Inteligencia de Negocio

Obligatorio final

Grupo 2
Estudiantes:

Bruno Cattáneo
Micaela Olivera
Dhiago Rivera
Matias Rodriguez

28/11/2021

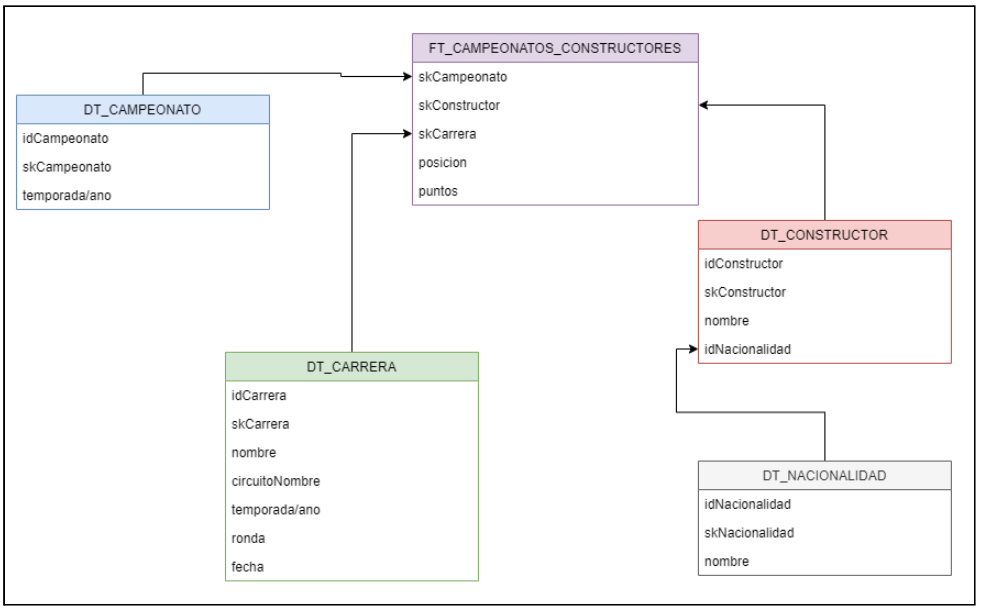
Índice

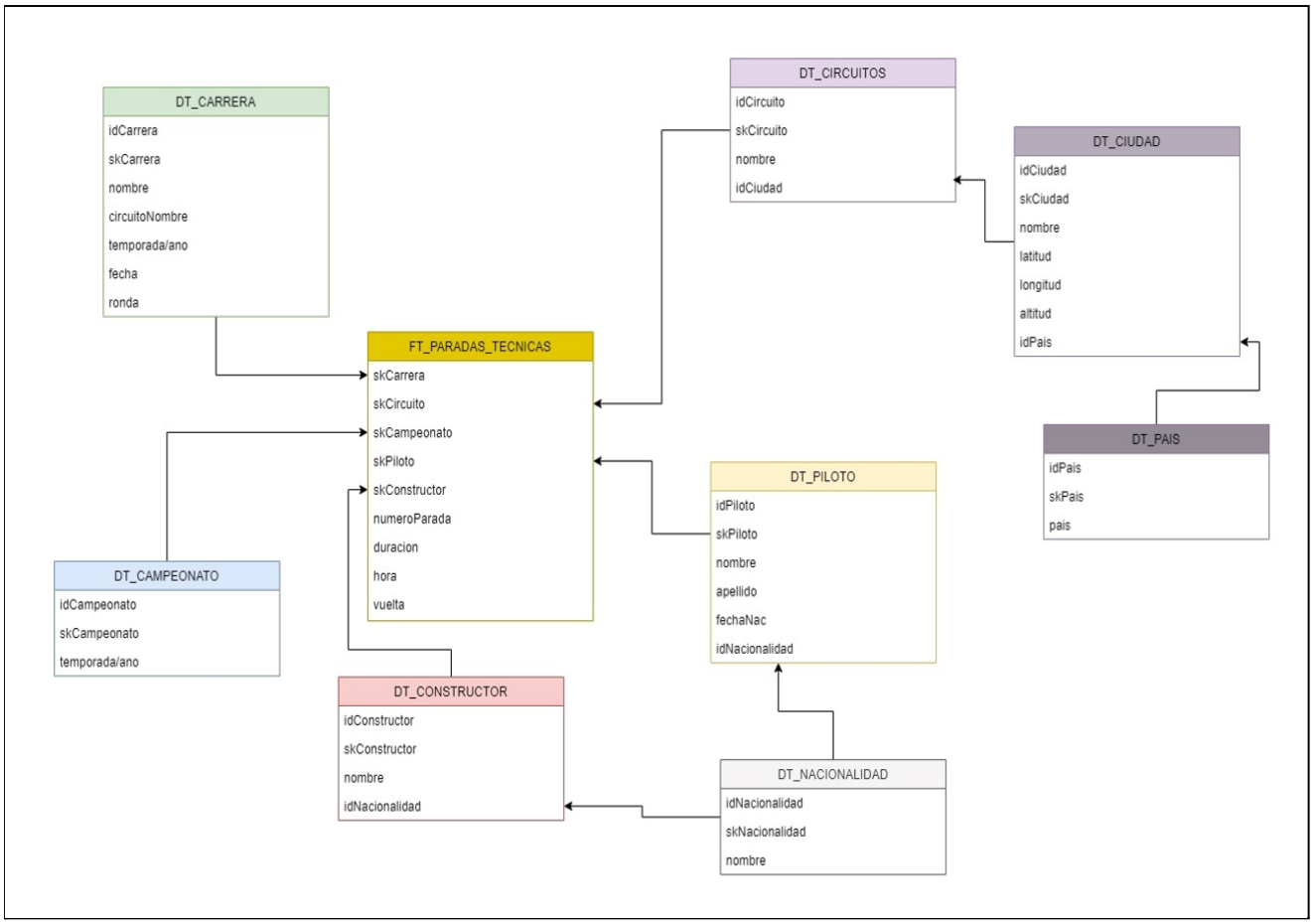
Introducción

El presente informe tiene como cometido plasmar el resultado de aplicar los conocimientos adquiridos durante el curso en el trabajo obligatorio final. El obligatorio como tal, tiene como objetivo diseñar e implementar un datawarehouse que permita conocer en mayor detalle la evolución de la Fórmula 1 y como han ido evolucionando diversos aspectos de este deporte a lo largo del tiempo. Para ello, se emplearon técnicas de ETL y se implantó un datawarehouse con la información necesaria para poder responder las preguntas del negocio. Finalmente, se generaron un conjunto de visualizaciones específicas que permiten cumplir el objetivo de conocer un poco más sobre este deporte. A nivel de herramientas, se utilizó Pentaho para los procesos ETL, Tableau como herramienta de visualización de datos interactivos y MySQL como motor de base de datos.

Diseño del DataWarehouse

Para el diseño del datawarehouse decidimos implementar un esquema de snowflake, dado que no queríamos perder la oportunidad de mantener los datos con una mayor granularidad, como así también evitar el exceso de información redundante, lo que contribuye a obtener un diseño más claro. Por otra parte, visto que los esquemas obtenidos no generaban problemas