UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

DESARROLLO Y GESTIÓN DE SOFTWARE

EXTRACCIÓN PARA CONOCIMIENTOS EN BASES DE DATOS

II.1. Reporte de limpieza de datos

DOCENTE:

ING. LUIS ENRIQUE MASCOTE CANO

PRESENTA:

DARON TARÍN GONZÁLEZ

MATRÍCULA: 1123250008

GRUPO:

IDGS91N

Chihuahua, Chih., 05 de octubre de 2025

Contenido

Introducción

En la era digital, los servicios de streaming representan una de las mayores fuentes de generación y consumo de datos. Tanto las plataformas de **contenido audiovisual** (como Netflix, Disney+ y Spotify) como las de **videojuegos en la nube** (como Xbox Cloud Gaming, GeForce Now o PlayStation Plus) producen información masiva a partir del comportamiento de los usuarios, las transacciones y los sistemas automatizados.
Este reporte analiza y compara la **procedencia**, **tipos y fuentes de datos**, así como las **técnicas de limpieza** aplicadas en ambos sectores, con el objetivo de comprender la relevancia del tratamiento y depuración de datos para mantener la calidad, precisión y utilidad de la información.

1. Procedencia de los datos

En los servicios de streaming, los datos provienen de diversas fuentes y procesos:

- **Datos generados por humanos:** son los que los usuarios proporcionan directa o indirectamente, como registros de cuentas, búsquedas, calificaciones, tiempo de visualización o juegos jugados. En Netflix, por ejemplo, estos datos ayudan a personalizar las recomendaciones (Netflix Tech Blog, 2024).
- **Datos de transacciones:** incluyen información sobre suscripciones, métodos de pago y consumo de contenido, fundamentales para el control financiero y la segmentación de mercado.
- **Datos máquina a máquina (M2M):** los sistemas de streaming recogen métricas automáticamente desde dispositivos conectados (consolas, televisores inteligentes o servidores) para monitorear rendimiento y latencia.
- **Datos web y redes sociales:** las plataformas rastrean menciones y opiniones en redes (Twitter/X, Reddit, YouTube) para evaluar tendencias o detectar problemas de servicio (Microsoft Azure, 2024).
- **Datos biométricos (caso emergente):** algunos servicios de juegos exploran métricas fisiológicas (como ritmo cardíaco o movimiento ocular) para experiencias inmersivas, lo que incrementa la sensibilidad del manejo de datos (NVIDIA Research, 2024).

En síntesis, ambas industrias recogen datos tanto de interacción humana como de procesos automáticos, pero los **servicios de videojuegos tienden a generar más datos en tiempo real** por la naturaleza interactiva del juego.

2. Tipos y fuentes de datos

Los datos en el contexto del streaming se clasifican según su naturaleza y formato:

Tipo de dato	Descripción	Ejemplo en streaming audiovisual	Ejemplo en streaming de videojuegos
Cuantitativo	Numéricos, medibles.	Tiempo de reproducción, número de vistas.	FPS, ping, tiempo de sesión.
Cualitativo	Descriptivos o categóricos.	Opiniones y reseñas de usuarios.	Comentarios de jugadores.
Estructurados	Organizados en bases de datos tabulares.	Información de usuarios y facturación.	Estadísticas de partidas.
No estructurados	Formato libre o multimedia.	Imágenes, audios, subtítulos.	Clips de juego, transmisiones.
Nominales	Categorizan sin jerarquía.	Género de serie o película.	Título del juego.
Ordinales	Implican orden o nivel.	Clasificación por edad o popularidad.	Nivel o rango del jugador.

Ambos servicios utilizan una combinación de estos tipos. Sin embargo, el **streaming de videojuegos maneja volúmenes más altos de datos en tiempo real** y requiere mayor sincronización entre usuario, red y servidor.

3. Técnicas de limpieza de datos

El proceso de limpieza de datos es crucial para garantizar la calidad del análisis posterior. En los servicios de streaming se aplican diversas técnicas, entre ellas:

1. **Eliminación de valores nulos o faltantes:** los sistemas pueden registrar errores de carga o interrupciones de conexión; estos registros deben completarse o eliminarse para evitar distorsiones.
2. **Detección y corrección de duplicados:** especialmente en registros de usuarios o transacciones recurrentes, que pueden repetirse por fallos en la sincronización.
3. **Estandarización de formatos:** convertir fechas, horas y nombres de archivo a formatos uniformes (ISO 8601, UTF-8, etc.) es esencial para la integración entre plataformas (Google Cloud, 2024).

