Propuesta de Proyecto de Big Data

Empresa XYZ

Integrantes:

Rodrigo Andrés Espina León

Antonio Alejandro Jara Rozas

Bernardo Adolfo Riffo Carrillo

Néstor Patricio Rojas Ríos

Juan Pablo Ureta Fernández

Índice

Contenido

Objetivos

El objetivo de este informe es proponer un diseño de arquitectura que pueda obtener los datos generados por los sitios web (compras, clics, etc.), información de redes sociales, entre otros, y disponibilizarlos para nuestros científicos de datos y analista de datos, quienes generarán un modelo de recomendación de productos, que será desplegado por el usuario cuando esté navegando por alguno de los sitios en "tiempo real".

Para ello, en el presente informe se entregan las bases, tiempos y alcances para el desarrollo de este sistema, separando en tres puntos que requiere este proyecto: Proyecto, Diseño y Recursos y planificación.

La arquitectura expuesta tendrá un diseño escalable, permitiendo cualquier modificación, sin alterar su estructura. Además, contará con seguridad en el acceso de los datos y estará disponible en tiempo real.

El sistema cuenta con al menos tres roles de usuario, el primero el administrador del sistema, segundo el científico de datos y finalmente el gerente de la empresa.

El proyecto

Se definió dar énfasis a la base del trabajo, dar un orden a la ejecución de las tareas con el fin de generar las prioridades. En este sentido se tiene en cuenta de que existe una gran cantidad de datos históricos los cuáles son muy convenientes para segmentar a los clientes y generar mejores campañas de marketing, que es el objetivo final de este proyecto. En un inicio se llegó a pensar que se podría hacer una arquitectura híbrida en donde servicios de la nube podían conectarse a datos históricos on-premise. Sin embargo, considerando que era mucha información, se decidió que sería mejor subirla a la nube y analizarla allí, considerando además que la puesta en marcha de este proyecto en algún momento va a generar datos históricos. Se determinó, finalmente, que el flujo de trabajo tendrá el siguiente orden:

1. Generar arquitectura en la nube
2. Generar arquitectura de respaldo on Premise
3. Modelar los repositorios de información
4. Carga inicial de datos históricos a la nube
5. Programación de funcionalidades
6. Diseño de los reportes de usuario
7. Marcha blanca
8. Puesta en producción
9. Implementación de políticas de purga y políticas de respaldos de la data fría

Diseño

Esta sección abordará el diseño de las arquitecturas y los componentes que las integran, además del criterio asumido para cada selección.