de performance al efectuar las consultas en la base de datos, y que algunas dimensiones se deseaban relacionar en diversas tablas de hechos, optamos por mantener esta alternativa de diseño por sobre el esquema estrella. Como resultado del análisis realizado determinamos la necesidad de contar con 4 tablas de hechos (FT) y 8 tablas de dimensiones (DT) relacionadas entre sí de la siguiente manera:





Procesos de ETL

Para la construcción de los procesos de ETL se tuvieron en cuenta las buenas prácticas aprendidas en el curso que refieren a:

- Analizar y entender los data sources que se van a utilizar.

  Esto es parte del entendimiento previo que debe hacerse del negocio y sus procesos a fin de identificar la información que se debe extraer de los data sources. Para este caso de estudio puntual se dispuso de 13 archivos .csv conteniendo información estadística de las carreras, pilotos, campeonatos y demás información relativa a este deporte desde 1950 hasta 2020.

- Preguntarse con qué frecuencia se desea cargar los datos.

  Si bien para este caso, la carga se hace asíncronamente y por única vez, esta pregunta es clave ante un caso de BI real al que nos enfrentemos en la vida profesional.

- Tener una estrategia ante la falla de la carga de datos.

Al igual que el punto anterior, si bien a los efectos de este obligatorio no es determinante, sí lo será en cualquier sistema de BI en producción que da soporte a los procesos de un negocio en particular.

