Introducción

La gestión y análisis de grandes volúmenes de datos es crucial en el entorno digital actual. Entender las arquitecturas típicas, las estrategias de almacenamiento y las aplicaciones prácticas permite a las organizaciones tomar decisiones informadas y obtener ventajas competitivas. Además, conocer el perfil del científico de datos y sus habilidades complementa el entendimiento del proceso de transformación de datos en información valiosa. Esta actividad está diseñada para que los estudiantes trabajen de manera práctica en un proyecto que abarque todos estos temas de manera integral.

Objetivo

Aplicar de manera práctica e integrada los conocimientos sobre arquitecturas de proyectos de datos masivos, estrategias de almacenamiento, el perfil del científico de datos y las áreas de aplicación de la ciencia de datos.

Caso De Uso De La Industria De Las Telecomunicaciones

Objetivo del proyecto

Reducción de la Deserción (CHURN)
Los operadores de internet pierden varios miles de suscriptores al año debido a la deserción de aquellos que experimentan baja calidad en el servicio. La relevancia del proyecto se puede medir directamente en términos de la retención de suscriptores que están en peligro de deserción. Si un ARPU típico en Latinoamérica es de alrededor de 20 USD por suscriptor y se logra retener, por ejemplo, a un 10% de estos suscriptores, se puede entender que el operador estará ahorrando significativamente en términos de este porcentaje. Algunas empresas públicas reportan deserciones anuales de entre el 2% y el 6%, y con una base

de suscriptores generalmente por encima de 500 mil, los operadores podrían perder entre 2 y 6 millones de dólares al año. Si se logra retener el 10% de estos suscriptores, el ahorro podría ser considerable y el proyecto adquiriría mayor relevancia.

Fuentes de datos

- Comprender las **fuentes heterogéneas de datos:** identificar y clasificar diferentes tipos de fuentes de datos (estructurados, semiestructurados, no estructurados) y comprender sus características y desafíos.
  Tabla de Suscriptores
  - Datos Estructurados
  - MySQL

Muestra

CUST_ID	CITY	CMTS_ID	CABLE_MAC	KB_DOWN	KBPS_DOWN	PROFILE_MAX_BW_DOWN
345D9E981995	Metropolis	MET01-E6K	cable-mac 36	136,067	302	1,126,400
C005C266BF1B	Gotham_City	GOT02-E6K	cable-mac 31	5,849	6	320,000
400D108094DB	Gotham_City	GOT02-E6K	cable-mac 33	260,926	97	425,000
ACF8CC5BA755	Emerald_City	EMC02-E6K	cable-mac 15	1,737	1	43,000
50A5DC2B5BD7	Hogsmeade	HOG03-E6K	cable-mac 32	143,424	46	210,000
C005C2673B3B	Minas_Tirith	MNT02-E6K	cable-mac 15	2,353,881	872	76,800
ACF8CC6E682C	Minas_Tirith	MNT02-E6K	cable-mac 15	259,654	82	43,000
6455B13C9802	Minas_Tirith	MNT02-E6K	cable-mac 15	732	0	33,000

Información Personal de Suscriptores

- Datos Semi-Estructurados
- mongoDB

Muestra


```
{
  "_id": {
    "$oid": "669ad620ca290ed38f86e1fa"
  },
  "CM_MAC": "ACF8CC6E682C",
  "PATH_2": "Minas_Tirith",
  "latitude": 36.058198038228625,
  "longitude": -117.65747009818804,
  "first_name": "Maria",
  "last_name": "Stewart",
  "street_address": "Matthew Crossroad",
  "street_number": "7608"
},
{
  "_id": {
    "$oid": "669ad620ca290ed38f86e1fb"
  },
  "CM_MAC": "6455B13C9802",
  "PATH_2": "Minas_Tirith",
  "first_name": "Joshua",
  "last_name": "Martinez",
  "street_address": "Smith Springs",
  "street_number": "99132"
},
{
  "_id": {
    "$oid": "669ad620ca290ed38f86e1fc"
  },
  "CM_MAC": "ACF8CC161F8A",
  "PATH_2": "Zion",
  "first_name": "Carol",
  "last_name": "Miller"
},
```


