UTEC Posgrado

MAESTRÍA

De Especialización en Ciencia de Datos e Inteligencia Artificial

Geraldo Colchado
Docente

UTEC Posgrado

CS8081 - Cloud Computing (Ciclo 2025-2)

Proyecto Final

Semana 10 - Enunciado

Contenido

Proyecto Final

Proyecto Final
Instrucciones

- Debe mantener el mismo grupo del Proyecto Parcial
- Se sugiere mantener la misma empresa elegida del Proyecto Parcial

Contenido

Proyecto Final

Proyecto Final
Parte A: Diseño Data Lakehouse

1. Tome como base su Mapa de Fuentes de Datos e identifique **entidades de datos relevantes** y las Fuentes de datos de donde obtenerlas:

Ejemplo:

Entidad de Datos	Fuentes de Datos
Clientes	BD Core Seguros RRGG (Interna, Estructurada) BD Core Seguros Vida (Interna, Estructurada) BD Core Seguros Salud (Interna, Estructurada) BD CRM (Interna, Estructurada)
Pólizas de Seguros	BD Core Seguros RRGG (Interna, Estructurada) BD Core Seguros Vida (Interna, Estructurada) BD Core Seguros Salud (Interna, Estructurada)
Pagos	BD Core Pagos (Interna, Estructurada)
Siniestros	BD CRM (Interna, Estructurada) BD Core Seguros RRGG (Interna, Estructurada) BD Core Seguros Vida (Interna, Estructurada) BD Core Seguros Salud (Interna, Estructurada) Fotos y Videos del siniestro (Externa, No Estructurado) Denuncia Policial (Externa, No Estructurado)
Comisiones	BD Comisiones (Interna, Estructurada)
Competencia	Redes Sociales (Externa, No Estructurada) Información pública de la competencia (Externa, Estructurada)

Proyecto Final
Parte A: Diseño Data Lakehouse

2. Plantee estrategias para transformar la Raw Data (Datos en bruto) y hacer analítica. Ejemplo:

Entidad de Datos (¿Qué?)	Fuentes de Datos	Estrategia para Transformar la Raw Data (Dejarlo listo para Analítica)	Estrategia de Analítica (¿Para Qué?)
Clientes	BD Core Seguros RRGG (Interna, Estructurada) BD Core Seguros Vida (Interna, Estructurada) BD Core Seguros Salud (Interna, Estructurada) BD CRM (Interna, Estructurada)	Se tienen 4 fuentes de datos y cada una tiene un código único para el cliente que usualmente es un correlativo. Se tiene que estandarizar (Ejemplo: Tipo Documento (RUC, DNI, CE, Pasaporte) + Número Documento)	Clusterización o Segmentación de clientes por geografía (Zonas (Lima, norte, sur, centro) o Departamentos), por volumen de compras (US$) de pólizas anuales (platino, oro, plata, bronce), etc.
Pólizas de Seguros	BD Core Seguros RRGG (Interna, Estructurada) BD Core Seguros Vida (Interna, Estructurada) BD Core Seguros Salud (Interna, Estructurada)	Se tienen 3 fuentes de datos y cada una tiene un código único para la Póliza que usualmente es un correlativo. Se tiene que estandarizar (Ejemplo: Tipo Póliza (RRGG, Vida, Salud) + Correlativo del Core)	Se aplicará Modelo de Machine Learning de Predicción de cancelación de pólizas de seguros. Se incluirán también otras entidades de datos.
Pagos	BD Core Pagos (Interna, Estructurada)	Incorporar el código estándar del Cliente y de Póliza de Seguros al cual corresponde el pago	Se aplicará Modelo de Machine Learning de Predecir los impagos. Se incluirán también otras entidades de datos.
Siniestros	BD CRM (Interna, Estructurada) BD Core Seguros RRGG (Interna, Estructurada) BD Core Seguros Vida (Interna, Estructurada) BD Core Seguros Salud (Interna, Estructurada) Fotos y Videos del siniestro (Externa, No Estructurado) Denuncia Policial (Externa, No Estructurado)	• Incorporar el código estándar del Cliente y de Póliza de Seguros al cual corresponde el siniestro. • Estandarizar el código del siniestro (Podría usarse el del CRM que es único). • Agregar metadata (Códigos estándar de cliente, Póliza de Seguros y Siniestro) a cada foto y denuncia policial como parte del nombre del archivo incluyendo un correlativo en caso haya más de una foto o denuncia policial en siniestro.	Se aplicará Machine Learning para procesar las fotos y videos del siniestro y reconocer automáticamente la placa del vehículo y el porcentaje de daños físicos al vehículo. Se aplicará Machine Learning (OCR) para extraer automáticamente texto de la denuncia policial y determinar su validez y relación con el siniestro.
Comisiones	BD Comisiones (Interna, Estructurada)	Incorporar el código estándar de Póliza de Seguros al cual corresponde la comisión	Se aplicará Modelo de Machine Learning de Detección de probabilidad de Fraude en cobro de comisiones. Se incluirán también otras entidades de datos.
Competencia	Redes Sociales (Externa, No Estructurada) Información pública de la competencia (Externa, Estructurada)	Establecer un identificador único de cada empresa aseguradora de la competencia y otros códigos para cada tema que se quiera almacenar de la competencia	Comparación de diferentes variables de la competencia con la empresa aseguradora.

Contenido

Proyecto Final

Proyecto Final
Parte B: Implementación Data Lakehouse

Elija 3 entidades de datos de la Parte A:

Catalog Layer (AWS Glue Data Catalog)
- Cree un Database
- Cree 3 tablas con su esquema en json

Storage Layer (AWS S3)
- Cree un archivo de texto en formato CSV por cada entidad y agregue como mínimo 100 registros ficticios.
- Cree 3 carpetas (una por entidad) en un bucket de S3 y suba los 3 archivos

Consumption Layer (AWS Athena)
- Muestre evidencia de ejecución de 3 queries que hagan join de 2 a 3 tablas del Data Catalog
- Muestre evidencia de 1 vista creada y consultada

Comentario: Si desea puede implementar algo equivalente en “Microsoft Fabric”, GCP, https://www.databricks.com/, https://www.snowflake.com/

Contenido

Proyecto Final

Proyecto Final
Plazo de Ejecución y Entrega

- **Plazo de Ejecución:** Del Miércoles 15-Octubre-2025 al Martes 4-Noviembre-2025 (3 semanas)
- **Entrega:** Máximo el Martes 4-Noviembre-2025 23:59 por correo al docente gcolchado@utec.edu.pe (sólo 1 persona del grupo lo debe enviar y copiarle al resto del grupo)

Contenido

Proyecto Final

1. Instrucciones
2. Parte A: Diseño Data Lakehouse
3. Parte B: Implementación Data Lakehouse
4. Plazo de Ejecución y Entrega
5. **Fecha de Exposición**
6. Rúbrica

Proyecto Final
Fecha de Exposición

- **Fecha de Exposición:** Miércoles 5-Noviembre-2025 19h a 21h (Virtual)
- **Tiempo:** Máximo 20 minutos x grupo (Hay 5 Grupos)

Contenido

Proyecto Final

Proyecto Final
Rúbrica

- **Parte A - Diseño Data Lakehouse:** 12 puntos
- **Parte B - Implementación Data Lakehouse:** 6 puntos
- **Exposición:** 2 puntos

Gracias

Docente: Geraldo Colchado


UTEC Posgrado