API N°1

Materia: Base de datos y Big Data

Fecha de entrega: 12/4/2024

Alumno: Carro, Marco Alexis

Consignas:

1. Antes de elegir una de las 3 nubes, es relevante realizar un benchmark de ellas. No hay que olvidar que el análisis debe ser realizado sobre los servicios de datos.

2. Es importante, como ingeniero de datos senior, cuestionarse lo siguiente: • ¿Cuáles son los distintos servicios que cada nube ofrece? ¿Son similares? • ¿Cuál es el valor de cada uno de estos servicios? ¿Cómo se pagan? • ¿Cuáles son las principales características de los servicios utilizados por los pipelines de datos de cada nube?

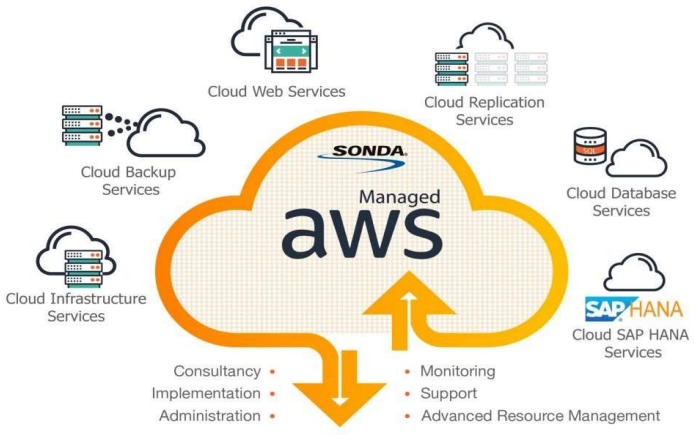
3. Una vez que se tiene el suficiente contexto para decidir, se debe seleccionar una de las nubes y fundamentar esta elección.

4. Luego de la selección de la nube, se deberá ejemplificar qué servicios se utilizan en la nube para realizar el data lifecycle management. Y, finalmente, se solicita realizar un diagrama que explique el pipeline ML para los modelos de analítica avanzada.

Respuesta.

Este informe consta de los servicios que ofrece cada nube y cuales son sus principales diferencias

Amazon Web Services.



Amazon Web Services (AWS) es un proveedor de servicios en nube donde nos dispone de almacenar grandes volúmenes de datos, recursos de aplicación, bases de datos, etc. AWS es un servicio de tipo suscripción mensual, esto, como en cualquier servicio en nube, nos ahorrara el costo de tener equipos físicos y ocupando espacio y

almacenamiento. Las ventajas que ofrece AWS son SEGURIDAD con certificaciones como: **PCI DSS nivel 1, FISMA Moderate, HIPAA Y SOC 1, ISO 27001 y auditoria SOC 2**, que hacen que sea 100% confiable, BASES DE DATOS: AWS permite acceder a bases de datos como **MySQL, ORACLE, Aurora, PostgreSQL, SQL Server, MongoDB**, etc. Dentro de los servicios que AWS destacan son: Almacenamiento. Redes. Bases de datos. Aplicaciones. Mensajería. Inteligencia artificial. Servicios móviles. Seguridad informática. Identidad.

