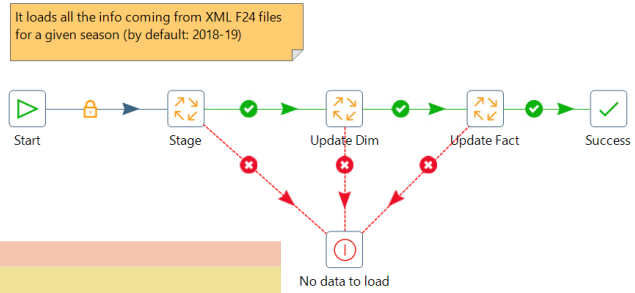
TRATAMIENTO DE DATOS OPTA

PFM - MASTER EN BIG DATA DEPORTIVO



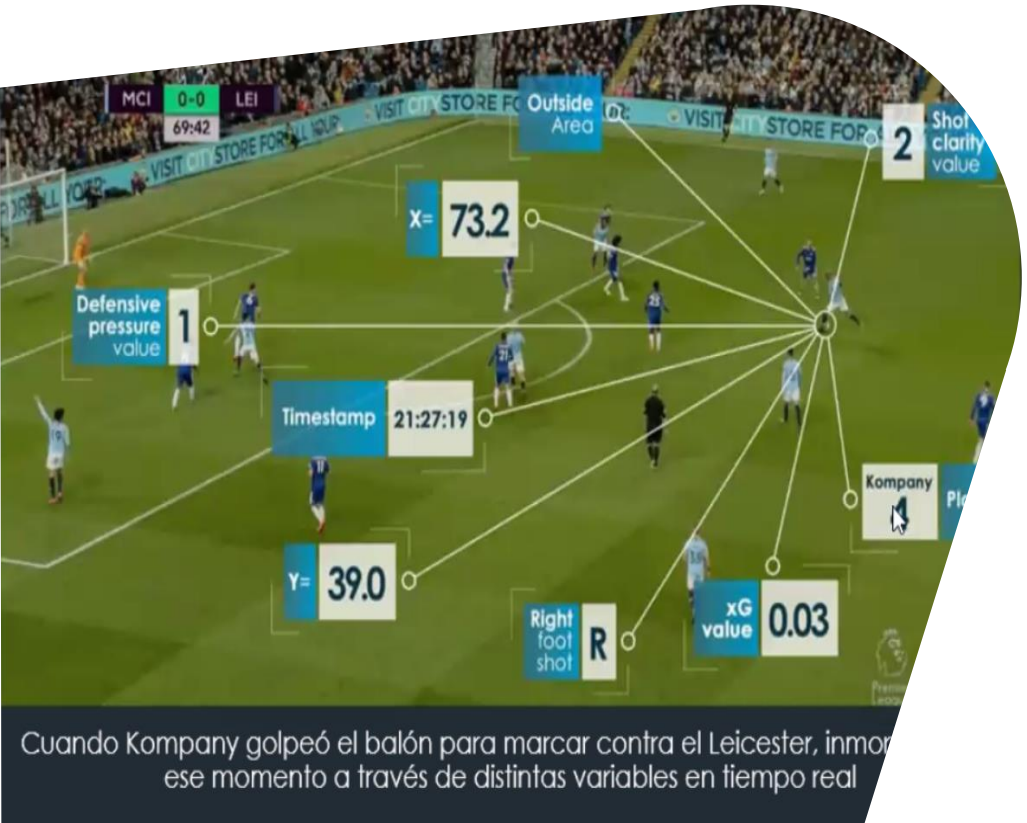
AUTOR: MARIO RODRIGÁLVAREZ GARCÍA
TUTOR: DAVID FOMBELLA
FECHA: JULIO 2020

ÍNDICE

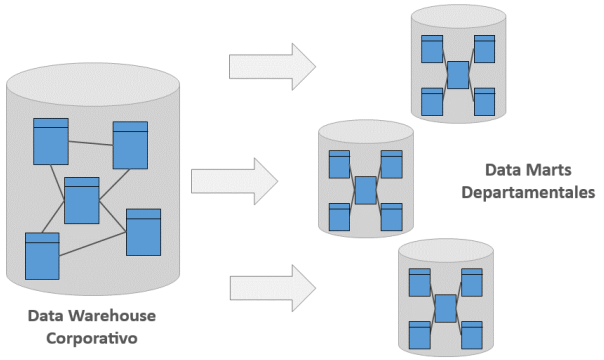
PLANTEAMIENTO DEL PROBLEMA

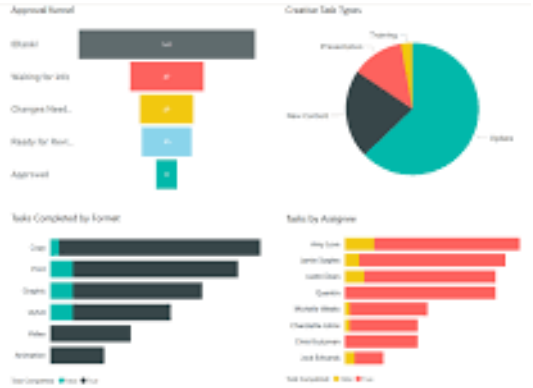
CONTEXTO Y MOTIVACIONES

- OPTA posee un amplio conjunto de **feeds de datos** de fútbol
- Sus clientes tienen acceso a estos datos en **formato XML**



- Se desea almacenar toda esta información de manera estructurada en una base de datos (**data warehouse/data mart**, según su alcance) que se pueda explotar mediante **visualizaciones analíticas**