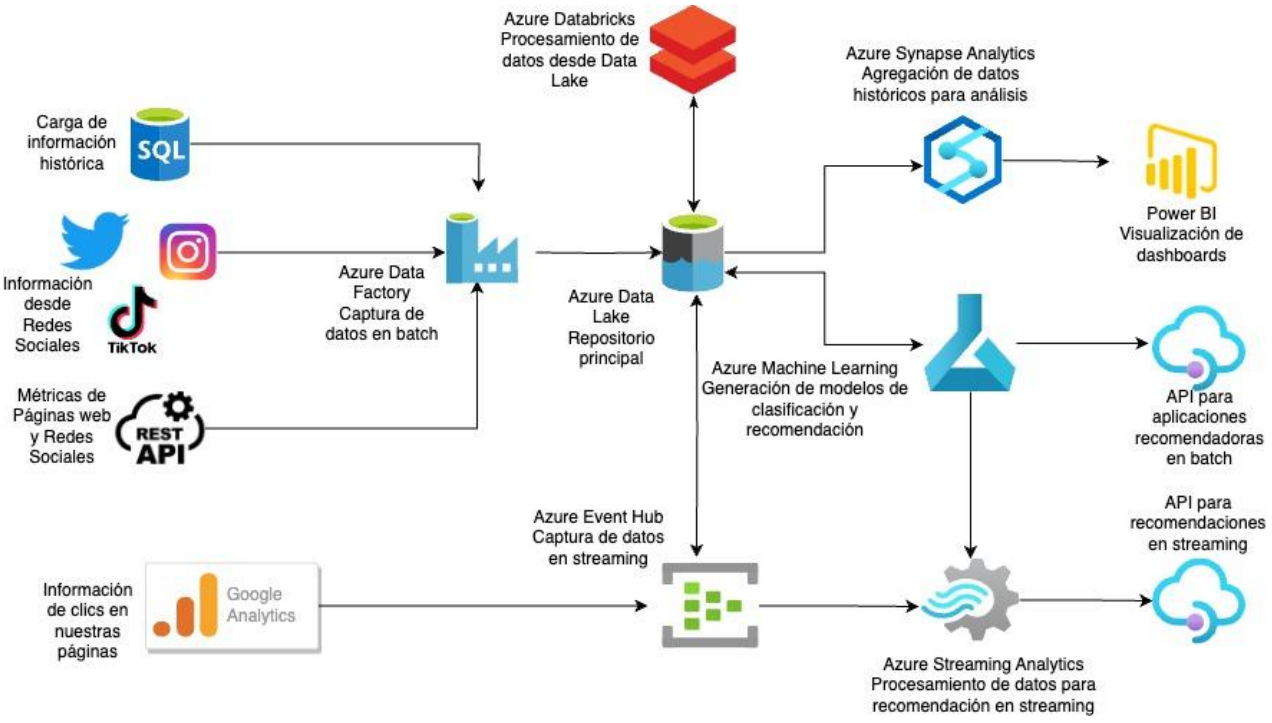
Dado que la mayoría de las fuentes de información en este proyecto provienen de internet se pensó que era conveniente realizar, por un lado, una capa en tiempo real (streaming) con la recolección de los clics y su recomendación en las distintas páginas de compra, junto a otra capa de carga batch con el grueso de la información desde redes sociales que permita el análisis de los datos y el entrenamiento de modelos de machine learning, proceso que requiere de más tiempo para obtener frutos; esto es lo que se conoce como "arquitectura Lambda". Para este proyecto se definió "tiempo real" como 3 segundos de respuesta.

En cuanto a las necesidades de almacenamiento, se realizó el cálculo para mantener el sistema funcionando por 3 años. Éste se basó en la cantidad de información histórica disponible más lo proyectado de acuerdo a la tasa de crecimiento diario reportada. Para un manejo holgado de la arquitectura se requerirá de 80 terabytes de almacenamiento, desglosados en los siguientes orígenes:

- Información proveniente de los clics (Google Analytics): 4,7 terabytes.
- Información proveniente de las compras: 12,8 terabytes.
- Información proveniente de redes sociales: 47 terabytes.

Teniendo en cuenta esto, a continuación se muestra el diagrama de la arquitectura en la nube y sus componentes; hay que tener en cuenta que las orientaciones de las flechas que se muestra no es al azar, sino que tiene que ver con el recorrido que hacen los datos entre los componentes:

Arquitectura en la Nube Azure



- ***APIs de Redes Sociales:*** Actualmente, algunas redes sociales ofrecen diversos planes para obtener información relativa a tópicos específicos, perfiles e información relativa sobre comentarios de los usuarios. Con el fin de obtener información estructurada y semiestructurada sobre el estado de la marca y tendencias de redes sociales, se contratarán planes en las siguientes APIs:
    - API Graph de Instagram: Ofrecida por Meta, busca entregar a las empresas la posibilidad de administrar su presencia en esta red social, se utilizará principalmente para identificar contenido multimedia en el que los usuarios hayan mencionado a la marca y obtener metadatos y métricas sobre hashtags y perfiles de Instagram.
    - API de Twitter: Se usará principalmente para acceder a información disponible de manera pública en esta red social, entre las cuales se destacan publicaciones, comentarios sobre la marca y/o temas a considerar para el sistema de recomendaciones y menciones por parte de los usuarios, además de publicaciones de otros perfiles de interés, como empresas competidoras.
- ***Data Factory***: Este es el servicio ELT de Azure y será el orquestador del flujo de los datos batch, haciendo los movimientos y procesamiento necesarios para que la información esté donde se requiere y de la manera óptima. Algunas de las tareas de la pieza son:
    - Permitir la captura de los datos provenientes de las redes sociales mediante procesos batch, pudiendo, además, hacer una primera depuración de la información antes de guardarla en el repositorio.
    - Realizar la carga inicial del sistema desde la información que está disponible en la arquitectura on-premise.
- ***Azure Data Lake***: Este es el repositorio del proyecto, especialmente diseñado para implementaciones de big data. Será usado para:
    - Almacenar el resultado de los datos procesados desde la recolección realizada por la capa batch con Data Factory.
    - Almacenar el resultado de los datos procesados desde la recolección realizada por la capa de streaming desde Event Hub.
    - Almacenar la información consultada por los sitios Web con las recomendaciones de productos, esto puede incluir las fotos y/o videos de los productos.
    - Alimentar el análisis de la información realizado por Azure Databricks.
    - Guardar los resultados generados por el análisis realizado por Azure Databricks.
    - Contener la “data fría”, aquella que no se consulta frecuentemente.

  Dentro del repositorio cada tipo de información deberá estar estructurada en bloques separados (dígase un sistema de carpetas) según su ciclo de vida y la naturaleza de su acceso. También se propone la encriptación de los datos para generar mayor seguridad a la información de los

clientes. Se considera que con lo anterior se puede manejar de manera adecuada la gobernanza de la información.

