupyter-notebook.readthedocs.io/en/stable/
- https://support.huawei.com/enterprise/es/category/cloud-computing-pid-7919788
- https://docs.aws.amazon.com/
- https://docs.microsoft.com/en-us/azure/?product=featured