

Ejercicio 2 Test GFB DS team

Tomando como referencia que el Ejercicio 1 se basó en servicios de AWS , se planteó la propuesta sobre estos servicios pero se proponen alternativas para responder siguientes preguntas:

1)

- a) Los Datos Identificados para el área operativa, serían los mínimos se deben extraer, pero la realidad es que conforme el equipo de Ciencia de Datos vaya analizando la información, se tendrán que aumentar los campos, por que el área operativa solo atiende un tema específico del banco y el equipo de datos debe tener un análisis global de cada uno de los negocios así como diferentes ópticas de cómo interpretar dicha información.
- b) Tomando como referencia que la intención es utilizar los servicios de AWS, utilizar el servicio de AWS Database Migration Service, por que esta toma de el log de transacciones para realizar las replicas por lo que no le impactaría a la transaccionalidad de las bases de fuentes de datos, hay que considerar el caso del del CRM para determinar la mejor estrategia.
 - i) Referencia : [Introducción a la replicación continua desde Amazon RDS para SQL Server mediante AWS Database Migration Service | Blog de base de datos de AWS](#)
 - ii) Se tendría que considera el uso de Apache Airflow para diseñar alguna carga y extracción de información del CRM
- c)
 - i) AWS-DMS , toma los datos de los log de transacciones
 - ii) AWS MWAA Amazon Managed Workflows for Apache Airflow :Referencia : [¿Qué es Amazon Managed Workflows for Apache Airflow \(MWAA\)? - Amazon Managed Workflows for Apache Airflow](#)
- d) Etapas a considerar para el ETL:
 - i) Lectura rápida e incremental
 - ii) Carga a Base de Paso con similar estructura a la del origen
 - iii) Limpieza de datos, implementando un proceso de mejora continua en las funciones de limpieza de datos dependiendo el tipo de campo
 - iv) Extracción de la Metadata de la información recuperada y agregación de metadata complementaria para su mejor manejo
 - v) Carga de información a Estructura homologada junto con su metadata
- e) AWS MWAA ó Apache Airflow
- f) Debería de ser una base similar al destino para simplificar la última inserción y no tener que aplicar nuevamente otra transacción
- g) AWS DMS ó AWS MWAA, Apache Airflow
- h) Se anexa archivo SVG

2) Seguridad

- a) Mediante la correcta administración e implementación de Roles, Políticas, Grupos de Seguridad, VPC,VPG junto con una VPN para enlazar la corporativo con la nube para proceso delicados.

Ejercicio 2 Test GFB DS team

3) Gobernanza de datos

- a) Mediante la implementación de Data Catalog para la implementación de un gobierno de datos, existen herramientas como Apache Atlas y Informatica Data Catalog.

4)

- a) Pros: Se debe considerar que Apache Atlas es una de las más eficientes herramientas para la creación de repositorios de metadatos porque permite recolectar la metadata, clasificarla, ordenar y validar el linaje de la misma como también el ciclo de vida de la información.
- b) Contras : Se debe considerar que al momento de implementar un Apache Atlas en el entorno AWS es posible que la integración no sea tan sencilla y transparente como se pudiera desear pero, considero que valdría la pena hacerlo.