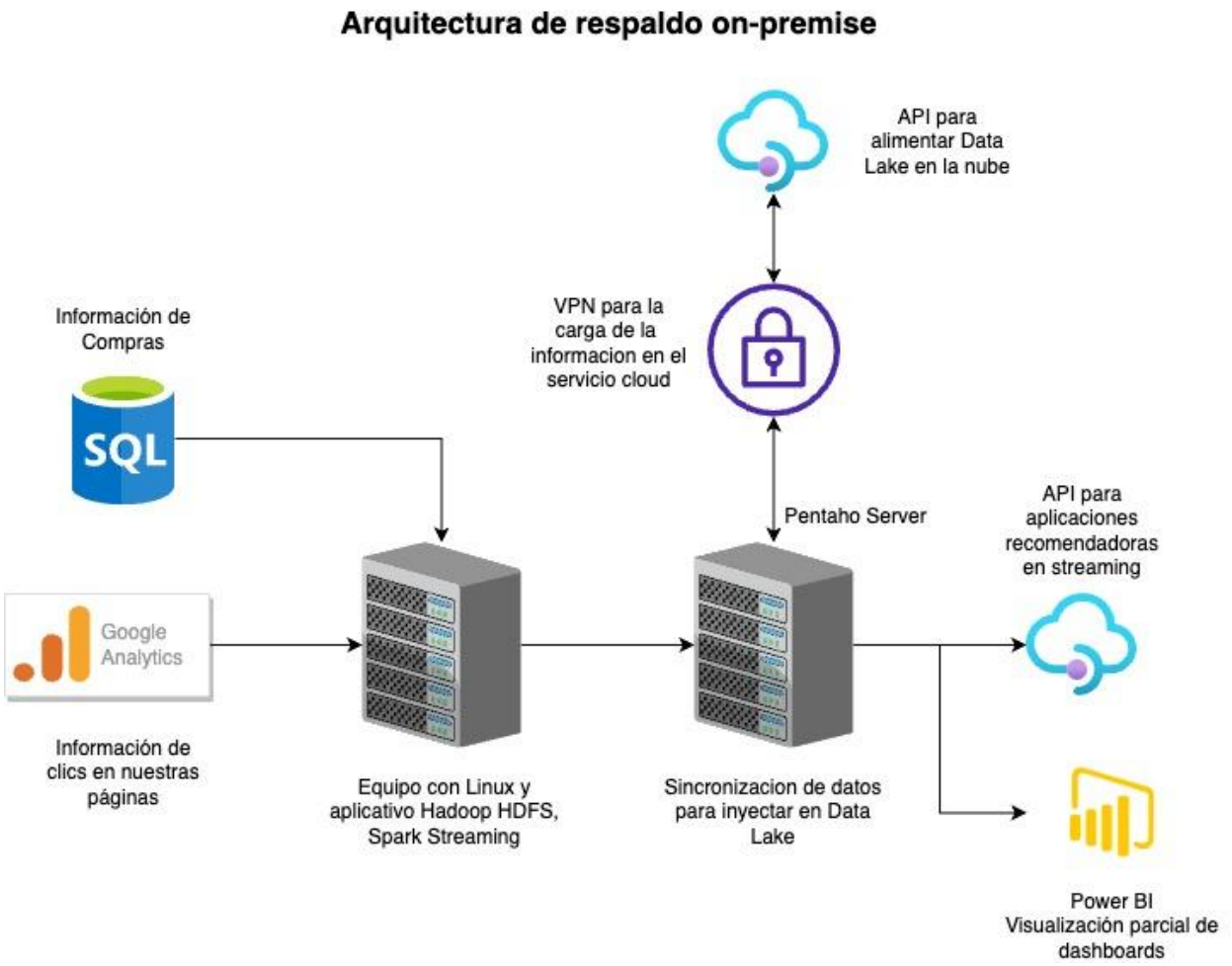
- ***Azure Databricks***: Esta herramienta permitirá ir a los datos obtenidos por las vías anteriormente descritas y ordenarlos, y generar tablas y consolidados de información para empezar a armar los primeros indicadores que requieren los científicos de datos y el analista de datos para el descubrimiento de patrones. En esta caja se incorpora información de otros servicios como el análisis de sentimientos.
- ***Azure ML***: Alimentado con los datos provenientes de Azure Data Lake, permitirá el trabajo de los científicos de datos para que éstos puedan categorizar a todos los clientes del sistema con el fin de poder personalizar las campañas de marketing que se tiene como objetivo final. El modelo está pensado para ser integrado a aplicaciones web que generen recomendaciones y otras que se puedan generar en el futuro mediante un sistema de APIs RESTful. A su vez, la última versión del modelo será almacenada en el Data Lake y, por medio de este, compartida con la arquitectura on-premise.
- ***Azure Synapse Analytics***: Alimentado con los datos provenientes desde Azure Data Lake, tendrá información agregada histórica y regularmente actualizada con KPI (key performance indicators) de análisis, exponiendo información valiosa con un lenguaje conocido por el usuario de negocio. En primera instancia la actualización será diaria, pudiendo variar según las necesidades de los usuarios.
- ***Microsoft Power BI***: Esta herramienta permite la visualización de la información, siendo una opción ideal para el autoservicio de reportes y la exploración de patrones. Su enfoque está ligado a usuarios finales, los cuales no intervendrán en el modelo, como es el caso de Gerencia y otras áreas de la empresa.
- ***Azure Event Hub***: En la capa en streaming, es el orquestador de los datos provenientes de los clics que sean realizados en las aplicaciones web. Las funciones que tendrá son:
    - Depurar y organizar los datos para que puedan ser analizados por Azure Streaming Analytics.
    - Compartir la información de clics con el repositorio principal para que puedan alimentar el entrenamiento de modelos de machine learning.
- ***Azure Streaming Analytics***: Esta herramienta permitirá procesar los datos provenientes en tiempo real desde los clics y, utilizando la información anteriormente generada con los modelos de machine learning, brindar las recomendaciones de compra en tiempo real mientras el usuario utiliza las aplicaciones web.

Arquitectura on-premise

Este modelo ilustra la arquitectura de respaldo on-premise, en caso de que el servicio en la nube presente alguna indisponibilidad. Dado que se cuenta con equipos físicos sin utilizar y con una capacidad

limitada, se decidió que estos serán útiles para recibir la información de las compras de los clientes, más el comportamiento de los clics que se realicen en los sitios webs durante el período en que el servicio de Azure cloud presente problemas.