4. **Identificación de valores atípicos:** detectar comportamientos anómalos, como picos de visualización inusuales o actividad de bots.
5. **Validación de integridad:** verificación cruzada entre bases de datos de usuarios, contenido y servidores para asegurar coherencia.
6. **Anonimización de datos personales:** medida obligatoria por regulaciones de privacidad como el Reglamento General de Protección de Datos (GDPR) o la Ley Federal de Protección de Datos Personales en México.

En la práctica, plataformas como Netflix utilizan procesos automáticos de *data cleaning pipelines* integrados con aprendizaje automático para corregir y validar grandes volúmenes de información (Netflix Tech Blog, 2024).

Figura 1. Comparación general de fuentes y tipos de datos en streaming audiovisual y de videojuegos

Categoría	Streaming audiovisual	Streaming de videojuegos
Fuentes de datos		
Datos generados por humanos	Registro de usuarios, preferencias de contenido, calificaciones, listas personalizadas.	Perfiles de jugadores, acciones dentro del juego, configuración de controles.
Datos de transacciones	Suscripciones, métodos de pago, historial de consumo.	Compras dentro del juego, suscripciones, tiempo de juego.
Datos máquina a máquina	Monitoreo de rendimiento del servidor y dispositivos de reproducción.	Métricas de conexión, latencia, rendimiento en tiempo real entre consola y servidor.
Datos web y redes sociales	Comentarios, reseñas, menciones en redes sociales.	Opiniones en foros de jugadores, transmisiones y chats.
Datos biométricos	No aplican de forma común.	Uso experimental en experiencias inmersivas (ritmo cardíaco, movimientos).
Tipos de datos		
Cuantitativos	Número de reproducciones, duración promedio de visualización.	FPS, tiempo de sesión, puntuaciones.
Cualitativos	Opiniones, géneros preferidos, etiquetas de contenido.	Experiencia de usuario, reseñas, nivel de satisfacción.
Estructurados	Bases de datos de usuarios y catálogos de contenido.	Tablas de estadísticas de jugadores y rendimiento.
No estructurados	Archivos multimedia, subtítulos, descripciones.	Capturas, videos de partidas, comentarios de voz.
Nominales	Categorías de contenido (película, serie, documental).	Tipos de juegos (acción, estrategia, deportes).

Ordinales	Clasificaciones por edad, popularidad o calificación.	Rangos de jugador, niveles de habilidad.

Fuente: Elaboración propia con base en Microsoft Azure (2024) y Netflix Tech Blog (2024).

4. Fundamentación

La limpieza de datos se encuentra en el núcleo del proceso de *data engineering* moderno. Según Google Cloud (2024), la mayoría de los errores en los análisis de datos corporativos provienen de conjuntos de datos sin depurar. En servicios de streaming, donde la experiencia del usuario depende de la precisión del sistema de recomendación, **la calidad de los datos determina directamente la fidelidad del cliente**.

En el caso de los videojuegos, la estabilidad de la conexión y el rendimiento dependen de datos de red limpios y consistentes, lo que exige una gestión continua y automatizada. Ambos sectores demuestran que **la limpieza no es una etapa opcional, sino un requisito estructural** del procesamiento de datos en la economía digital.

Conclusiones

La comparación entre servicios de streaming de contenido audiovisual y de videojuegos demuestra que, aunque ambos dependen de grandes volúmenes de información, difieren en la naturaleza y velocidad de generación de datos.

Los servicios audiovisuales manejan datos predominantemente de consumo y preferencias, mientras que los de videojuegos gestionan datos de interacción en tiempo real. En ambos casos, la aplicación de técnicas rigurosas de limpieza —como eliminación de duplicados, estandarización y detección de valores atípicos— resulta indispensable para garantizar la calidad analítica y la seguridad del usuario.

En conclusión, la **limpieza de datos constituye una práctica esencial para optimizar la toma de decisiones, personalizar servicios y asegurar la integridad de las plataformas digitales.**

Referencias

Google Cloud. (2024). Data cleaning best practices for scalable analytics. https://cloud.google.com/

Microsoft Azure. (2024). Big data management and governance in streaming systems. https://azure.microsoft.com/

Netflix Tech Blog. (2024). Building data pipelines for personalization. https://netflixtechblog.com/

NVIDIA Research. (2024). Real-time data streaming for gaming analytics. https://research.nvidia.com/

Statista. (2024). Global streaming market data and usage trends. https://www.statista.com/

DataCamp. (2023). Introduction to data cleaning techniques. https://www.datacamp.com/