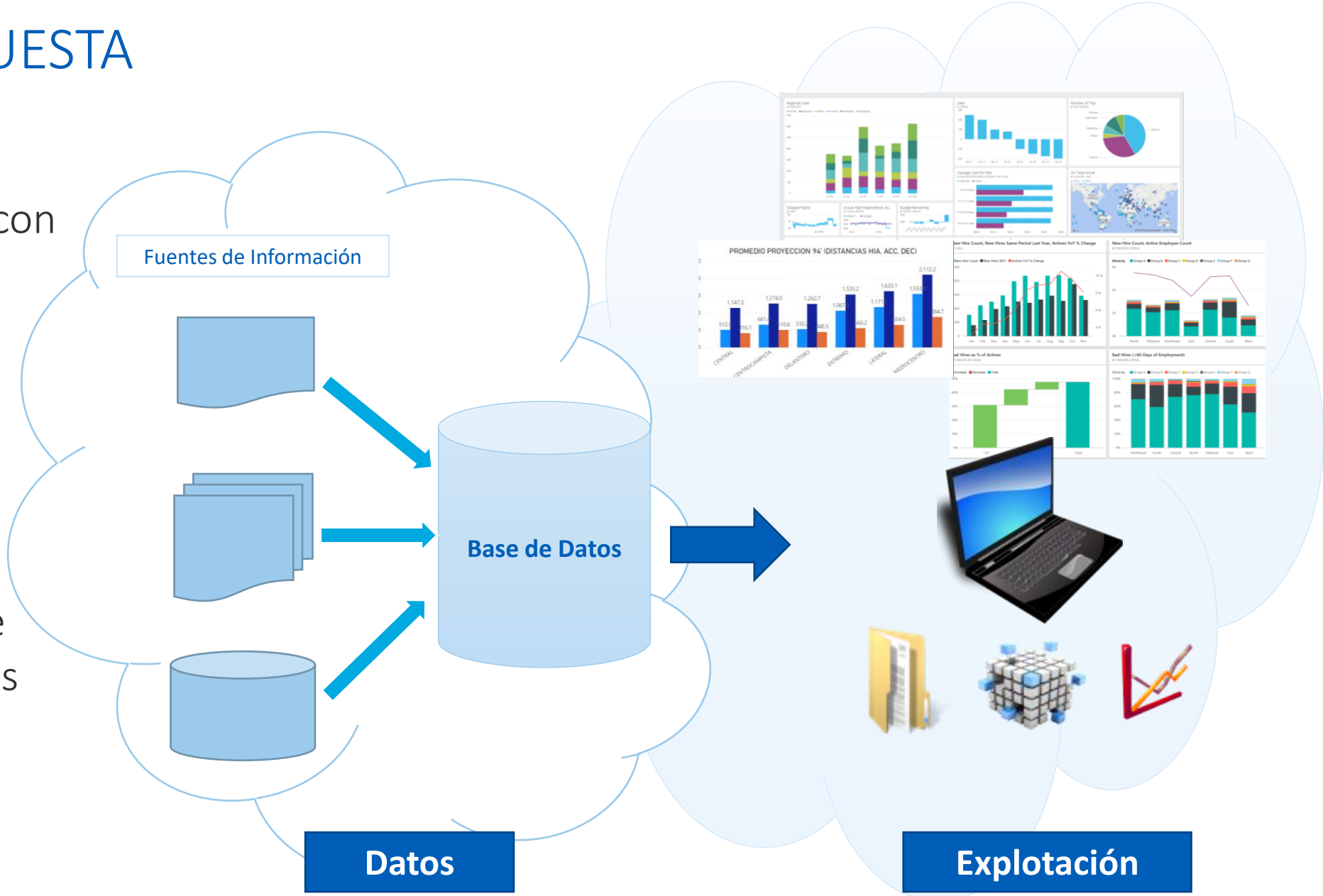


SOLUCIÓN PROPUESTA

_Nueva base de datos con información de OPTA

_Procesos de carga

_Visualización de la información mediante dashboards o informes



FUENTES DE DATOS

FICHEROS XML

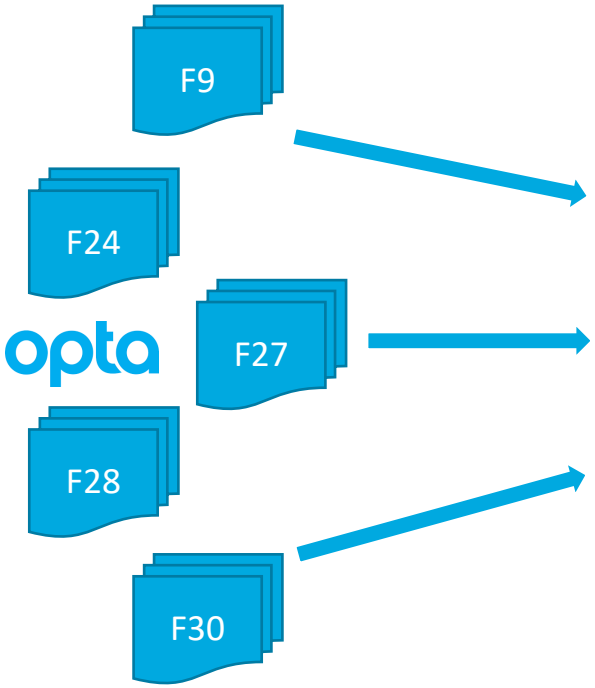
- Listado de ficheros que se desean cargar y analizar:

- F9 - Live Match Feed + Team and Player Stats → Estadísticas de equipos y jugadores en cada partido
