Limpieza y análisis exploratorio de datos de movilidad urbana

Llevamos a cabo un análisis del dataset de movilidad urbana con datos GPS de autobuses respecto al mes de marzo de 2024 con el objetivo de identificar patrones útiles para optimizar la red de autobuses urbanos y reducir tiempos de espera.

En el dataset se detectaron registros duplicados, se detectaron 8 outliers y valores nulos en las variables numéricas. Para corregir esto se eliminaron los duplicados, se imputaron nulos mediante mediana que es robusta frente a outliers y se tomó la decisión de dejar los valores atípicos, puesto que pueden aportar información sobre picos puntuales de demanda. También se normalizó la variable fecha_hora para poder llevar a cabo un análisis temporal, se comprobó que la variable parada sigue un formato homogéneos, se generaron las variables hora, dia_semana y laboral; y codificamos la variable evento_climatico mediante la técnica de one-hot encoding para evitar la introducción de órdenes artificiales.

- Hallazgos clave

El análisis de la afluencia de pasajeros por parada mostró una distribución muy homogénea entre la mayoría de las paradas, ya que salvo en la parada_8, la mediana del número de pasajeros es 8 en todas. No obstante, el análisis de la media por parada nos permite observar pequeñas diferencias que, probablemente se deban a picos de demanda puntuales, y que hacen que las patadas con mayor afluencia sean la Parada_6, la Parada_4 y la Parada_5, respectivamente.

En cuanto a la variabilidad del tiempo de espera según el clima, podemos observar que el comportamiento es similar en cualquiera de los escenarios, ya que la mediana es 6 minutos en todos los climas, excepto en tormenta que es de 6,05. No obstante, al analizar la media, podemos ver que si las condiciones meteorológicas son peores, si aumenta el tiempo de espera, sobre todo en tormentas, lo que refuerza la idea de que existan picos puntuales en nuestro dataset y uno de los principales motivos sea la climatología.

Si definimos el concepto congestión como mayor tiempo de espera, estudiaremos la media y mediana del tiempo de espera en franjas horaria de 3 horas, comenzando la primera de ellas a medianoche. La franja horaria con mayor media es de 18-21, no obstante la que mayor mediana presenta corresponde a 9-12.

Esta diferencia sugiere que en el intervalo de 18-21 se podrían haber producido picos puntuales que han disparado la media, pero que la franja horaria de 9 a 12 es la que presenta una mayor congestión de forma habitual y sostenida.

Por último, analizamos un heatmap cruzando franja horaria y clima, donde cada celda representa el tiempo medio de espera. Así, podemos identificar interacciones entre el momento del día y la meteorología, pudiendo observar que la mayor congestión se produce en días de tormenta entre las 18 y las 21. Las franjas nocturnas son menores y más estables, por lo que el impacto del clima se intensifica, sobre todo, en las horas punta.

En conclusión, la congestión en la red de autobuses no es uniforme, sino que depende de la combinación de factores temporales y externos. Las horas punta, especialmente entre las 9 y las 12 y entre las 18 y las 21, concentran los mayores tiempos de espera, siendo este efecto más significativo cuando se dan condiciones meteorológicas adversas como la tormenta.

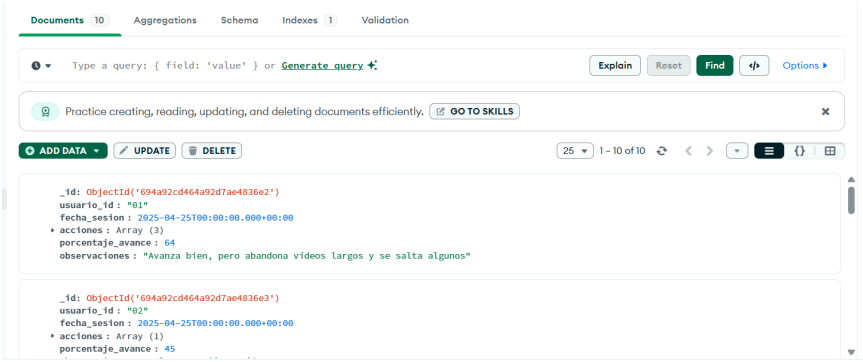
Por tanto, las estrategias de optimización del servicio deberían centrarse en refuerzos por franja horaria y en planes específicos según la previsión meteorológica, más que en cambios homogéneos en toda la red.