A continuación, el diagrama de la arquitectura de respaldo y sus componentes:

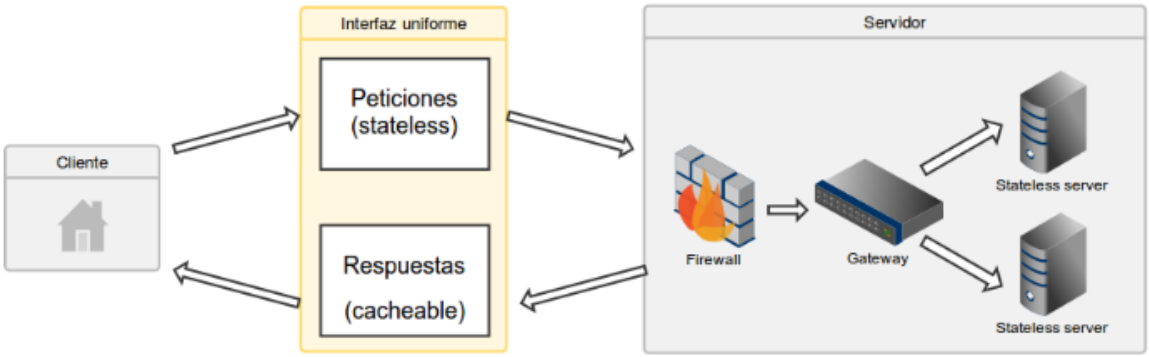


- Se utilizarán los equipos disponibles como servidores con sistema operativo Linux con dos contenedores donde en uno corre un aplicativo en Java, Python o Node.js con el código para la captura de los datos fuentes, datos colocados como mensajes en una cola Apache Kafka, la que cumple la función de transportar los datos al servidor que los procesa.
- El segundo contenedor corresponde a equipo con Apache Spark en una infraestructura Hadoop para generar los KPI básicos tales como el análisis de sentimiento para los clics que se realicen en nuestras páginas o total de compras por cliente. Los resultados quedarán en los nodos de almacenamiento HDFS ordenados en un modelo donde los datos procesados y los datos originales recolectados sin procesar se encuentren separados. Igualmente y de manera separada, se alojarán la última versión del modelo recomendador de forma tal que trabaje realizando recomendaciones en tiempo real.

- Se eligen servicios en contenedores, ya que es una unidad escalable y usa sus propios recursos logrando un aislamiento entre componentes.
- Desde el segundo contenedor se realizará la integración a aplicaciones web que generen recomendaciones en tiempo real mediante un sistema de APIs RESTful.
- Pentaho Data Integration orquesta todos los movimientos de información requeridos desde el Hadoop hacia la instancia de API que responde con recomendaciones para los clientes según su comportamiento, cuando sea requerido, además que puede generar indicadores post-proceso útiles para el cientista de datos o los usuarios finales.
- Tanto el usuario final como el Cientista de Datos podrá disponer de reportes parciales mediante Power BI desktop. Esto pues al actuar como contingencia, se podrá mostrar los indicadores en base a las compras y/o clics de los clientes.
- En el gráfico de arquitectura se puede observar un componente muy importante llamado VPN Gateway. Este es el nodo que comunica la infraestructura de la empresa con el desarrollo en la nube por donde pasarán los flujos de datos en ambos sentidos haciendo parecer que todo se encuentra en la misma RED, por lo mismo deberá tener configurada un buen control seguridad y accesos estrictamente necesarios.

Exposición de la información

La exposición de los datos para su consumo por otros sistemas (los datos de sugerencias de compras para los portales de la empresa) será a través de APIs RESTful, tanto para la infraestructura en la nube mediante servicios Lambda (Azure Functions) como para la infraestructura de respaldo usando lenguaje Node.JS en uno o varios contenedores. Estas interfaces son estándares fáciles de consumir, que abstraen la complejidad de cómo está almacenada la información. Se le deberá adicionar alguna capa de seguridad para un acceso controlado. Existen muchos servicios de software libre disponibles para este fin; también se sugiere colocar un certificado de comunicación https de los que seguramente ya tiene la empresa para sus propios portales.