Grabación de Llamadas Centro de Atención Telefónica

- Datos No Estructurados

- MySQL

Muestra

Archivo de Audio	Date	Tamaño MB
001A2B3C4D5E_4.aac	20/07/24	1.8
112B3C4D5E6F_8.aac	21/07/24	3.1
223C4D5E6F7G_2.aac	22/07/24	6.1
334D5E6F7G8H_2.aac	23/07/24	2.8
445E6F7G8H9I_1.aac	19/07/24	4
556F7G8H9I0J_1.aac	20/07/24	4
667G8H9I0J1K_6.aac	21/07/24	5.2
778H9I0J1K2L_1.aac	22/07/24	8

Necesita procesamiento específico, conversión de audio a texto y análisis de sentimiento para poder usarse adecuadamente.

Procesos ETL

- Diseñar **procesos ETL:** crear un esquema detallado del proceso ETL, describiendo métodos de extracción, transformación y carga de datos, y seleccionar herramientas ETL adecuadas para diferentes tipos de datos.

  - Transformaciones de características,
  - normalización
  - estandarización
  - manejo de valores faltantes.

De entre las herramientas de transformación que se aplicarían en el proceso se encuentran las siguientes:

- Dremio
- KNIME
- Python Pandas
- Python SKlearn

Ejemplo de procesamiento en KNIME

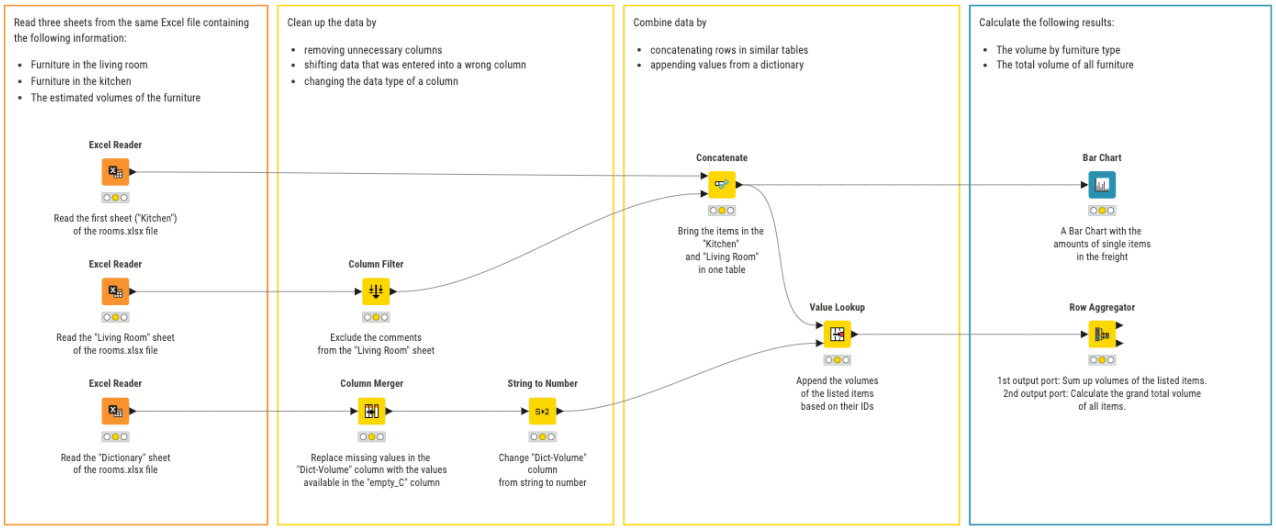
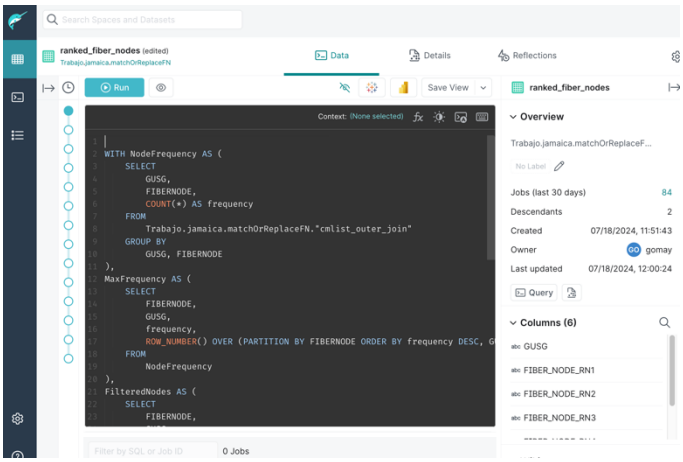


