Proyecto 02 Python

Máster Universitario en Ciencia de Datos
Extracción, Transformación y Carga

CUNEF Universidad

Estudiantes:
Jazmín Fernández Ramírez
Jacqueline Fernández Ramírez
Nina M. Odoux

Docente:
Víctor Ramos Fuentes

23 de diciembre, 2024

Fase I: Elección de datos

➔ **Descripción del Formula 1 Official Data (1950-2022):**

La Fórmula 1 es el nivel más alto del deporte de motor internacional para monoplazas de ruedas abiertas aprobado por la Fédération Internationale de l'Automobile (FIA). Desde su temporada inaugural en 1950, el Campeonato Mundial de Pilotos se ha convertido en una de las series de automovilismo más prestigiosas del mundo. La palabra "Fórmula" se refiere a las reglas especiales que deben seguir todos los autos participantes. El conjunto de datos **Formula 1 Official Data** contiene datos detallados de todas las carreras, sprints, calificaciones, sesiones de práctica, pilotos, equipos y vueltas de todos los campeonatos desde 1950 hasta 2022. De esta forma, constituye una herramienta invaluable para el análisis histórico y comparativo de las carreras de Fórmula 1, proporcionando una descripción completa y detallada de las décadas de carreras y eventos clave en este apasionante deporte.

➔ **Componentes seleccionados del dataset:**

Este proyecto de ETL se enfocará en tres componentes clave de este dataset:

- **driver_standings.csv:** Contiene la clasificación de los pilotos para todas las carreras de Fórmula 1 desde 1950 hasta 2022. Incluye las posiciones finales de los pilotos en cada temporada, permitiendo analizar el rendimiento histórico de los conductores y compararlos a lo largo de las décadas.

- **race_details.csv:** Incluye los resultados detallados de las carreras de Fórmula 1 para todas las pistas desde 1950 hasta 2022. Proporciona información sobre cada carrera, como las posiciones finales y tiempos de vuelta, permitiendo entender el desarrollo y los resultados de cada evento.

- **driver_details.csv:** Ofrece detalles sobre los pilotos que han competido en todas las carreras desde 1950 hasta 2022. Incluye datos personales de los

conductores, estadísticas de carrera y demás información relevante, que añade contexto al análisis del rendimiento de los pilotos a lo largo del tiempo.

➔ **Relevancia del Dataset para el proceso de Extracción, Transformación y Carga:**

El dataset **"Formula 1 Official Data (1950-2022)"** constituye una fuente integral y precisa que recopila información detallada sobre el campeonato mundial de Fórmula 1, abarcando más de siete décadas de historia. Su estructura, que incluye variables clave como fechas, circuitos, ubicaciones y resultados de carreras, ofrece una base idónea y óptima para desarrollar un proceso de ETL (Extracción, Transformación y Carga) robusto y eficiente, mediante el uso del lenguaje de programación Python.

Así pues, es importante añadir que la aplicación de este procedimiento con Python es especialmente significativa debido a la flexibilidad y potencial de sus **librerías especializadas** como Pandas, NumPy, entre otros. Estas facilitan la extracción de grandes volúmenes de información, su tratamiento a través de procesos de limpieza, normalización y mejora, y al final, su carga en bases de datos apropiadas para su análisis. Esta técnica no solo asegura la calidad y coherencia de la información, sino que también simplifica su uso analítico mediante herramientas relacionales y visualizaciones sofisticadas. En este sentido, al integrar la capacidad de Python, se pueden producir informes detallados que resaltan tendencias pasadas, patrones de rendimiento de pilotos y equipos, así como la evolución del deporte en diferentes escenarios temporales. Así, el dataset se convierte en un insumo clave para proyectos de investigación académica, permitiendo un análisis efectivo y metodológicamente riguroso del **automovilismo deportivo.**

➔ **Aplicaciones y potencial análisis:**

Finalmente, es preciso mencionar que, este dataset proporciona una base sólida para:

- Evaluar el rendimiento histórico de los pilotos a lo largo de las décadas.

- Estudiar la evolución de las carreras de Fórmula 1, en torno al cambio de las estrategias y las dinámicas de las competiciones.
- Analizar el impacto de los cambios en las reglas, y su incidencia en los resultados de las carreras y el rendimiento de los equipos.
- Evaluar la influencia de los equipos y los constructores en el éxito de los pilotos y el desarrollo del deporte.
- Elaborar estudios comparativos entre pilotos.

Fase II: Documentación

En este apartado, se expondrán los objetivos del proyecto, información adicional del dataset seleccionado, las características de los datos, la calidad de los datos, la limpieza de los datos y, por último, los problemas y las siguientes acciones a considerar.

1. Objetivos del proyecto:

- **Propósito del proyecto:**

Desarrollar un sistema de análisis histórico del campeonato mundial de Fórmula 1 mediante un proceso de ETL (Extract, Transform, Load), utilizando Python como lenguaje de programación. El propósito principal es extraer, transformar y cargar datos del dataset *Formula 1 Official Data (1950-2022)*, consolidando y estandarizando la información para crear una base de datos estructurada y precisa, que permita realizar análisis detallados sobre el desempeño de pilotos, carreras y circuitos a lo largo del tiempo, identificando tendencias históricas, patrones de rendimiento y factores clave que han influido en la evolución del automovilismo deportivo.