Almacenamiento NoSQL con MongoDB

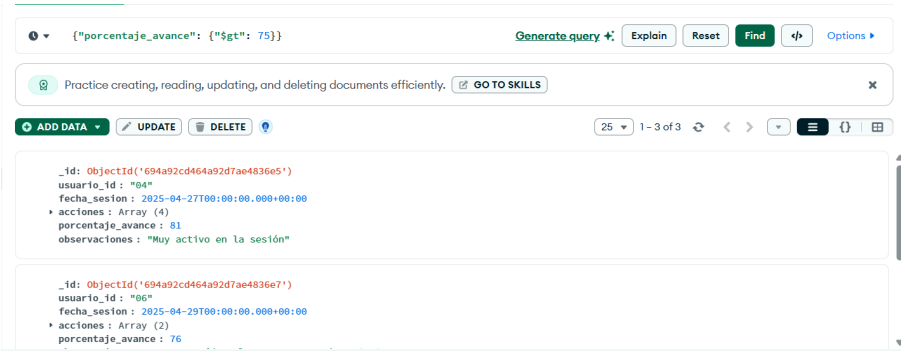
El objetivo de este análisis es diseñar un modelo de almacenamiento NoSQL en MongoDB y analizar el comportamiento de los usuarios durante sus cursos, de manera que puedan convertir los datos en información valiosa.

El diseño del modelo de datos se ha basado en una estructura de documento JSON flexible para adaptarse a la naturaleza semiestructurada de la interacción de los usuarios. Este es un modelo de ejemplo de los documentos en JSON utilizados para nuestra base de datos:

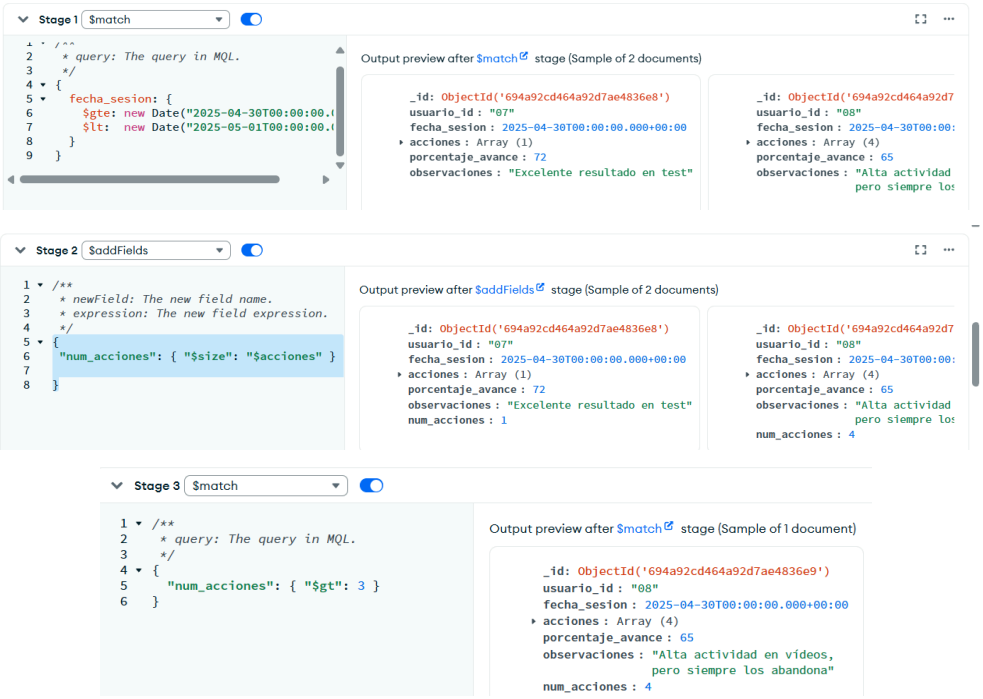
```
{
  "_id": ObjectId("..."),
  "usuario_id": "01",
 "fecha_sesion": { "$date": "2025-04-25T00:00:00Z" },
    "acciones": [
      { "tipo": "video", "id_contenido": "V101", "duracion_min": 12 },
      { "tipo": "video", "id_contenido": "V103", "duracion_min": 17 },
      { "tipo": "test", "id_contenido": "T021", "duracion_min": 15, "resultado": 80 }
    ],
    "porcentaje_avance": 64,
    "observaciones": "Avanza bien, pero abandona vídeos largos y se salta algunos"
}
```