Ilustración 1 Grafica Tipo Proceso de ETL

Ejemplo de Procesamiento en Dremio

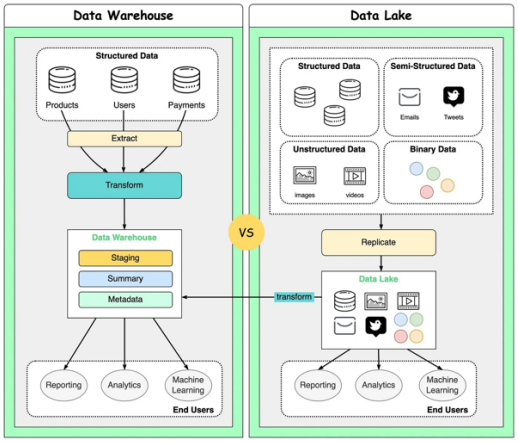


Arquitectura de Almacenamiento

- Planificar la **arquitectura de almacenamiento:** diseñar una arquitectura de almacenamiento de datos que incluya soluciones como Data Lake y Data Warehouse, explicando las ventajas y aplicaciones de cada uno.

Aquí podemos poner las fuentes que irían en el warehose:

- Tablas limpias y transformadas de las fuentes de monitoreo
- Análisis de sentimiento de las grabaciones
- Tablas Limpias del sistema de alarmes de Interrupción de servicio