- **Objetivos específicos:**

1. Seleccionar información relevante de las fuentes de datos del dataset *Formula 1 Official Data* (driver_standings.csv, race_details.csv y driver_details.csv), con el fin de obtener una visión integral y estructurada del desempeño de los pilotos, los resultados de las carreras, y detalles valiosos asociados con el campeonato mundial de Fórmula 1.
2. Extraer los datos contenidos en estos archivos como primer paso del proceso de ETL, asegurando la recopilación completa y precisa de la información pertinente para su posterior transformación y análisis.

3. Identificar y corregir errores, duplicados y datos inconsistentes presentes en los archivos, aplicando técnicas de limpieza y normalización que permitan garantizar la calidad, integridad y fiabilidad de los datos.
4. Diseñar un modelo dimensional que facilite el análisis eficiente y comparativo del desempeño de pilotos, equipos y circuitos, posibilitando segmentar la información a través de tablas de hechos y de dimensiones.
5. Transformar los datos en formatos estandarizados y normalizados, para facilitar su integración y comparación eficiente durante las fases de análisis.
6. Cargar los datos procesados en un repositorio centralizado, asegurando un almacenamiento organizado y accesible, que sirva como base para la realización de análisis avanzados y creación de informes.

- **Justificación de la importancia de realizar un proceso ETL:**

Los procesos de ETL (Extract, Transform, Load) son fundamentales para gestionar eficazmente grandes cantidades de datos, como los contenidos en el dataset *Formula 1 Official Data (1950-2022).* Este proceso garantiza una integración consistente de datos de diferentes fuentes, como carreras, clasificaciones e información de los corredores, y los transforma en un formato de análisis estandarizado y práctico para el análisis. La coherencia y precisión de estas integraciones de datos, son fundamentales para obtener información confiable y útil para futuras investigaciones y estudios posteriores.

En cuanto a los procedimientos específicos, la extracción de datos de archivos como race_details.csv, driver_standings.csv y driver_details.csv facilita la recolección de información dispersa. De este modo, se recopila toda la información relevante en un solo lugar, lo cual es esencial para cualquier análisis integral. Por su parte, el proceso de transformación implica limpiar, normalizar y estandarizar los datos para garantizar la precisión y consistencia de los mismos. Este proceso permite preparar los datos para su análisis y visualización manteniendo su integridad y calidad.

Finalmente, en la etapa de carga, los datos transformados se integran en un depósito de datos centralizado. Esto facilita el acceso a datos fiables y optimizados

para realizar análisis detallados, generar informes precisos y tomar decisiones informadas basadas en datos. Cabe añadir que, un proceso de ETL bien organizado es esencial para asegurar que la información utilizada en el análisis sea de excelente calidad, lo cual es primordial para obtener percepciones valiosas y exactas sobre la historia, rendimiento y evolución en la Fórmula 1.

2. Dataset:

- **Enlace al dataset desde la web de Kaggle, plataforma recomendada por el docente:**

https://www.kaggle.com/datasets/debashish311601/formula-1-official-data-19502022/data?select=race_details.csv

- **Descripción de las fuentes de datos del Formula 1 Official Data (1950-2022*)* y Data Catalog con las principales características de cada campo:**

driver_standings.csv

Nombre de la columna	Descripción	Tipo de dato
Pos	Posición del piloto en el campeonato.	int
Driver	Nombre del piloto.	str
Nationality	Nacionalidad del piloto.	str
Car	Nombre del coche o escudería.	str
PTS	Puntos obtenidos en la temporada.	float
DriverCode	Código abreviado del piloto.	str
Year	Año de la temporada.	int

race_details.csv

Nombre de la columna	Descripción	Tipo de dato
Pos	Posición del piloto en la carrera.	int
No	Número asignado al piloto en la carrera.	int
Driver	Nombre del piloto.	str
Car	Nombre del coche o escudería.	str
Laps	Número de vueltas completadas en la carrera.	int
Time/Retired	Tiempo registrado o estado (retirado)-	str
PTS	Puntos obtenidos en la carrera.	float
Year	Año en que se realizó la carrera.	int
Grand Prix	Nombre del Gran Premio.	str
Detail	Información adicional de la carrera.	str
DriverCode	Código abreviado del piloto.	str

driver_details.csv

Nombre de la columna	Descripción	Tipo de dato (Python)
Car	Nombre del coche o escudería.	str

Date	Fecha de la carrera (día, mes y año).	str
Driver	Nombre del piloto.	str
Grand Prix	Nombre del Gran Premio.	str
PTS	Puntos obtenidos en la carrera.	float
Race Position	Posición del piloto en la carrera.	int
Year	Año de la temporada.	int

- **Frecuencia de actualización de los datos:**

La frecuencia de actualización del dataset ***Formula 1 Official Data (1950-2022)***, aunque se indica como semanal, presenta una última actualización registrada hace 2 años. Aunque los datos no están actualizados en tiempo real, su cobertura histórica (desde 1950 hasta 2022) proporciona un **horizonte temporal amplio y consolidado** de más de siete décadas, lo cual resulta pertinente para cumplir con los objetivos del análisis planteado en este proyecto. Por lo tanto, la frecuencia de actualización, aunque limitada, **no compromete la validez del análisis histórico**, que constituye el núcleo de este estudio. Esta cobertura proporciona una base de datos robusta y estable, adecuada para realizar un análisis riguroso y consistente del rendimiento de pilotos, equipos y circuitos durante más de 70 años, aspectos determinantes en la evolución del campeonato mundial de Fórmula 1.