- Realizar una transformación en Pentaho (archivo .ktr) por cada tabla de datos a cargar (FTs y Dts).

  Para este caso, se implementaron un total de 12 transformaciones (4 FTs y 8 DTs) y se hizo uso de un conjunto importante de herramientas de Pentaho como ser: lookup, joins, filtros, fórmulas, entre otros.

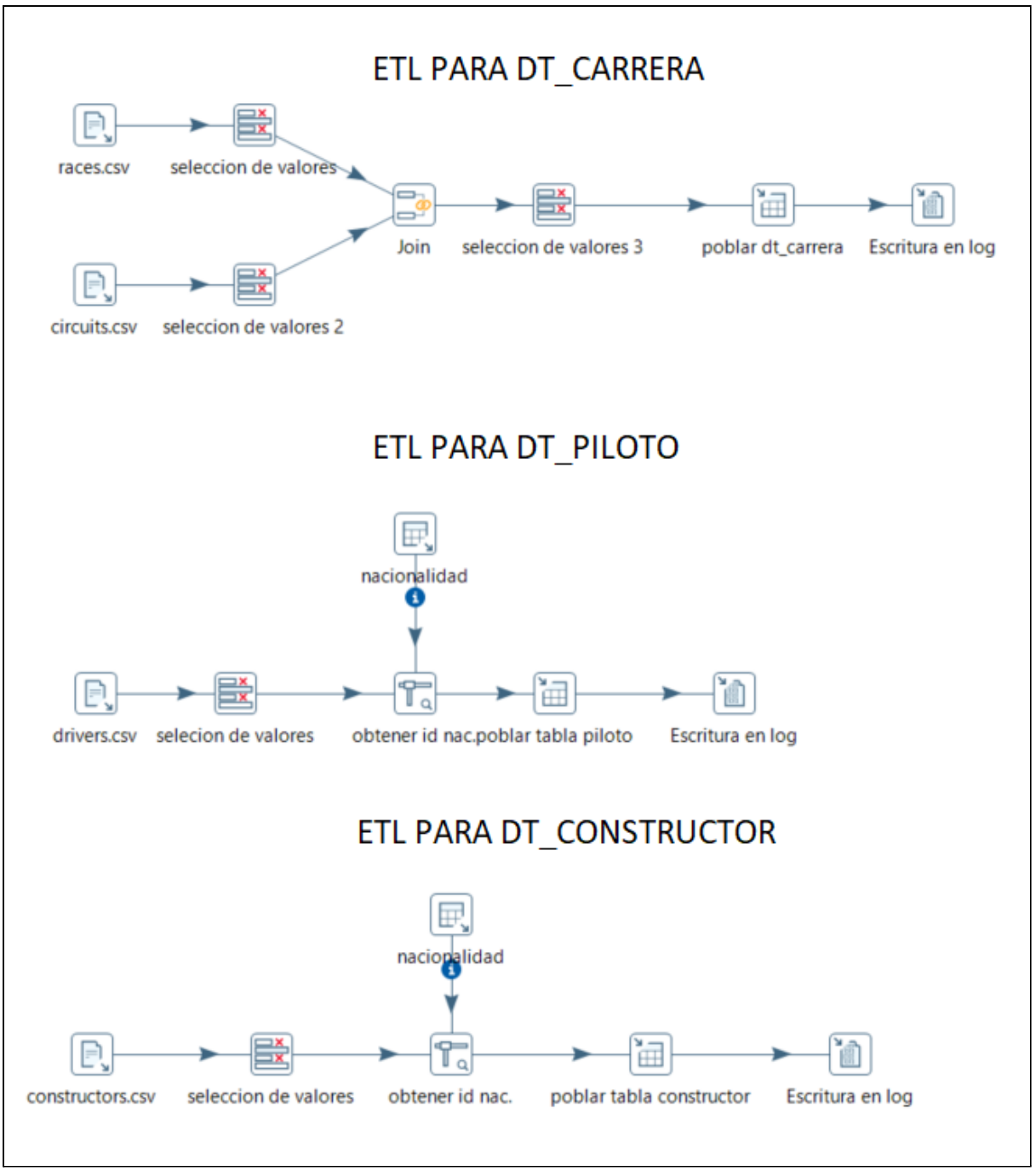
- Utilizar uno o varios jobs (archivo .kjb) para ejecutar las transformaciones encadenadas y de esta forma automatizar el proceso.