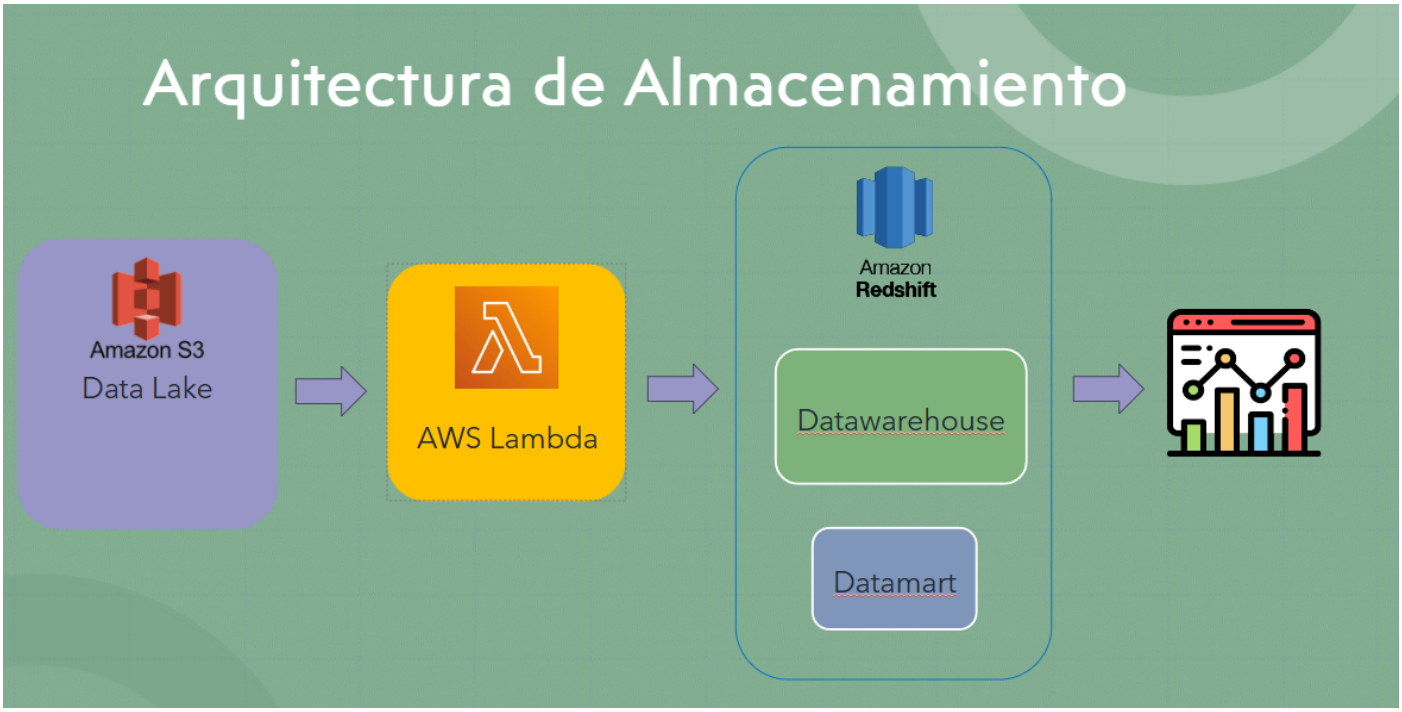


Aquí podemos poner las fuentes que irían en el datalake :

- Archivos de Audio del Centro de atención telefónica
- Menciones Twitter:
- Estadísticas de calidad
- Net promoter score
- Fuentes de monitoreo
- Subscriber Geolocation
- Alarmas de Interrupción de Servicio. Resumen meteorológico histórico por población.

Tablas procesadas, limpias con políticas de tratamiento de datos faltantes, relacionables y relevantes al proyecto

ETL



Fuentes de datos de diferentes fuentes y áreas dentro de la organización.

Resultados disponibles para el consume desde otras áreas de la organización

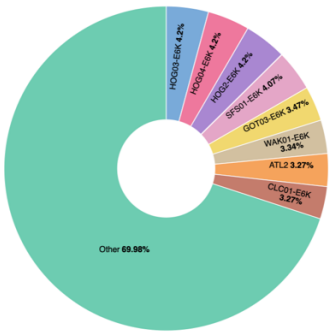
Estrategias de Tratamiento de Datos

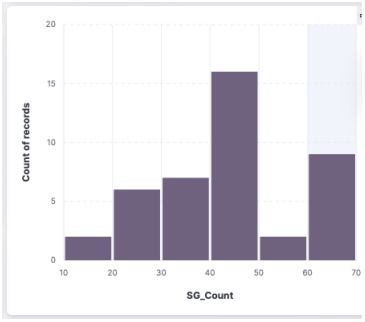
- Definir **estrategias de tratamiento de datos:** proponer estrategias para la limpieza, integración y análisis de datos, asegurando la calidad y coherencia de los datos para su posterior análisis.

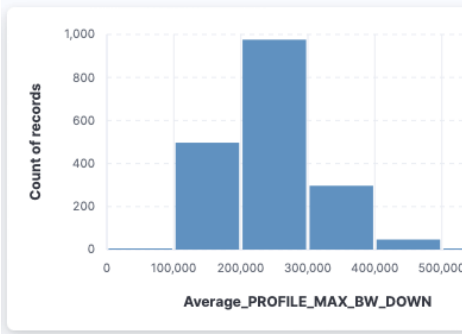
Temporalidad de datos según la fuente
Entorno significativo de términos
Conversión de campos para su comparación global
Mapa de valores para proteger identidad de usuarios en los análisis.