3. Características de los datos:

- **Descripción de los tipos de datos manejados:**

El **Formula 1 Official Data (1950-2022)** se caracteriza por ser un conjunto de datos organizados con gran precisión y consistencia. Este dataset está integrado, entre otros, por tres archivos CSV, driver_standings.csv, race_details.csv, y

driver_details.csv, todos presentados en un formato tabular en el que cada fila corresponde a un registro único y cada columna a un atributo específico. Este tipo de disposición facilita el entendimiento y manejo de la información, lo que resulta imprescindible para cualquier estudio en el contexto de la Fórmula 1, donde la precisión y la claridad son fundamentales.

Uno de los elementos clave de estos archivos CSV es la uniformidad en su organización. Cada archivo sigue un modelo preestablecido con nombres de columnas bien definidos y permanentes. Por ejemplo, en ***driver_standings.csv***, las columnas como **Pos** y **PTS** almacenan la posición del piloto en el campeonato y los puntos obtenidos en la temporada, respectivamente. En ***race_details.csv***, se encuentran columnas como **Laps** (número de vueltas completadas en la carrera) y **Grand Prix** (nombre del Gran Premio), mientras que ***driver_details.csv*** incluye **Date** (fecha de la carrera) y **Race Position** (posición final del piloto en la carrera). Esta uniformidad no sólo simplifica el proceso de análisis mediante lenguajes de programación como Python, sino que también asegura que todas las filas de los archivos conservan un formato constante, lo que fortalece la fiabilidad y robustez de los datos.

Además, los datos en estos archivos son concretos y precisos. Cada valor en una columna corresponde a un tipo de dato claramente definido. Por ejemplo, los identificadores de posiciones (**Pos**) y años (**Year**) están representados como enteros (**int**), los nombres de pilotos (**Driver**) y escuderías (**Car**) son cadenas de texto (**str**), y los puntos obtenidos en carreras (**PTS**) están representados como números decimales (**float**). También se utilizan fechas en formato de cadena de texto (**str**) para representar la fecha de cada carrera, especialmente en **driver_details.csv.** Esta clasificación precisa de los tipos de datos facilita la validación de la información y la ejecución de análisis sin ambigüedades, lo que resulta fundamental para la exactitud y utilidad de los resultados obtenidos.

En particular, **driver_standings.csv** proporciona una visión global del rendimiento de los pilotos a lo largo de cada temporada, destacando la posición final de los mismos y su rendimiento acumulado en puntos. Por su lado, **race_details.csv** ofrece un análisis detallado de cada Gran Prix, permitiendo una

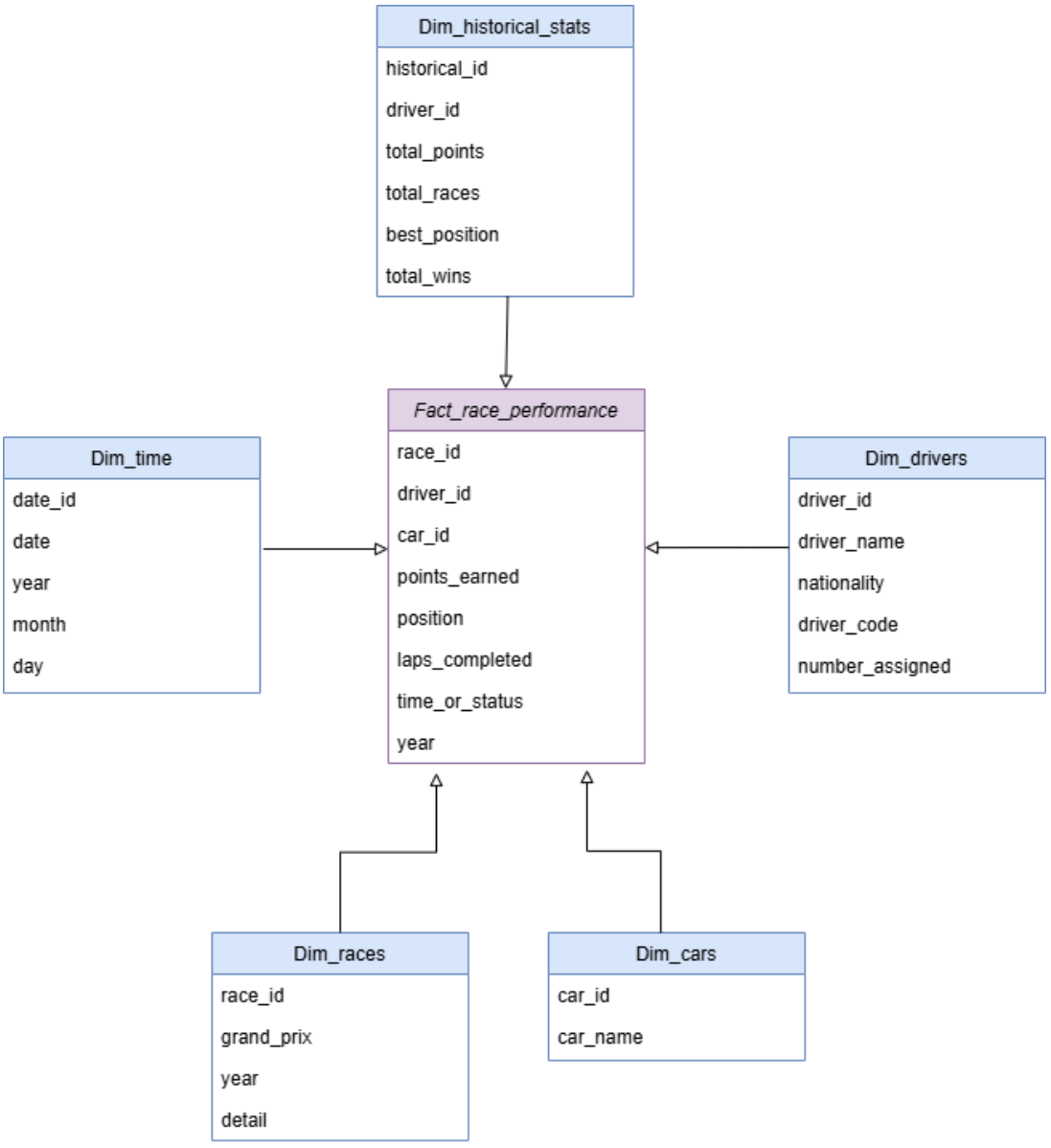
evaluación más profunda del rendimiento individual de los pilotos en cada carrera, con información relevante como las vueltas completadas y el estado final (tiempo registrado o retirada). Por último, **driver_details.csv** se encarga de registrar el desempeño de los pilotos en cada carrera específica, detallando fechas y posiciones, lo que permite realizar un análisis exhaustivo del comportamiento de cada piloto a lo largo de la temporada.