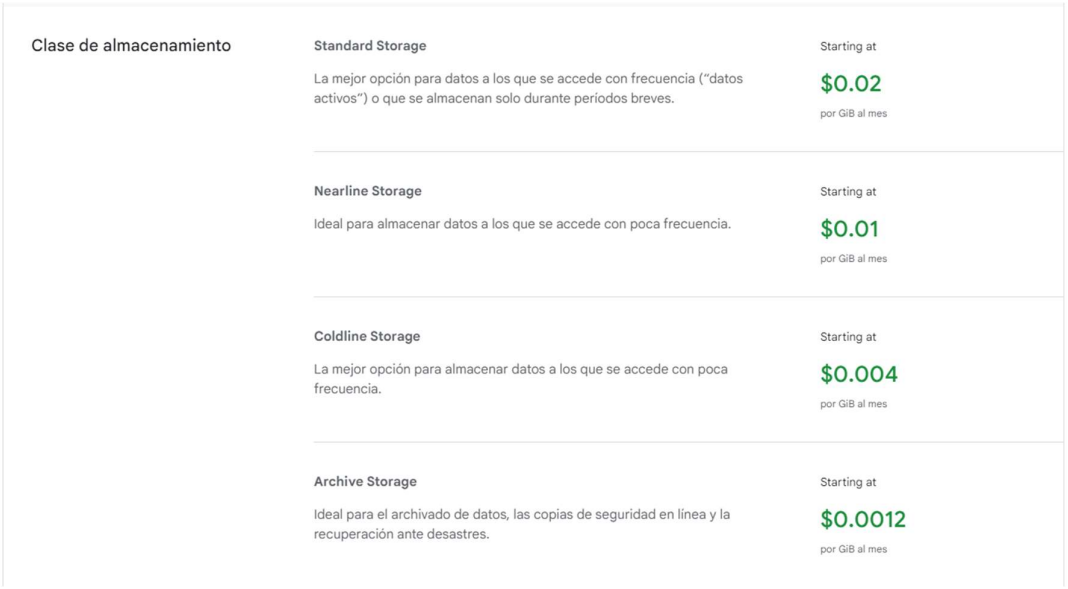
Algunos beneficios al utilizar AWS es la integración con un módulo de programación, base de datos fáciles de usar. Es rentable. Ofrece facturación y gestión centralizada, computación hibrida e instalación y desinstalación rápida. AWS ofrece gran variedad de tipos de almacenamiento como Amazon S3, EBS, Storage Gateway, EFS, etc.

Google Cloud Storage

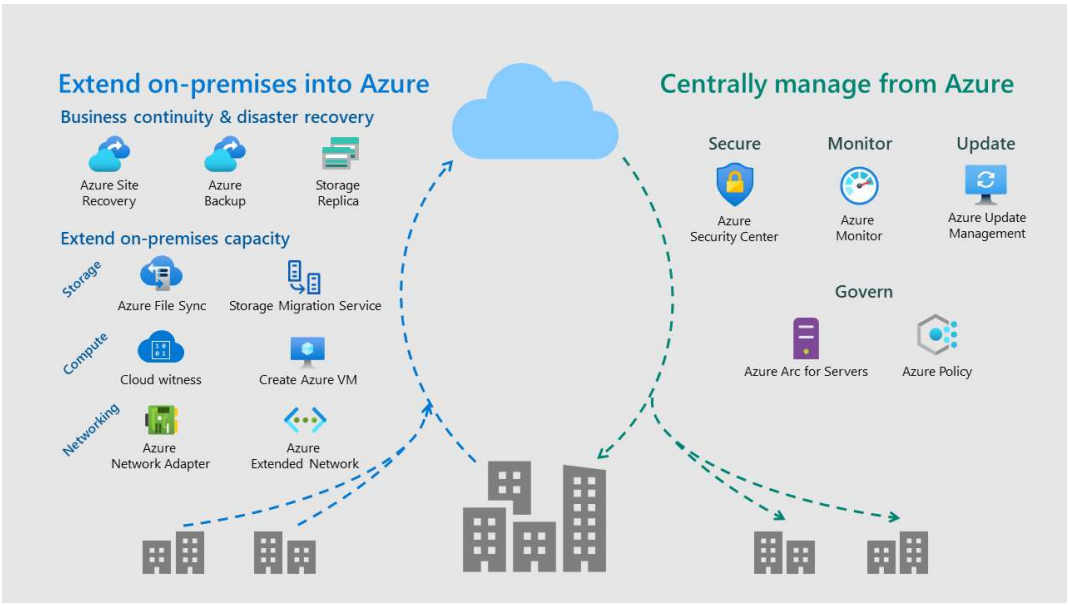
Cloud Storage es un servicio administrado para almacenar datos no estructurados. Almacena grandes volúmenes de datos y de cualquier tipo. Una de vez que los datos están dentro de Cloud Storage se puede conectar con facilidad con cualquier herramienta de Google Cloud como BigQuery, también puede ejecutar análisis de código abierto con Dataproc, o crear e implementar modelos de aprendizaje automático. Google Cloud Storage cuenta con 4 tipos de opciones de almacenamiento