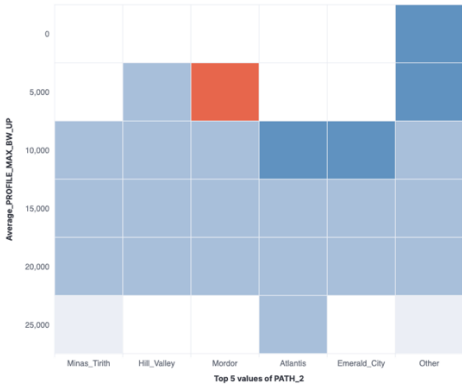
- Desarrollar **planes de visualización de datos:** diseñar *dashboard* interactivos y planificar la visualización de datos utilizando herramientas como Tableau, Power BI o matplotlib para representar métricas clave y tendencias relevantes.

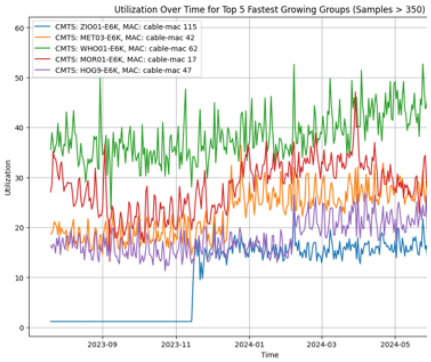
Valor del histograma y los diagramas relacionales.
PowerBI o Kibana, Grafana











Perfil del científico de datos:

- Comprender el **perfil del científico de datos:** analizar las competencias necesarias en ciencias de la computación, matemáticas, comunicación y negocios que forman el perfil del científico de datos.

KPI o indicadores clave de desempeño en las estadísticas de Telecomunicaciones.
Competencias:

- Estadística
- Telecomunicaciones
- Ciencia de datos
- Mercadotecnia

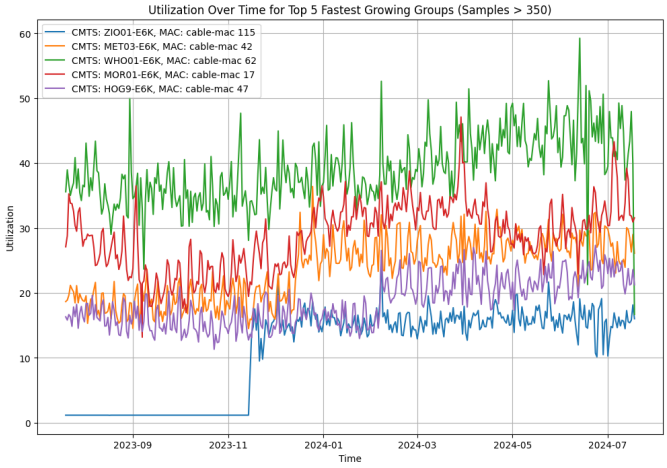
- Explorar **estrategias en almacenamiento masivo:** investigar y entender las diferencias entre Data Mart, Data Warehouse y Data Lake, y nuevas tendencias en almacenamiento masivo.

La estrategia es empezar a mover la estructura de almacenamiento a la nube para poder sacar provecho del rendimiento, la flexibilidad, las herramientas de acceso y colaboración que ofrecen los proveedores de servicios en la nube

Estrategias De Ciencia De Datos Y Datos Masivos

- Aplicar **estrategias de ciencia de datos y datos masivos:** proponer aplicaciones prácticas de la ciencia de datos en inteligencia de negocio, analítica de negocio, minería de datos, aprendizaje automático e inteligencia artificial en distintos sectores.

Regresión lineal de mediciones de utilización por Grupos de servicio, análisis predictivo y alarmas de proactividad, que permitan al operador de internet reaccionar con tiempo para incrementar los recursos asignados a un grupo de servicio.



Regresion lineal de mediciones de utilización por Grupos de servicio, analisis predictivo y alarmas de proactividad, que permitan al operador de internet reaccionar con tiempo para incrementar los recursos asignados a un grupo de servicio.