- F24 - Event Details → Eventos de cada partido
- F27 - Pass Matrix & Average Formation → Matriz de Pases de cada partido
- F28 - Possession → Datos de Posesión en campo propio/rival
- F30 - Season Stats → Estadísticas agregadas de equipos y jugadores en la temporada

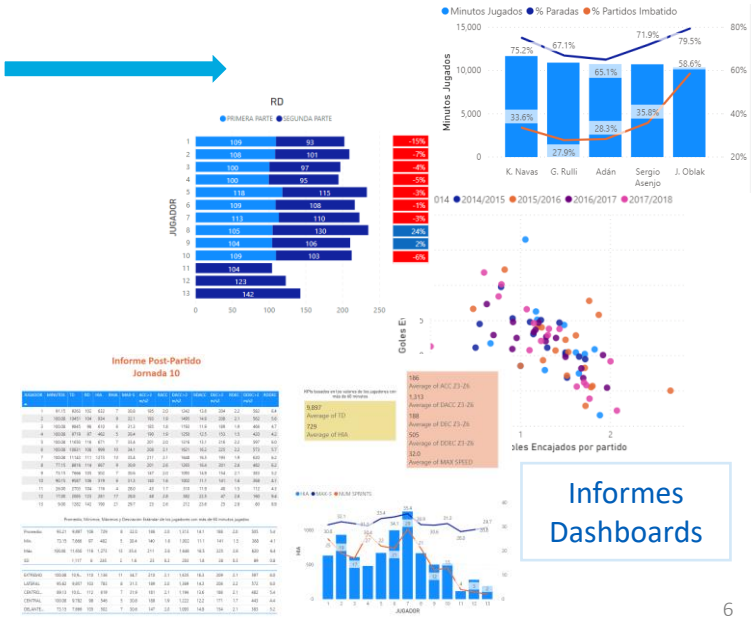
ARQUITECTURA DEL PROYECTO

TECNOLOGIAS UTILIZADAS

Fuentes de Información | Extract, Transform & Load | Data Warehouse | Explotación del Dato