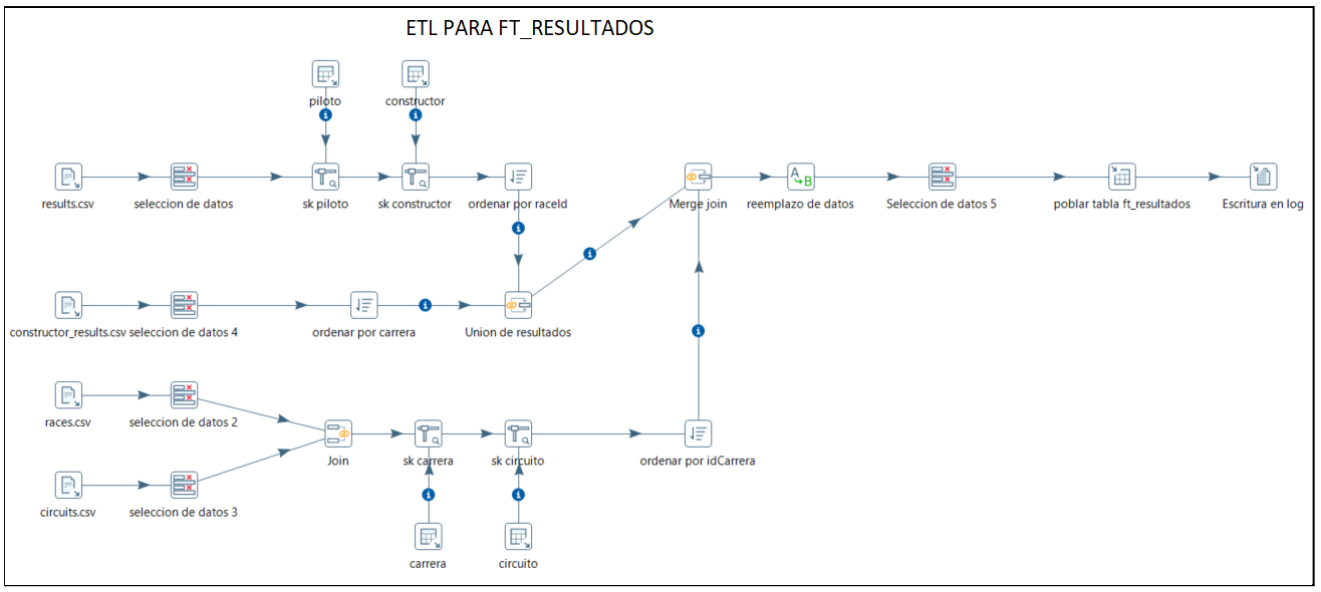
  En nuestro caso, se implementó un job para la carga de todas las FTs, un job para la carga de todas las DTs y finalmente un último job que encadena los jobs anteriores.