Tipo de almacenamiento	Descripción	Ideal para
Standard Storage	Almacenamiento para datos a los que se accede con frecuencia (datos "activos") o que se almacenan solo durante períodos breves.	Datos activos, incluidos sitios web, videos en streaming y apps para dispositivos móviles.
Nearline Storage	Servicio de almacenamiento de alta durabilidad y bajo costo para almacenar datos a los que se accede con poca frecuencia.	Datos que se pueden almacenar por 30 días.
Coldline Storage	Un servicio de almacenamiento muy duradero y de bajo costo para almacenar datos a los que se accede con poca frecuencia.	Datos que se pueden almacenar durante 90 días.
Archive Storage	El servicio de almacenamiento de alta durabilidad y más bajo costo para el archivado de datos, copias de seguridad en línea y recuperación ante desastres.	Datos que se pueden almacenar durante 365 días.

¿Como funciona Google Cloud? Primero se crea un “Bucket” que son los contenedores donde se almacenan los datos, una vez subidos los datos se pueden descargar, compartir o administrar. Los precios de Cloud Storage varían según el tipo de almacenamiento



AZURE



Azure es el servicio de Nube desarrollado por Microsoft, ofrece servicios de Infraestructura para virtualizar Windows y Linux, generar copias de seguridad, almacenamiento, etc; Servicios de desarrollo de apps modernas, Azure permite crear gran variedad de aplicaciones, como soluciones web, multimedia, móvil y de línea de negocios, ofrece características de escalado automático integradas que nos ayudan a aumentar y reducir el escalado en función de las necesidades; Información basada en datos, a través de Azure podemos extraer información ya sean grandes o bajos volúmenes

de datos, nos proporciona servicios de SQL y NoSQL, compatibilidad integrada para realizar análisis y que nos ayuden a extraer la máxima información de los datos.

Azure agrupa sus servicios en 3 categorías, estos servicios pueden variar según la zona, las categorías son, Fundamentales: están disponibles en todas las regiones recomendadas y alternativas cuando la región está disponible con carácter general, o bien en un plazo de 90 días después de que un nuevo servicio fundamental esté disponible con carácter general. Estándar: disponibles en todas las regiones recomendadas en un plazo de 90 días a partir de la disponibilidad general de una región. Controlados por la demanda en regiones alternativas y muchos ya están implementados en un gran subconjunto de regiones alternativas. Estratégicos (anteriormente Especializados): ofertas de servicio dirigidas, a menudo centradas en el sector o respaldadas por hardware personalizado. Disponibilidad controlada por la demanda entre regiones, y muchos ya están implementados en un gran subconjunto de regiones recomendadas.

En conclusión y en base a mi investigación yo opto por elegir los servicios de AWS, es el servicio en nube con mayor experiencia en cuanto a años, los servicios que ofrecen abarcan todas las necesidades de FARISPLEY, desde el almacenamiento hasta su propio análisis de datos y machine learning. AWS nos proporciona una interfaz de usuario intuitiva y es de fácil adopción.

Data Lifecycle Management en AWS

AWS nos proporciona **AWS Glue** para nuestra integración de datos desde diversas fuentes y su transformación para su análisis. Luego escalaremos a Amazon S3 donde almacenaremos nuestros datos por su durabilidad y bajo costo. En el procesamiento de datos elegiremos **AWS Glue, Amazon EMR y Amazon Athena**, según los requisitos de la tarea. Y como punto final para el análisis y visualización de datos utilizaremos Amazon RedShift y AWS QuickSight

Si queremos realizar una pipeline en AWS lo primero que debemos realizar es la preparación de los datos, para ello utilizaremos **AWS Glue** para su limpieza y **Amazon S3** para su posterior almacenamiento. Luego podremos utilizar **Amazon SageMaker** para el entrenamiento de modelos predictivos. Para largar a producción los modelos entrenados podremos utilizar **AWS Lambda o Amazon ECS** para la interferencia en tiempo real. Como ultimo para el monitoreo y su optimización utilizaremos **Amazon Cloud Watch** donde veremos el rendimiento y realizar ajustes