Grafica del “top 5” grupos de servicio con mayor crecimiento en en la Operación.

Utilización del módulo de Python Solear

Resumen del Módulo sklearn

El módulo sklearn (también conocido como Scikit-learn) es una biblioteca de Python diseñada para realizar tareas de aprendizaje automático y minería de datos. Ofrece herramientas eficientes y fáciles de usar para:

- Preprocesamiento de Datos:
- Algoritmos de Aprendizaje Supervisado:
- Algoritmos de Aprendizaje No Supervisado:
- métodos de conjunto como Bagging, Boosting, y Random Forest.
- Permite construir pipelines que integran múltiples etapas de preprocesamiento y modelado de manera secuencial.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830. Retrieved from https://scikit-learn.org/stable/

Un ejemplo de pseudocódigo para el propósito del proyecto sería el siguiente:

Predicción de Deserción con Bosques Aleatorios

```
1. Cargar Datos:
    - Leer archivo CSV 'customer_data.csv' en un dataframe llamado 'data'.
```

```
2. Ingeniería de Características:
    - Para cada 'user_id' en 'data':
            - Calcular 'avg_daily_usage' (uso diario promedio).
            - Calcular 'usage_variance' (varianza en el uso diario).
      - Calcular 'quality_score' como 'speed' dividido por 'latency'.
      - Calcular 'downtime_count' como la cantidad de tiempos de inactividad.

3. Preparación de Etiquetas:
    - Crear una nueva columna 'churn' en 'data':
        - Si 'subscription_status' es 'churned', asignar 1.
        - Si no, asignar 0.

4. Selección de Características y Etiquetas:
    - Definir 'features' como ['avg_daily_usage', 'usage_variance', 'quality_score',
'downtime_count', 'geolocation'].
    - Definir 'X' como 'data[features]'.
    - Definir 'y' como 'data['churn']'.

5. División de Datos:
    - Dividir 'X' y 'y' en conjuntos de entrenamiento ('X_train', 'y_train') y prueba
('X_test', 'y_test').

6. Entrenamiento del Modelo:
    - Crear un modelo RandomForestClassifier con 100 árboles y random_state=42.
    - Entrenar el modelo usando 'X_train' y 'y_train'.

7. Evaluación del Modelo:
    - Predecir 'y_pred' usando 'X_test'.
    - Imprimir el reporte de clasificación y el ROC-AUC.

8. Importancia de las Características:
    - Obtener la importancia de las características del modelo.
    - Imprimir la importancia de cada característica.
```

Conclusión

El conocimiento al que podemos llegar con el procesamiento de la información recolectada nos puede llevar a reducir considerablemente la tasa de deserción del operador logrando hacer análisis predictivos para determinar que suscriptores van a desertar e incluso que probabilidad de deserción se le puede atribuir a cada cliente y buscar las medidas y la asignación de recursos correspondientes.

Churn mensual y anual:

En la industria de telecomunicaciones, las tasas de churn mensual suelen estar alrededor del 1.9% para servicios postpago, lo que puede traducirse en una tasa anual cercana al 20% .

Para los servicios prepago, las tasas de churn pueden ser significativamente más altas, llegando hasta un 67% anual en algunos casos .

Estas tasas pueden variar dependiendo de diversos factores como la calidad del servicio, la satisfacción del cliente, y las estrategias de retención aplicadas por la empresa.

Xoxoday. (2024). Customer Churn Rate: The Complete Guide for 2024. Recuperado de xoxoday.com
Heavy.AI Team. (2021). Strategies for Reducing Churn Rate in the Telecom Industry. Recuperado de heavy.ai