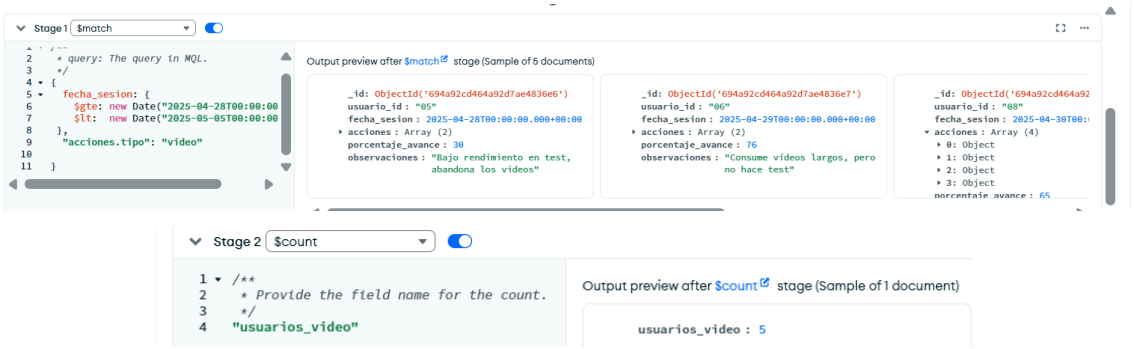
Tras la inserción de los 10 documentos simulados (ver anexo) y la ejecución de consultas mediante MQL (MongoDB Query Language), se extraen los siguientes hallazgos:



- Usuarios con alto rendimiento: detectamos que los usuarios 04, 06 y 09 han avanzado más de un 75%, lo cual significa que van muy adelantados y que visualizan los contenidos de la plataforma de forma progresiva.



- Acciones de usuarios en una fecha concreta: detectamos que en la sesión del 30/04/2025 tan sólo un usuario realizó más de tres acciones. Este usuario es el usuario 08. Esto podría indicar que nuestros usuarios cuando se conectan a la plataforma no realizan muchísimas acciones en un mismo día, sino que lo hacen más de forma progresiva.



- En el análisis, la “semana pasada” se considera los últimos 7 días respecto a la última fecha del dataset, es decir del 28 de abril al 4 de mayo de 2025. Realizando las agregaciones necesarias, podemos comprobar que 5 usuarios vieron al menos un vídeo durante la semana pasada, es decir, un 50% de los usuarios de los cuales tenemos datos. Esto podría implicar que los vídeos es el principal atractivo de los usuarios cuando interactúan con nuestra plataforma.

Por tanto, podemos concluir con este análisis de las interacciones de los usuarios que hay grandes diferencias en cómo los usuarios interactúan con la plataforma y los ritmos en la que cada uno sigue el avance del curso.

Además, debemos estar en alerta de los usuarios que no interactuaron con ningún vídeo la semana pasada, pues puede ser que estén pensando en abandonar la plataforma. Sería una buena práctica considerar la posibilidad de personalizar una campaña de re-engagement específica para ellos).

Por último, podríamos premiar de alguna forma a los usuarios más activos, ya sea con descuentos en otros cursos, de manera que los fidelizamos a nuestra plataforma sabiendo que son usuarios muy activos y que pueden generar interés por los cursos en su propio círculo.

Procesamiento distribuido con Hadoop (MapReduce)

Se renuncia a la evaluación de esta parte.