En síntesis, la organización precisa y la uniformidad de los atributos en el Dataset Oficial de Fórmula 1 no solo respaldan su **clasificación como datos estructurados**, sino que también permiten una integración eficiente en sistemas de bases de datos relacionales. Este tipo de estructuración facilita el proceso de ETL (Extracción, Transformación y Carga), proporcionando una base sólida para realizar análisis complejos sobre el **rendimiento histórico de los pilotos** y **las escuderías en el campeonato.** En este sentido, la clara organización de los datos asegura que estos puedan ser utilizados de manera efectiva para obtener exploraciones significativas, lo que los convierte en un recurso valioso para estudios y análisis dentro del contexto de la Fórmula 1.

- **Definición del modelo de datos:**

El modelo dimensional diseñado para el análisis histórico del campeonato mundial de fórmula 1 adopta una ***estructura en estrella,*** optimizada para simplificar consultas y análisis relacionados con el rendimiento de pilotos, coches y carreras a lo largo del tiempo. En el centro del modelo, la tabla de hechos Fact_race_performance centraliza métricas clave como los puntos obtenidos, las posiciones alcanzadas, el número de vueltas completadas y el estado final de los pilotos en cada carrera. Alrededor de esta tabla se estructuran las dimensiones que enriquecen el análisis, como Dim_drivers, Dim_races, Dim_cars, Dim_time y Dim_historical_stats. Cada una de estas dimensiones aportan atributos específicos que proporcionan contexto detallado, como el nombre y la nacionalidad de los pilotos, la información de cada Gran Prix, las características de los coches y las estadísticas históricas acumuladas.

Este diseño permite un equilibrio entre la simplicidad y la profundidad analítica, garantizando que los datos sean fácilmente accesibles para identificar patrones de desempeño, tendencias por temporada o comparaciones entre escuderías. La inclusión de dimensiones como Dim_historical_stats proporciona un valor añadido, permitiendo el análisis acumulado y a largo plazo del rendimiento de los pilotos. Además, este enfoque respeta las prácticas del diseño dimensional y asegura un uso efectivo para futuros análisis, proporcionando así una herramienta valiosa para entender mejor la evolución del Campeonato Mundial de Fórmula 1 y tomar decisiones informadas basadas en datos históricos.



Fuente: Elaboración propia.

- **Sistema de gestión de grandes volúmenes de datos: Data Warehouse**

El presente modelo corresponde a la arquitectura de un **Data Warehouse**, diseñado para almacenar grandes volúmenes de datos estructurados y organizados de manera eficiente. Utilizando un modelo dimensional, como el esquema en estrella propuesto, esta estructura permite realizar consultas rápidas y análisis detallados. Esto es fundamental para desarrollar un sistema de análisis histórico del campeonato de Fórmula 1 mediante un proceso de ETL.

El proyecto tiene como objetivo principal extraer, transformar y cargar datos de diferentes fuentes del Formula 1 Official Data (1950-2022), creando una base de datos integrada, precisa y estandarizada. Esta base de datos centralizada permitirá realizar **análisis detallados** del rendimiento de pilotos, escuderías y Grandes Premios, proporcionando una **herramienta clave** para el estudio de tendencias, patrones de desempeño y evolución histórica del campeonato.

En este contexto, la arquitectura de un Data Warehouse proporciona un entorno confiable, escalable y optimizado para el análisis de datos. Así pues, al establecer este sistema, se posibilita un acceso más detallado y relevante a los datos históricos del campeonato, respaldando no solo el estudio minucioso de sucesos individuales, sino también la **comprensión estratégica** de la evolución de uno de los deportes automovilísticos más importantes a nivel mundial. En conclusión, este proyecto de Data Warehouse combina la precisión y consistencia de un proceso ETL apropiadamente diseñado con la capacidad analítica de un modelo dimensional, creando una **solución robusta y eficiente** para la gestión y el análisis de datos históricos del campeonato mundial de Fórmula 1.