Recursos y Planificación

Para cumplir con nuestro objetivo debemos contar con diversidad de profesionales, los cuales ayudarán en la implementación y puesta en marcha del proyecto. Se trabajará en dólares, con una tasa de cambio de 850 pesos chilenos (CLP) por dólar (USD).

- ***Jefe de Proyecto***: Profesional con conocimientos y nociones de desarrollo de proyectos de ciencias de datos, su labor principal será coordinar y elaborar un plan de trabajo que permita el desarrollo e implementación del proyecto. Además de ser un puente entre los distintos agentes involucrados, que reporte a la empresa sobre el estado del proyecto (USD $2000 mensuales).
- ***Ingeniero de Operaciones de infraestructura on-premise***: Encargado de gestionar y mantener la infraestructura física utilizada en el proyecto (USD $2000 mensuales).
- ***Ingeniero de Operaciones de infraestructura cloud***: Encargado de gestionar y mantener la infraestructura en la nube de Azure (USD $2300 mensuales).
- ***Ingeniero de Datos Senior***: Experto en el diseño y desarrollo de soluciones de procesamiento y almacenamiento de datos a gran escala (USD $3000 mensuales).
- ***Desarrolladores de Software***: Se requerirán con el fin de permitir la integración entre los sitios de ventas y los resultados del sistema recomendador implementado. Esta integración se hará mediante servicios y conexión con la API expuesta por el sistema, desplegando la información en las vistas e incorporándose a la experiencia del usuario en las plataformas (2 personas, USD $1600 mensuales por cada uno).
- ***Científicos de Datos***: Encargados de aplicar técnicas y algoritmos avanzados para analizar los datos y extraer información. Deberán entrenar el modelo recomendador de machine learning (2 personas, aportados por la empresa por lo que no entran en el presupuesto).
- ***Analista de Datos***: Responsable de recopilar, limpiar y analizar los datos para obtener información relevante (USD $2100 mensuales).
- ***Analista de negocio***: Encargado de comprender las necesidades y los objetivos comerciales, y traducirlos en requisitos y soluciones de análisis de datos (aportado por la empresa por lo que no entra en el presupuesto).

Plazos

Se estima que es posible realizar la migración completa en un plazo de 12 meses, desglosada de la siguiente forma:

Procesamiento de datos

La primera parte, el procesamiento de datos, se descompone en los siguientes plazos:

- Recolección de información: 2 meses.
- Diseño de algoritmos: 3 meses.
- Entrenamiento del modelo de machine learning: 2 meses.
- Diseño de API Rest de consumo de datos: 1 mes.

Generación de reportes para el usuario

La visualización interactiva, la cual estará disponible a través de una plataforma Web, será desarrollada en un plazo de 3 meses, descompuesto en lo siguiente:

- Integración con sitios de ventas: 2 meses.
- Elaboración de reportes y dashboards en PowerBI: 1 mes.
- Etapa de pruebas: 1 mes.

Entregables del Proyecto

Los entregables definidos en función de los objetivos del proyecto son los siguientes:

Servicio

- Servicio para análisis de comportamiento de clientes y generación de propuestas de productos para obtener compras.
- Reportes de KPI de funcionamiento de la plataforma.
- Reportes de KPI de resultados de la campaña.
- Reportes de KPI del comportamiento de consumo de los clientes.
- Interfaces API REST para las respuestas de las propuestas de compras a los clientes.

Modelo

- Modelo de machine learning clasificador de clientes entregado, que permita generar recomendaciones según comportamiento de los clientes y tendencias, con una tasa de aceptación de 90%.

  Nota: en caso de visualizar la necesidad de más de un modelo, se entregarán todos los que el sistema requiera.

Propuesta Económica

La oferta económica por este proyecto, considerando los recursos requeridos para su ejecución, se presenta en la siguiente tabla.