Rúbrica

Criterios	Descripción	Puntuación máxima (puntos)	Peso %
Diseño del proyecto		3	30 %
Definición del problema y objetivos	1-2: Objetivos y problemas mal definidos o inexistentes. 3-4: Objetivos definidos, pero problemas no claros o incompletos. 5-6: Objetivos y problemas claramente definidos, con algunos detalles menores faltantes. 7-8: Objetivos y problemas bien definidos y coherentes. 9-10: Objetivos y problemas perfectamente definidos y coherentes con el contexto del sector elegido.	1	10 %
Diagrama del Flujo ETL	1-2: Diagrama incompleto o confuso. 3-4: Diagrama claro, pero con algunos pasos ETL faltantes o mal descritos. 5-6: Diagrama claro y completo con pequeñas mejoras necesarias. 7-8: Diagrama detallado y completo, con detalles menores mejorables. 9-10: Diagrama detallado, claro y completo, sin errores.	1	10 %

Estructura de almacenamiento y tratamiento de datos	1-2: Estructura de almacenamiento no adecuada o inexistente. 3-4: Estructura básica, pero con falta de detalles importantes. 5-6: Estructura adecuada con la mayoría de los detalles necesarios. 7-8: Estructura bien diseñada y justificada, con pequeños detalles mejorables. 9-10: Estructura bien diseñada y justificada, con todos los detalles necesarios.	1	10%
Justificación de las decisiones tomadas		3	30 %
Selección de herramientas ETL	1-2: Herramientas no adecuadas o sin justificación. 3-4: Herramientas adecuadas, pero con justificación insuficiente. 5-6: Herramientas bien elegidas con justificación adecuada. 7-8: Herramientas bien elegidas y justificadas con pequeños detalles mejorables. 9-10: Herramientas perfectamente elegidas y justificadas con argumentos sólidos.	1	10 %
Solución de almacenamiento	1-2: Solución de almacenamiento no adecuada o sin justificación. 3-4: Solución adecuada, pero con justificación insuficiente. 5-6: Solución bien elegida con justificación adecuada. 7-8: Solución bien elegida y justificada con pequeños detalles mejorables. 9-10: Solución perfectamente elegida y justificada con argumentos sólidos.	1	10 %
Selección de herramientas de visualización	1-2: Herramientas no adecuadas o sin justificación. 3-4: Herramientas adecuadas, pero con justificación insuficiente. 5-6: Herramientas bien elegidas con justificación adecuada. 7-8: Herramientas bien elegidas y justificadas con pequeños detalles mejorables. 9-10: Herramientas perfectamente elegidas y justificadas con argumentos sólidos.	1	10 %
Calidad de la presentación y comunicación		2	20 %

Informe ejecutivo	1-2: Informe confuso, incompleto o mal estructurado. 3-4: Informe adecuado, pero con falta de claridad o detalle. 5-6: Informe claro y bien estructurado con la mayoría de los detalles. 7-8: Informe claro y bien estructurado, con pequeños detalles mejorables. 9-10: Informe perfectamente claro, detallado y bien estructurado.	1	10 %
Presentación oral en vídeo	1-2: Presentación desorganizada, poco clara o incompleta. 3-4: Presentación adecuada, pero con falta de claridad o detalle. 5-6: Presentación clara y organizada con la mayoría de los detalles. 7-8: Presentación clara y organizada con pequeños detalles mejorables. 9-10: Presentación clara, organizada y detallada, con excelente comunicación.	1	10 %
Innovación y Creatividad		2	20 %
Aplicación de técnicas de ciencia de datos	1-2: Aplicación inapropiada o inexistente. 3-4: Aplicación básica con falta de innovación. 5-6: Aplicación adecuada con algunos elementos innovadores. 7-8: Aplicación adecuada con innovación notable, con pequeños detalles mejorables. 9-10: Aplicación creativa y bien fundamentada, con alta innovación.	1	10 %
Propuestas de mejora y nuevas tendencias	1-2: Propuestas inexistentes o irrelevantes. 3-4: Propuestas básicas con falta de detalle o relevancia. 5-6: Propuestas adecuadas con buenos argumentos. 7-8: Propuestas adecuadas con innovación notable, con pequeños detalles mejorables. 9-10: Propuestas innovadoras, bien detalladas y altamente relevantes.	1	10 %
		10	**100 %**