4. Calidad de los datos:

En esta sección se presenta la evaluación de la calidad de los datos del Formula 1 Official Data (1950-2022), abarcando tres DataFrames principales, driver_details, driver_standings, y race_details. Así pues, se aplicarán diversas métricas de calidad a las columnas de cada DataFrame, para obtener un porcentaje de calidad por columna y por DataFrame. Finalmente, se calculará un porcentaje global de calidad de los tres DataFrames y, también, se presentará una ***calificación final*** para el Formula 1 Official Data, con el propósito de brindar una visión integral y precisa de la calidad de los datos disponibles.

- **Resumen de las métricas para cada uno de los DataFrames:**

Driver_details

	Precisión	Linaje	Semántica	Estructura	Completitud	Consistencia	Moneda	Puntualidad	Razonabilidad	Identificabilidad
Car	100.000000	100.0	100.0	100.000000	99.964671	100.000000	100.0	100.0	100.000000	100.0
Date	100.000000	100.0	100.0	13.248208	100.000000	100.000000	0.0	0.0	100.000000	100.0
Driver	100.000000	100.0	100.0	100.000000	100.000000	100.000000	100.0	100.0	100.000000	0.0
Grand Prix	100.000000	100.0	100.0	100.000000	100.000000	100.000000	100.0	100.0	100.000000	100.0
PTS	99.949531	100.0	100.0	100.000000	99.949531	100.000000	100.0	100.0	99.949531	100.0
Race Position	100.000000	100.0	100.0	100.000000	99.919249	100.000000	100.0	100.0	100.000000	100.0
Year	100.000000	100.0	100.0	100.000000	100.000000	13.248208	100.0	100.0	100.000000	100.0

El análisis de las métricas de calidad de "Driver_details" revela que, en general, los datos mantienen una alta calidad, con la mayoría de los atributos alcanzando valores cercanos al 100% en casi todas las métricas. Sin embargo, se identifican áreas de mejora, como en las métricas de "Moneda" y "Puntualidad", que muestran valores de 0 en el atributo de "Date". Además, en el

atributo de "Driver", se presenta este valor de 0 en la métrica de identificabilidad, lo que sugiere que los datos pueden no estar actualizados, ser oportunos o suficientemente únicos para su análisis efectivo. Por su parte, la "Estructura" del atributo "Date" presenta un valor significativamente bajo (13.25%), indicando posibles inconsistencias en el formato de las fechas. Por último, la "Completitud" del atributo "Car" y la "Razonabilidad" del atributo "PTS" también son ligeramente inferiores al 100%, señalando la presencia de datos faltantes o valores atípicos.

Driver_standings

	Precisión	Linaje	Semántica	Estructura	Completitud	Consistencia	Moneda	Puntualidad	Razonabilidad	Identificabilidad
Pos	100.0	100.0	100.0	100.0	100.000000	100.0	100.0	100.0	100.0	100.0
Driver	100.0	100.0	100.0	100.0	100.000000	100.0	100.0	100.0	100.0	0.0
Nationality	100.0	100.0	100.0	100.0	100.000000	100.0	100.0	100.0	100.0	100.0
Car	100.0	100.0	100.0	100.0	99.320148	100.0	100.0	100.0	100.0	100.0
PTS	100.0	100.0	100.0	100.0	100.000000	100.0	100.0	100.0	100.0	100.0
DriverCode	100.0	100.0	100.0	100.0	100.000000	100.0	100.0	100.0	100.0	0.0
Year	100.0	100.0	100.0	100.0	100.000000	100.0	100.0	100.0	100.0	100.0

En el análisis de “Driver_standings” se muestra un alto nivel de calidad general en la mayoría de los atributos, con valores de 100% en métricas como precisión, linaje, semántica, estructura, consistencia y razonabilidad, lo cual indica una excelente integridad y coherencia de los datos. No obstante, la métrica de “Completitud” para el atributo "Car" es de 99.32%, sugiriendo la presencia de algunos datos faltantes, y la “Identificabilidad” de los atributos "Driver" y "DriverCode" muestra un notable valor de 0%, lo que implica que estos datos no son únicos y que podrían afectar la capacidad para identificar de manera precisa a los

pilotos. Por tanto, aunque la calidad general es alta, estos puntos específicos requieren de atención para asegurar una base de datos fiable y adecuada para futuros análisis.

Race_details

	Precisión	Linaje	Semántica	Estructura	Completitud	Consistencia	Moneda	Puntualidad	Razonabilidad	Identificabilidad
Pos	100.0	100.0	100.0	100.000000	100.000000	100.0	100.0	100.0	100.0	100.0
No	100.0	100.0	100.0	100.000000	100.000000	100.0	100.0	100.0	100.0	100.0
Driver	100.0	100.0	100.0	100.000000	100.000000	100.0	100.0	100.0	100.0	0.0
Car	100.0	100.0	100.0	100.000000	99.891567	100.0	100.0	100.0	100.0	100.0
Laps	100.0	100.0	100.0	100.000000	99.136709	100.0	100.0	100.0	100.0	100.0
Time/Retired	100.0	100.0	100.0	100.000000	99.966636	100.0	100.0	100.0	100.0	100.0
PTS	100.0	100.0	100.0	100.000000	100.000000	100.0	100.0	100.0	100.0	100.0
Year	100.0	100.0	100.0	100.000000	100.000000	100.0	100.0	100.0	100.0	100.0
Grand Prix	100.0	100.0	100.0	100.000000	100.000000	100.0	100.0	100.0	100.0	100.0
Detail	100.0	100.0	100.0	100.000000	100.000000	100.0	100.0	100.0	100.0	100.0
DriverCode	100.0	100.0	100.0	99.962466	100.000000	100.0	100.0	100.0	100.0	0.0