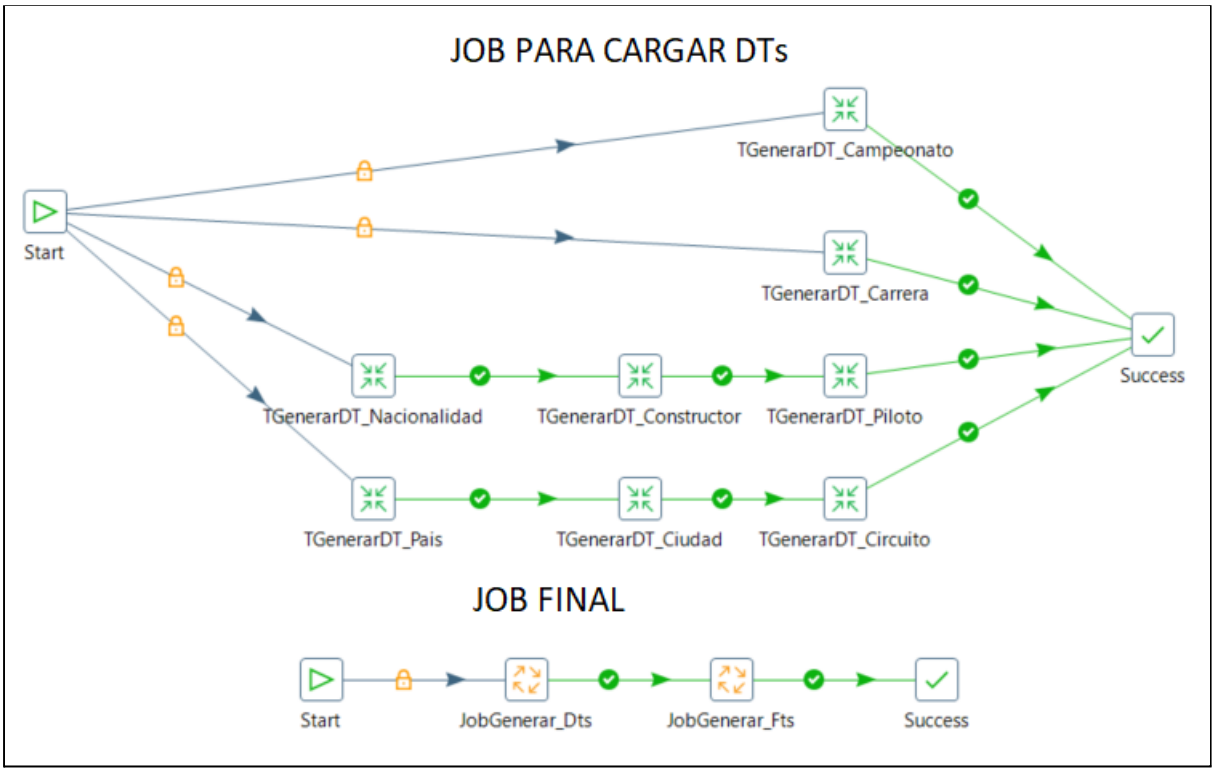
- Mantener un log de actividad de todas las acciones realizadas dentro de la herramienta.

  Se implementó un registro de la actividad completa en cada etapa de las transformaciones lo que permite generar trazas de auditoría y contribuye con un adecuado nivel de control de las operaciones.

A continuación se presentan capturas de pantalla de algunas de las transformaciones y jobs implementados en Pentaho a modo de ejemplo:


ETL PARA DT_CARRERA
races.csv
seleccion de valores
Join
seleccion de valores 3
poblar dt_carrera
Escritura en log
circuits.csv
seleccion de valores 2
ETL PARA DT_PILOTO
nacionalidad
drivers.csv
selecion de valores
obtener id nac.
poblar tabla piloto
Escritura en log
ETL PARA DT_CONSTRUCTOR
nacionalidad
constructors.csv
seleccion de valores
obtener id nac.
poblar tabla constructor
Escritura en log





Consultas SQL

A modo de validar el adecuado diseño del datawarehouse, se presentan un conjunto de consultas SQL que responden a diversas necesidades particulares del negocio:

Corredores más ganadores de carreras por país:

```
SELECT count(*), dt_nacionalidad.nacionalidad, dt_piloto.nombre,
dt_piloto.apellido FROM ft_resultados, dt_piloto, dt_nacionalidad
WHERE ft_resultados.posicion = 1
AND ft_resultados.sk_piloto = dt_piloto.sk_piloto
AND dt_nacionalidad.id_nacionalidad = dt_piloto.id_nacionalidad
GROUP BY dt_nacionalidad.nacionalidad, dt_piloto.nombre,
dt_piloto.apellido
ORDER BY count(*) DESC;
```

Equipos más ganadores de carreras por país:

```
SELECT count(*), dt_nacionalidad.nacionalidad,
dt_constructor.nombre FROM ft_resultados, dt_piloto,
dt_nacionalidad, dt_constructor
WHERE ft_resultados.posicion = 1
AND ft_resultados.sk_piloto = dt_piloto.sk_piloto
AND dt_nacionalidad.id_nacionalidad = dt_piloto.id_nacionalidad
AND dt_constructor.sk_constructor = ft_resultados.sk_constructor
GROUP BY dt_nacionalidad.nacionalidad, dt_constructor.nombre
ORDER BY count(*) DESC;
```

Evolución ordenada de duración de carreras a lo largo del tiempo:

```
SELECT ft_resultados.campeonato_anio,
SUM(ft_resultados.milisegundos) FROM ft_resultados, dt_piloto,
dt_nacionalidad, dt_constructor
AND ft_resultados.sk_piloto = dt_piloto.sk_piloto
AND dt_nacionalidad.id_nacionalidad = dt_piloto.id_nacionalidad
AND dt_constructor.sk_constructor = ft_resultados.sk_constructor
GROUP BY ft_resultados.campeonato_anio
ORDER BY ft_resultados.campeonato_anio ASC;
```