Visualización



Informes Dashboards

(*) Se ha optado por PostgreSQL como gestor de base de datos pero se puede adaptar a cualquier otro. Pentaho Data Integration (PDI) ofrece conectores para gran cantidad de proveedores

LIMPIEZA, INGESTA Y PROCESAMIENTO DE DATOS

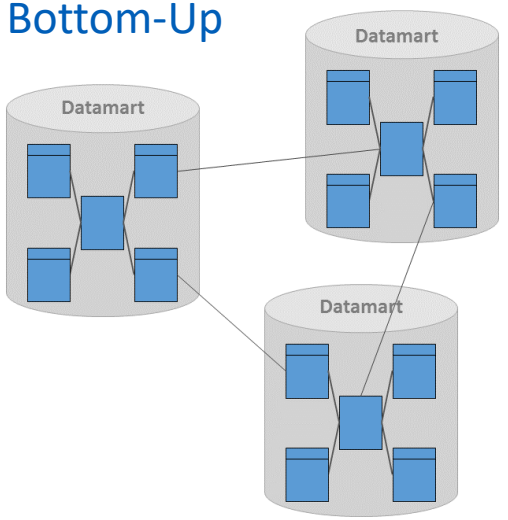
CONCEPTOS TÉCNICOS Y METODOLOGÍAS

_DATA WAREHOUSE (DW): repositorio de datos corporativos estratégicos, tácticos y operativos. Información de interés generada por la actividad de una empresa u organización

_DATA MART (DM): subconjunto físico/lógico del DW preparado para la consulta y análisis de la información de un área específico del negocio

_PROCESOS ETL (Extract, Transform y Load)

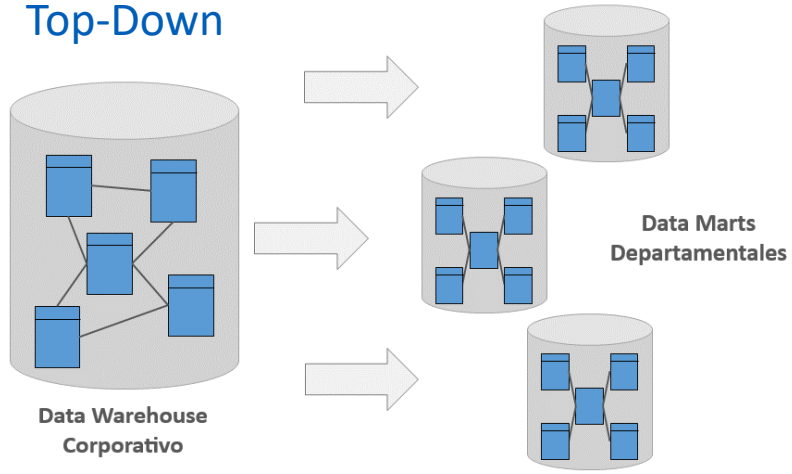
Bottom-Up



Características:

- Orientado a temas
- Integrado
- Variante en el tiempo
- No volátil

Top-Down

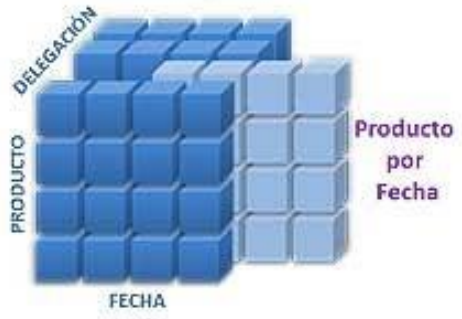


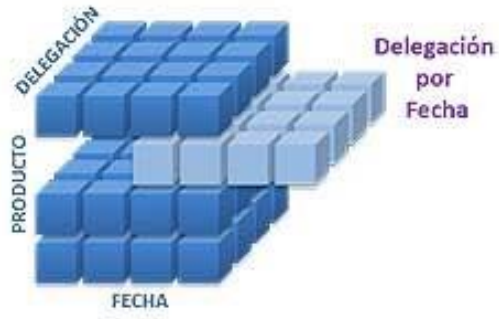
LIMPIEZA, INGESTA Y PROCESAMIENTO DE DATOS

CONCEPTOS TÉCNICOS Y METODOLOGÍAS

_MODELO MULTIDIMENSIONAL

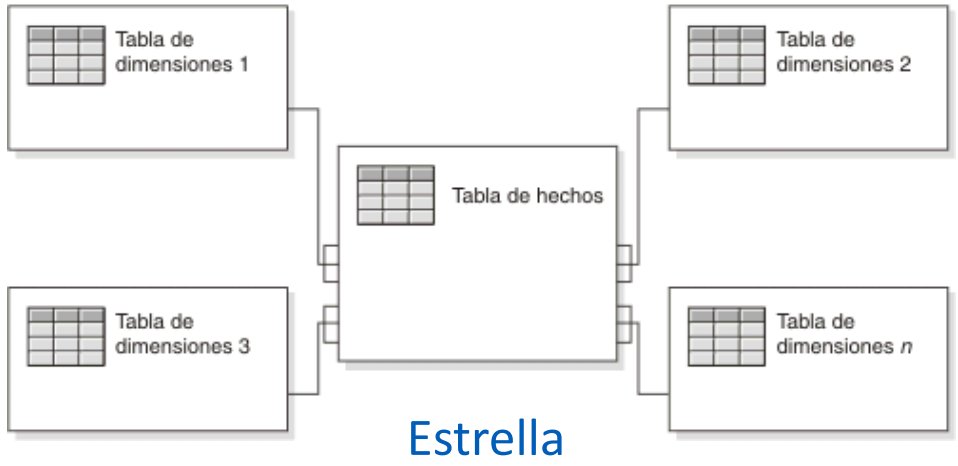
- Hechos
- Dimensiones



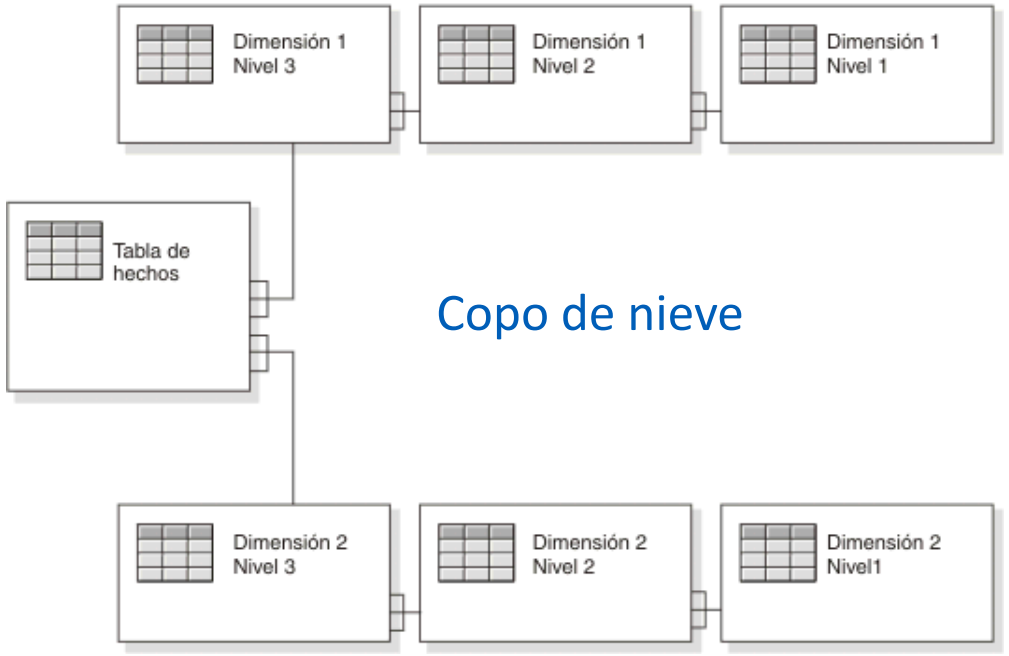


_ESQUEMAS DE DISEÑO

- Estrella
- Copo de nieve



Estrella



Copo de nieve

LIMPIEZA, INGESTA Y PROCESAMIENTO DE DATOS

PROCESOS ETL