En el análisis de "Race_details" se muestra una alta calidad general, con la mayoría de los atributos alcanzando el 100% en precisión, linaje, semántica, moneda, puntualidad y consistencia, lo que evidencia la actualización, relevancia y coherencia de los datos. Sin embargo, se identifican áreas de mejora en la completitud del atributo "Car" con un 99.89%, así como en los atributos de “Laps” y “Time/Retired”. Esto indica la presencia de algunos datos faltantes que deben ser tratados apropiadamente. Además, la baja identificabilidad de los atributos "Driver" y "DriverCode" (0%) destaca la necesidad de mejorar la unicidad de los identificadores para un análisis más preciso. De este modo, aunque la calidad general de los datos es alta, es fundamental abordar estos puntos específicos para garantizar datos confiables que posibiliten análisis detallados y pertinentes.

- **Resumen global de las métricas:**

Driver_details

Precisión	99.992790
Linaje	100.000000
Semántica	100.000000
Estructura	87.606887
Completitud	99.976207
Consistencia	87.606887
Moneda	85.714286
Puntualidad	100.000000
Razonabilidad	99.992790
Identificabilidad	100.000000

Media general para driver details: 96.09

Driver_standings

Precisión	100.000000
Linaje	100.000000
Semántica	100.000000
Estructura	100.000000
Completitud	99.902878
Consistencia	100.000000
Moneda	100.000000
Puntualidad	100.000000
Razonabilidad	100.000000
Identificabilidad	100.000000

Media general para driver_standings: 99.99

Race_details

Precisión	100.000000
Linaje	100.000000
Semántica	100.000000
Estructura	99.996588
Completitud	99.908628
Consistencia	100.000000
Moneda	100.000000
Puntualidad	100.000000
Razonabilidad	100.000000
Identificabilidad	100.000000

Media general para race_details: 99.99

Al analizar las tres tablas de métricas presentadas, se observa un rendimiento notablemente alto en todos los componentes evaluados, aunque con algunas variaciones significativas. La tabla de "driver_details" muestra las puntuaciones más bajas del conjunto, con una media de **96.09,** afectada principalmente por los valores reducidos en las categorías de **moneda** (85.71), **estructura** (87.60) y **consistencia** (87.60). En contraste, tanto "driver_standings" como "race_details" exhiben un desempeño sobresaliente, alcanzando medias generales de 99.99, donde prácticamente todos sus indicadores se acercan o alcanzan la perfección (100.00). Es particularmente interesante notar que la **estructura y la completitud** son los aspectos que más frecuentemente presentan valores por debajo del 100% en las tres tablas, lo cual sugiere que estas áreas podrían beneficiarse de una **revisión más detallada**, para alcanzar el nivel de calidad observado en las demás métricas.

Métricas generales para todos los datasets combinados:

```
Métricas generales para todos los datasets combinados:
Precisión             99.997981
Linaje               100.000000
Semántica            100.000000
Estructura            96.528427
Completitud           99.925940
Consistencia          96.529928
Moneda                96.000000
Puntualidad          100.000000
Razonabilidad         99.997981
Identificabilidad    100.000000
dtype: float64

Calificación general de calidad de datos: 98.90
```

El análisis de las métricas de calidad de los datasets "Driver_details", "Driver_standings" y "Race_details" indica un alto rendimiento general, con la mayoría de los componentes evaluados alcanzando valores cercanos al 100%. Sin embargo, se observan ligeras variaciones en áreas específicas como la estructura, completitud, consistencia y moneda. Por su parte, la calificación general de la calidad de los datos del Formula 1 Official Data es de **98,90.** Esto indica que, a pesar de algunas oportunidades de mejora en ciertos atributos, los datos presentan una alta precisión y consistencia, que asegura su fiabilidad y utilidad para próximos análisis en el ámbito de la Fórmula 1.

- **Problemas detectados:**

El análisis de las métricas de calidad de los datasets "Driver_details", "Driver_standings" y "Race_details" destaca un rendimiento general muy elevado, aunque existen algunas áreas específicas que requieren atención para asegurar datos consistentes y adecuados. Las principales áreas de mejora incluyen las métricas de "Moneda" y "Puntualidad" en el atributo "Date", y la "Identificabilidad" en el atributo "Driver", todas con valores del 0%, lo que indica que los datos podrían no estar actualizados, ser puntuales o suficientemente únicos. Además, se observaron inconsistencias en la "Estructura" del atributo "Date" en "Driver_details" con un valor de 13.25%, así como datos faltantes en la "Completitud" del atributo "Car" y la "Razonabilidad" del atributo "PTS". En "Driver_standings", la "Identificabilidad" de los atributos "Driver" y "DriverCode" también presenta un valor de 0%, afectando la precisión en la identificación de los pilotos.

Métricas empleadas

Se utilizaron **todas las métricas disponibles** para evaluar la calidad de los datos, incluyendo precisión, linaje, semántica, estructura, completitud, consistencia, moneda, puntualidad, razonabilidad e identificabilidad. La aplicación de estas métricas es crucial para asegurar que los datos sean precisos, completos y coherentes, lo cual es fundamental para cualquier análisis posterior. La evaluación integral de todas las métricas permite **identificar áreas específicas** que requieren mejoras, facilitando el proceso de limpieza de datos. En este caso, fue posible aplicar todas las métricas, lo que proporciona una **visión exhaustiva de la calidad de los datos** y asegura que estén en condiciones óptimas para su uso en futuros análisis y estudios.