Anexo - Documentos insertados

- Código inserción de documentos:


```
[
 {
  "usuario_id": "01",
  "fecha_sesion": { "$date": "2025-04-25T00:00:00Z" },
  "acciones": [
   { "tipo": "video", "id_contenido": "V101", "duracion_min": 12 },
   { "tipo": "video", "id_contenido": "V103", "duracion_min": 17 },
   { "tipo": "test", "id_contenido": "T021", "duracion_min": 15, "resultado": 80 }
  ],
  "porcentaje_avance": 64,
  "observaciones": "Avanza bien, pero abandona vídeos largos y se salta algunos"
 },
 {
  "usuario_id": "02",
  "fecha_sesion": { "$date": "2025-04-25T00:00:00Z" },
  "acciones": [
   { "tipo": "video", "id_contenido": "V103", "duracion_min": 45 }
  ],
  "porcentaje_avance": 45,
  "observaciones": "Solo consumió un vídeo"
 },
 {
  "usuario_id": "03",
  "fecha_sesion": { "$date": "2025-04-26T00:00:00Z" },
  "acciones": [
   { "tipo": "test", "id_contenido": "T202", "duracion_min": 10, "resultado": 65 },
   { "tipo": "test", "id_contenido": "T203", "duracion_min": 12, "resultado": 70 }
  ],
  "porcentaje_avance": 60,
  "observaciones": "No ve videos"
 },
 {
  "usuario_id": "04",
  "fecha_sesion": { "$date": "2025-04-27T00:00:00Z" },
  "acciones": [
   { "tipo": "video", "id_contenido": "V104", "duracion_min": 34 },
   { "tipo": "video", "id_contenido": "V105", "duracion_min": 40 },
   { "tipo": "video", "id_contenido": "V106", "duracion_min": 25 },
   { "tipo": "test", "id_contenido": "T204", "duracion_min": 18, "resultado": 90 }
  ],
  "porcentaje_avance": 81,
  "observaciones": "Muy activo en la sesión"
 },
 {
  "usuario_id": "05",
  "fecha_sesion": { "$date": "2025-04-28T00:00:00Z" },
  "acciones": [
   { "tipo": "video", "id_contenido": "V107", "duracion_min": 6 },
   { "tipo": "test", "id_contenido": "T205", "duracion_min": 9, "resultado": 55 }
  ],
  "porcentaje_avance": 30,
  "observaciones": "Bajo rendimiento en test, abandona los videos"
 },
 {
  "usuario_id": "06",
  "fecha_sesion": { "$date": "2025-04-29T00:00:00Z" },
  "acciones": [
   { "tipo": "video", "id_contenido": "V108", "duracion_min": 25 },
   { "tipo": "video", "id_contenido": "V109", "duracion_min": 18 }
  ],
  "porcentaje_avance": 76,
  "observaciones": "Consume vídeos largos, pero no hace test"
 },
 {
```

```
  "usuario_id": "07",
  "fecha_sesion": { "$date": "2025-04-30T00:00:00Z" },
  "acciones": [
   { "tipo": "test", "id_contenido": "T206", "duracion_min": 14, "resultado": 95 }
  ],
  "porcentaje_avance": 72,
  "observaciones": "Excelente resultado en test"
 },
 {
  "usuario_id": "08",
  "fecha_sesion": { "$date": "2025-04-30T00:00:00Z" },
  "acciones": [
   { "tipo": "video", "id_contenido": "V110", "duracion_min": 10 },
   { "tipo": "video", "id_contenido": "V111", "duracion_min": 10 },
   { "tipo": "video", "id_contenido": "V112", "duracion_min": 10 },
   { "tipo": "video", "id_contenido": "V113", "duracion_min": 10 }
  ],
  "porcentaje_avance": 65,
  "observaciones": "Alta actividad en vídeos, pero siempre los abandona"
 },
 {
  "usuario_id": "09",
  "fecha_sesion": { "$date": "2025-05-01T00:00:00Z" },
  "acciones": [
   { "tipo": "video", "id_contenido": "V114", "duracion_min": 27 },
   { "tipo": "test", "id_contenido": "T207", "duracion_min": 16, "resultado": 72 },
   { "tipo": "test", "id_contenido": "T208", "duracion_min": 11, "resultado": 78 }
  ],
  "porcentaje_avance": 91,
  "observaciones": "Progreso muy alto, cercano a finalizar"
 },
 {
  "usuario_id": "10",
  "fecha_sesion": { "$date": "2025-05-01T00:00:00Z" },
  "acciones": [
   { "tipo": "video", "id_contenido": "V115", "duracion_min": 9 },
   { "tipo": "video", "id_contenido": "V116", "duracion_min": 13 },
   { "tipo": "test", "id_contenido": "T209", "duracion_min": 20, "resultado": 85 }
  ],
  "porcentaje_avance": 74,
  "observaciones": "Ve los videos poco, pero bien en test"
 }
]
```