Item	Valor USD
Desarrollo de Plataforma (8 meses)	**$178.584**
Plataforma Interactiva (4 meses)	**$110.492**
Total del proyecto (12 meses)	**$289.076**

Nota 1: Los valores son presentados en USD y están exentos de IVA por la naturaleza del servicio.

Detalle de aproximación de costos para servicios en la nube

Microsoft Azure Estimate			
Service type	**Region**	**Description**	**Estimated monthly cost**
Azure Data Factory	East US	Azure Data Factory V2 Tipo, SQL Server Integration Services tipo de servicio, Estándar Nivel, 4 D4V3 Máquinas virtuales 730 Horas	$2.829,48
Storage Accounts	East US	Redundancia Data Lake Storage Gen2, Estándar y LRS, Acceso frecuente Nivel de acceso, Estructura de archivos Espacio de nombres jerárquico, Capacidad: 80 TB - Pago por uso, operaciones de escritura: 4 MB x 10 operaciones, operaciones de lectura: 4 MB x 10 operaciones, 10 operaciones de lectura iterativas, 10 operaciones de escritura iterativas, 10 operaciones de otro tipo. 1000 GB Recuperación de datos, 1000 GB Escritura de datos, 1000 GB de almacenamiento de metadatos	$1.711,11
Azure Databricks	East US	Carga de trabajo Proceso multiuso, nivel Premium, 6 D16ADSV5 (16 vCPU, 64 GB DE RAM) x 250 Horas, Pago por uso, 4 DBU x 250 Horas	$4.536,00

Azure Synapse Analytics	East US	Nivel: Optimizado para Compute Gen2, grupos de SQL dedicados: DWU 500 x 120 Horas, 1 TB de almacenamiento con recuperación ante desastres con redundancia geográfica; Región Este de EE. UU., 100 GB de datos recopilados al día, 7 días de caché de acceso frecuente, 30 días de retención total, Compresión de datos estimada en 7 veces, 730 Horas de 2 x Extra pequeña (2 núcleos virtuales) instancias de motor, 730 Horas de 2 x 1 núcleo virtual instancias de administración de datos	$1.813,44
Azure Machine Learning	East US	4 D3 (4 núcleo(s), 14 GB de RAM) x 150 Horas, Pago por uso	$184,80
Event Hubs	East US	Nivel Basic: 2 unidades de procesamiento x 730 Horas, 1000 millones de eventos de entrada	$49,90
Azure Stream Analytics	East US	Tipo de Estándar, 4 unidades de streaming x 730 Horas; Stream Analytics en 1 dispositivos con IoT Edge	$322,20
Azure Functions	East US	Nivel Consumo, Pago por uso, 128 MB de memoria, 100 milisegundos de tiempo de ejecución y 10.000 ejecuciones/mes	$0,00
Power BI Embedded	East US	4 nodos x 125 Horas, tipo de nodo: A1, 1 nodos virtuales, 3 GB de RAM, 1-300 pico de representaciones por hora	$504,05
VPN Gateway	East US	Puertas de enlace de VPN, nivel básico de VPN, horas de puerta de enlace 0 10 túneles S2S, 128 conexiones P2S, 2TB, Transferencias de datos entre redes virtuales tipo de puerta de enlace de VPN de salida	$71,68
	Support	0	$0,00
	Licensing Program	**Microsoft Customer Agreement (MCA)**	
	Total		**$12.022,66**

	Consumo de APIs de redes sociales por mes	**$1.000**

Detalle aproximado de los sueldos de los colaboradores

Cargos	Tiempo de contrato (meses)	Costo mensual (USD)	Total (USD)
Jefe de proyecto	12	$2.000	$24.000
Ingeniero de Operaciones de infraestructura on-premise	12	$2.000	$24.000
Ingeniero de Operaciones de infraestructura cloud	12	$2.300	$27.600
Ingeniero de Datos senior	12	$3.000	$36.000
Desarrolladores de softwares (el costo es por los 2 cargos)	4	$3.200	$12.800
Analista de datos	4	$2.100	$8.400
Total			**$132.800**