5. Limpieza de los datos:

En el marco del proceso de ETL (Extracción, Transformación y Carga), se llevó a cabo la limpieza de los conjuntos de datos **driver_details.csv, driver_standings.csv** y **race_details.csv.** Este proceso se diseñó para preparar la información con un nivel óptimo de calidad, asegurando su coherencia, integridad y

utilidad para análisis posteriores. A continuación, se describe en detalle el proceso aplicado, justificando cada paso implementado en el código.

Renombrado y estandarización de columnas

Se normalizaron los nombres de las columnas de cada tabla. Esto se implementó de la siguiente manera,

- Se transformaron los nombres de las columnas a formato **snake_case**, eliminando espacios y convirtiendo los nombres a minúsculas.
- Ejemplo: Grand Prix se convirtió en **grand_prix**, y Race Position en **race_position**.
- Esta estandarización mejora la compatibilidad con herramientas analíticas, y reduce errores en consultas o transformaciones posteriores.

Eliminación de duplicados

Se eliminaron registros **duplicados identificados**, mejorando la precisión para los análisis. Esto se alcanzó mediante claves compuestas,

- En driver_details.csv, se consideraron las columnas driver, date y grand_prix.
- En driver_standings.csv, se utilizaron las columnas driver, year y car.
- En race_details.csv, se incluyeron driver, year y grand_prix. Esto asegura que los datos sean únicos y representativos.

Imputación de valores nulos

Los valores nulos fueron tratados de acuerdo con la naturaleza de cada columna,

- En campos categóricos como car, se reemplazaron los valores nulos con el marcador **"Unknown"**.

- En campos numéricos como race_position y pos, se utilizó un marcador especial **-1** para indicar **"No Disponible"**. Esto mantiene la integridad del esquema y permite distinguir entre valores nulos imputados y datos reales.
- En la columna pts, los valores nulos fueron reemplazados con **0** como valor por defecto, reflejando la ausencia de puntos en un evento o temporada.

Estandarización de formatos

Se estandarizaron los formatos de las fechas en driver_details.csv al formato **YYYY-MM-DD,** asegurando uniformidad y compatibilidad con sistemas analíticos. Además, se realizaron transformaciones en campos de texto para garantizar la consistencia, incluyendo la homogenización de los valores en columnas como car para mantener un formato uniforme.

Limpieza de valores atípicos

Se validaron y transformaron valores no estándar para garantizar consistencia y facilitar el análisis. En las columnas de posiciones (race_position, pos), valores como 'DNF', 'Retired', y 'NC' fueron convertidos al marcador -1, indicando que el piloto no finalizó o no clasificó (los valores como **"DNF"** y **"NC"** no son posiciones válidas numéricamente, ya que el propósito de la columna es representar el lugar final del piloto en una carrera). En otras columnas, como en time/retired, los valores nulos fueron reemplazados con 'Unknown'. Por su lado, los valores como **"DNF"** y **"Retired"**, contienen información semántica crucial sobre por qué el piloto no terminó, lo cual refuerza el objetivo de la columna para describir el estado o el tiempo final del piloto en la carrera, por esta razón se mantuvieron estos términos. Por último, en laps, los valores faltantes se imputaron con 0.

Verificación y validación de cambios

Cada tabla fue sometida a un análisis post-limpieza, verificando el número de **registros iniciales y finales**, así como los tipos de datos. Este paso permitió asegurar que no existían duplicados residuales, los valores se encontraban dentro

de los rangos esperados y la estructura era consistente con los requisitos del análisis.

Resultados

- En driver_details.csv, se eliminaron **6 duplicados**, y las posiciones no disponibles fueron marcadas con -1.
- En driver_standings.csv, **no se identificaron duplicados** adicionales, y 1 posición nula fue tratada con un marcador.
- En race_details.csv, **se eliminaron 34 duplicados**, y se imputaron valores nulos en pos (valor de -1), y time/retired (unknown).

De esta manera, este proceso de limpieza logró estandarizar, corregir y enriquecer los datos en las tres tablas. A través de la eliminación de inconsistencias, la imputación de valores nulos y la normalización de formatos, se logró garantizar que los datos cumplan con altos estándares de calidad y coherencia requeridos. Finalmente estas transformaciones no solo aseguran que la información esté preparada para su análisis posterior, sino que también facilitan su integración eficiente en sistemas analíticos, permitiendo obtener resultados confiables y maximizar su valor en contextos de toma de decisiones.

6. Problemas y próximos pasos:

● Descripción de los principales desafíos encontrados durante las fases de extracción y transformación:

En las fases de extracción y transformación de datos, se evidenciaron desafíos tales como,

Extracción

- **Calidad de los datos de origen:** datos incompletos o faltantes (por ejemplo, en columnas como Car, Laps y Time/Retired) y formato de datos

inconsistente (el caso de la columna “Date” que presenta inconsistencias en el formato de las fechas ya que combina números y texto: ***21 May 1950***).