Visualizaciones

Como ya se mencionó anteriormente, se utilizó Tableau para construir y presentar las siguientes visualizaciones interactivas:

1. Corredores y equipos más ganadores por país
2. Evolución de la duración de las carreras desde los inicios hasta la actualidad
3. Análisis de las paradas técnicas:
    a. Vuelta promedio en la que se realizan cada una
    b. Cantidad promedio por equipo
    c. Demora promedio

Adicionalmente, se elaboró un dashboard a fin de concentrar toda la información en una misma visualización.

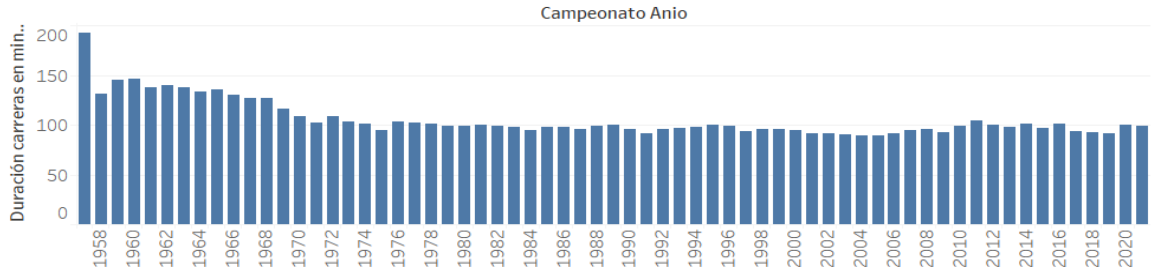
El resultado de este trabajo se presenta en 2 libros. El primero responde las preguntas 1 y 2, mientras que el segundo responde las siguientes 3 (3a, 3b y 3c). La decisión de este diseño refiere a un tema de claridad ya que las primeras 2 hacen uso de la FT_RESULTADOS, mientras que las restantes consultas surgen a partir de la FT_PARADAS_TECNICAS.

A continuación, se presentan capturas de las 5 visualizaciones y los dashboards correspondientes:

Análisis de Pilotos, Constructores y Carreras:

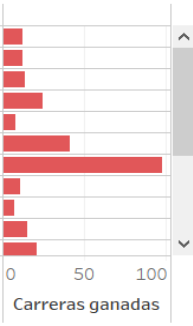
Evolución de la duración de las carreras a lo largo del tiempo.

Se tomo como criterio el tiempo que demora el primero en finalizar.



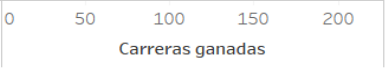
Corredores más ganadores de carreras por país

Nacionalidad	Apellido	Nombre
American	Andretti	Mario
Argentine	Reutemann	Carlos
Australian	Brabham	Jack
Austrian	Lauda	Niki
Belgian	Ickx	Jacky
Brazilian	Senna	Ayrton
British	Hamilton	Lewis
Canadian	Villeneuve	Jacques
Colombian	Pablo Montoya	Juan
Dutch	Verstappen	Max
Finnish	RÃ¤ikkÃ¶nen	Kimi



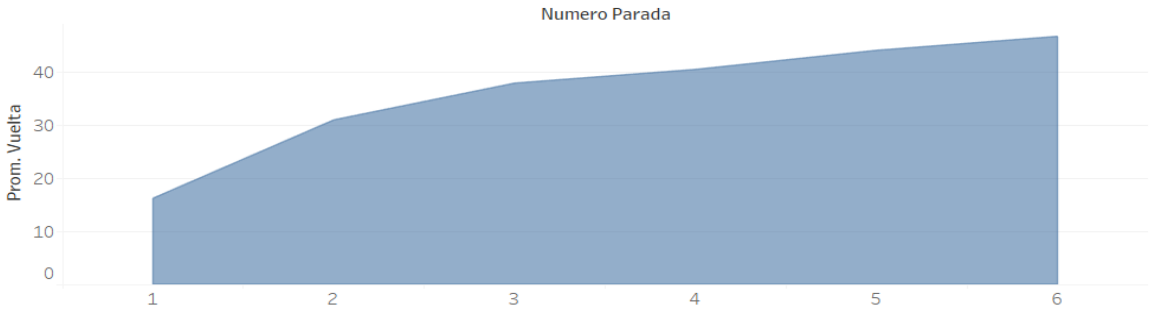
Equipos más ganadores de carreras por país

Nacional..	Nombre (..
American	Watson
Austrian	Red Bull
British	McLaren
Canadian	Wolf
French	Renault
German	Mercedes
Irish	Jordan
Italian	Ferrari
Japanese	Honda

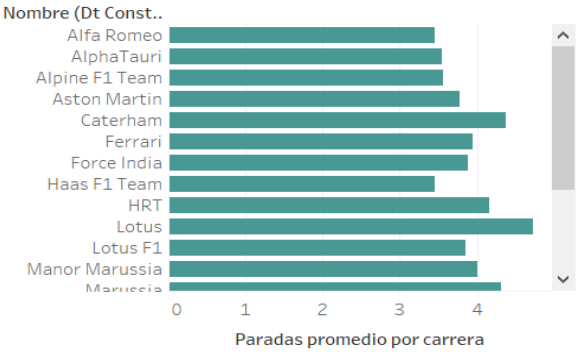


Análisis de las Paradas Técnicas:

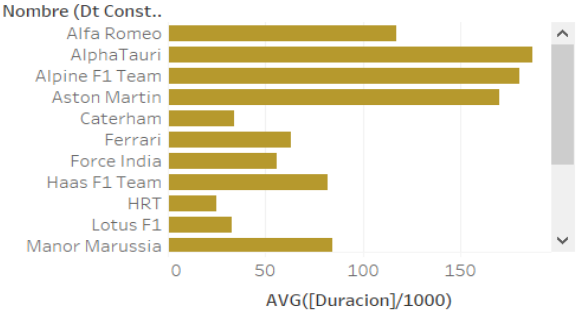
P3.1 Vuelta promedio parada



P3.2 Paradas por equipo



P3.3 Demora promedio paradas equipo



Lecciones aprendidas

A modo de conclusión general, se entiende completamente cumplido el objetivo de este obligatorio ya que fue posible aplicar la metodología aprendida durante el curso para diseñar e implementar un datawarehouse que permitió conocer en mayor profundidad el deporte de la Fórmula 1 respondiendo a las preguntas del negocio.

No obstante, existieron algunas dificultades que pudieron ser sorteadas y que como consecuencia aportaron las siguientes lecciones:

- Durante la creación de las visualizaciones en Tableau se identificaron algunas ausencias de dimensiones en las tablas, lo que implicó realizar iteraciones sobre el diseño y las modificaciones correspondientes en la base de datos. Entendemos que este proceso iterativo es “habitual” y que en último caso la experiencia de trabajo con distintos modelos de datos puede colaborar a obtener un diseño más exacto desde una etapa temprana.
- La herramienta Tableau presentó problemas a la hora de cargar la base de datos PostgreSQL en equipos con sistema operativo MacOS. Luego de varias pruebas, fue necesario migrar la base de datos a una MySQL ya que no fue posible solucionar la integración.
- Dentro de Tableau, es necesario realizar joins entre tablas. Esto se realiza arrastrando y soltando tablas que comparten una clave primaria dentro de la pestaña “Fuente de datos”. En algunas ocasiones dicha relación no funciona automáticamente y fue necesario vincular las tablas en forma manual indicando el campo en común entre las tablas.
- En cuanto a la utilización de la herramienta PENTAHO, existieron dificultades para conectarse a la base de datos MySQL debido a que se necesitaba la instalación de un determinado driver el cual no funcionaba en todas sus versiones.
- Otro punto a recalcar es que la herramienta tiene una gran variedad de funcionalidades para realizar transformaciones, lo cual requirió de un periodo de adaptación por parte de los integrantes del equipo al momento de interactuar con el mismo, y mucha investigación al momento de abordar particularidades que surgieron al momento de tratar los datos.

Bibliografía

- Multidimensional Data Modeling in Pentaho. (2021b, noviembre 5). Hitachi Vantara Lumada and Pentaho Documentation. https://help.hitachivantara.com/Documentation/Pentaho/9.2/Work_with_data/Multidimensional_Data_Modeling_in_Pentaho

- Free Training Videos - 2021.3. (2021). Tableau. https://www.tableau.com/learn/training/20213