Transformación

- **Inconsistencia en los nombres de las columnas**: Los nombres de las columnas variaban en formato, lo que requería normalización a snake_case para evitar errores en futuras consultas y transformaciones.
- **Presencia de registros duplicados**: Se detectaron y eliminaron registros duplicados utilizando claves compuestas específicas para cada tabla, garantizando la unicidad de los datos.
- **Valores nulos en campos clave**: Se encontraron valores nulos en varias columnas, como 'car' y 'race_position', los cuales fueron imputados con valores representativos como "Unknown" o -1 para mantener la integridad del esquema de datos.
- **Formato de fechas no estandarizado**: Las fechas en driver_details no seguían un formato uniforme, por lo que se estandarizaron a YYYY-MM-DD para asegurar la compatibilidad con sistemas analíticos.
- **Valores atípicos en columnas de posiciones**: Se identificaron valores no estándar como "DNF" y "Retired", los cuales fueron transformados a -1, en el caso de race_position y pos, para permitir un análisis uniforme.
- **Baja identificabilidad**: La falta de unicidad en los atributos "Driver" y "DriverCode" implicaba problemas para identificar de manera precisa a los pilotos, lo cual fue una preocupación constante en el proceso.

El notebook de transformación se centró en resolver problemas de calidad de datos, incluyendo la **eliminación de duplicados** mediante claves compuestas, la **imputación de valores nulos** con marcadores adecuados, y la **estandarización** de formatos de fechas a YYYY-MM-DD. Además, se abordaron inconsistencias en los nombres de las columnas, se realizaron **transformaciones en campos de texto** para garantizar la consistencia, y **se gestionaron valores atípicos** en columnas de posiciones.

- **Propuestas para futuros pasos y mejoras:**

Para mejorar la calidad y disponibilidad de los datos del Formula 1 Official Data, se recomienda implementar un **sistema de monitoreo continuo** de la calidad de los datos que detecte automáticamente valores atípicos, inconsistencias y duplicaciones cuando se carga nueva información. El sistema debe integrarse en el proceso de validación en tiempo real para garantizar que los datos cumplan ciertos estándares de calidad, antes de usarse para análisis o informes. Además, sería adecuado establecer reglas más estrictas para la identificación única de registros, especialmente en los campos "Driver" y "DriverCode", para asegurar la precisión en la identificación de pilotos y eventos.

Otra mejora importante es la automatización del proceso de estandarización de formatos y la imputación de valores nulos, utilizando algoritmos de *machine learning* para predecir valores faltantes basados en patrones históricos. Además, la actualización periódica de la base de datos y la integración de otras fuentes de datos pueden enriquecer el conjunto de datos, y proporcionar una descripción general más completa y precisa. Estos pasos no sólo mejorarán la integridad y coherencia de los datos, sino que también facilitarán su uso en análisis y toma de decisiones informadas en el ámbito de la Fórmula 1.

Fase III: Codificación

Los códigos de extracción, transformación y carga están distribuidos en tres notebooks separados,

1. **Extract:** Se encuentra el código de extracción de datos, realizado en un notebook de Google Colab. Se confirmó y validó que los registros extraídos de los csv fueran los mismos de los csv originales.
2. **Transform:** Mediante códigos desarrollados, se aplicaron las métricas de calidad de datos, la limpieza de datos y la verificación y control de cambios de los datos limpiados.

3. **Load:** Se crearon la tabla de hechos y las tablas de dimensiones del modelo dimensional propuesto, y se cargaron los datos.

- **Entrega final csv:** Los csv de la entrega final que son congruentes con las tablas planteadas en el modelo dimensional, están ubicados en la carpeta correspondiente del proyecto, denominada entrega_final_csv. Esta carpeta contiene el sistema de almacenamiento final del dataset después de haber pasado por el proceso de ETL.

Conclusiones

En conclusión, el proyecto de ETL aplicado al "Formula 1 Official Data (1950-2022)" demostró ser esencial para entender la complejidad y diversidad de los datos relacionados con las carreras, temporadas y pilotos de Fórmula 1. A través de un procedimiento riguroso de extracción, transformación y carga, se logró establecer una base de datos sólida y confiable, que facilita análisis exhaustivos y apoya la toma de decisiones en el área de estudio. La evaluación integral de las métricas de calidad y el proceso de limpieza, permitió identificar y abordar problemas críticos como la presencia de duplicados, la inconsistencia en los formatos, y la imputación de valores nulos, asegurando de este modo datos coherentes y útiles para análisis futuros.

Por otro lado, una de las conclusiones más relevantes del proyecto es la importancia de mantener procesos de aseguramiento de calidad constantes en cualquier tipo de análisis de datos. Aunque los datasets iniciales presentaban un alto grado de integridad, se detectaron áreas de mejora en las métricas de estructura, completitud y consistencia. Esto subraya la necesidad de implementar sistemas de monitoreo continuo y algoritmos de *machine learning*, para la imputación de valores faltantes y la estandarización de formatos.

Finalmente, se considera que estas mejoras garantizarán que los datos permanezcan precisos y relevantes a lo largo del tiempo, fortaleciendo así la base de datos para futuros estudios y análisis en el ámbito de la Fórmula 1. Conviene

indicar que, la calidad de la información es clave para obtener resultados confiables y maximizar su valor en contextos de toma de decisiones. Por este motivo, es esencial establecer procesos continuos de aseguramiento de calidad y actualización de los datos, que posibiliten explorar nuevas perspectivas y generar conocimientos más profundos, en el apasionante mundo de la Fórmula 1.