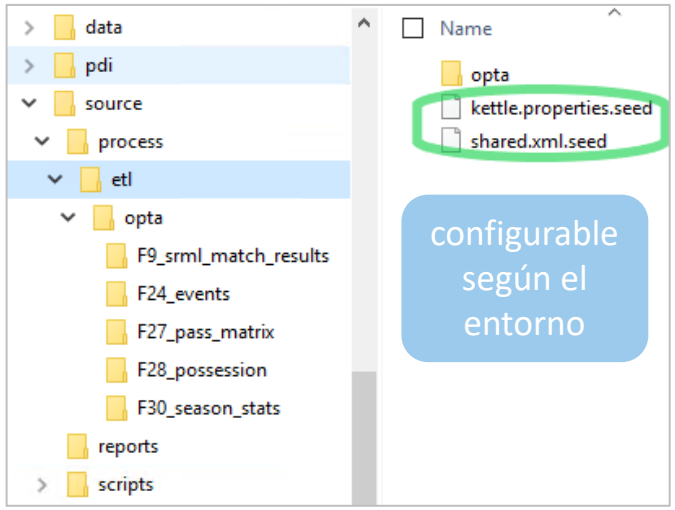
_Pentaho Data Integration (PDI)

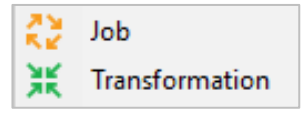
- Jobs / Trabajos
- Transformaciones

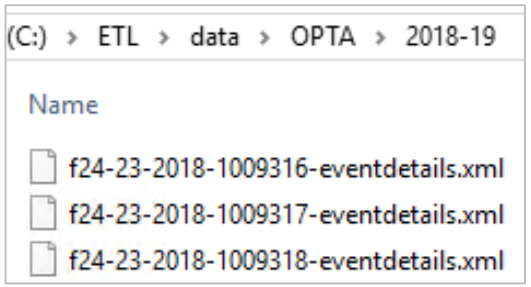
_Procesos parametrizados con los datos de la temporada (ej. *season = 2018-19*)

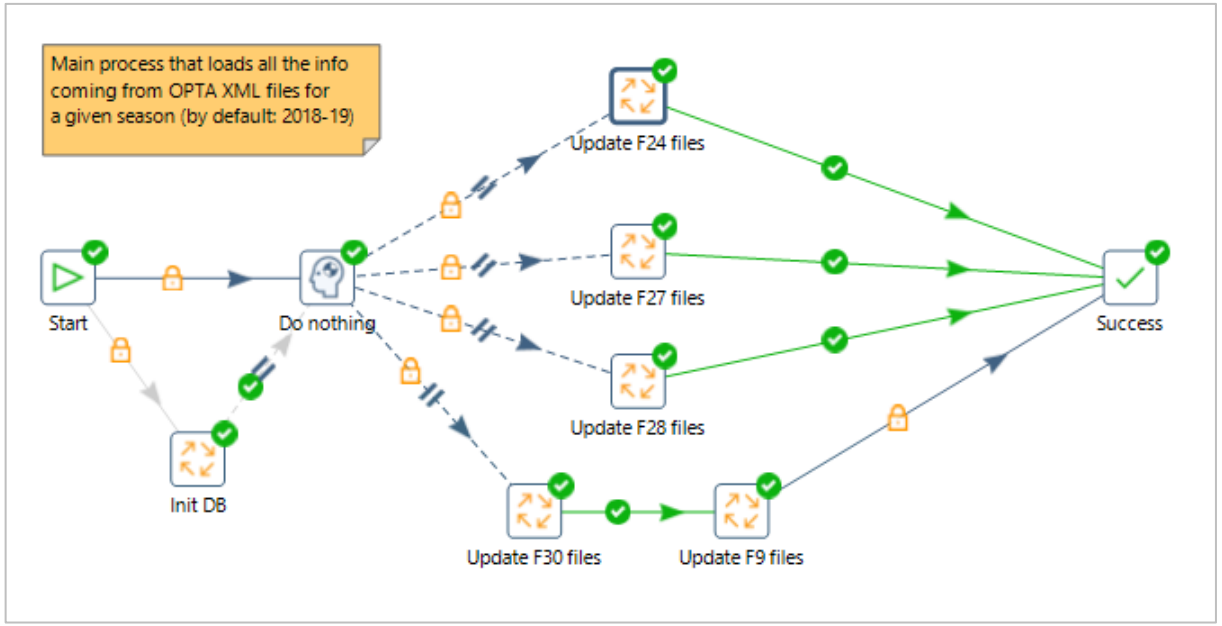
- El nombre del directorio que contiene los XML deberá ajustarse a este parámetro
- Ficheros *kettle.properties* y *shared.xml*

_Proceso/job principal "orquesta" la ejecución del resto de procesos









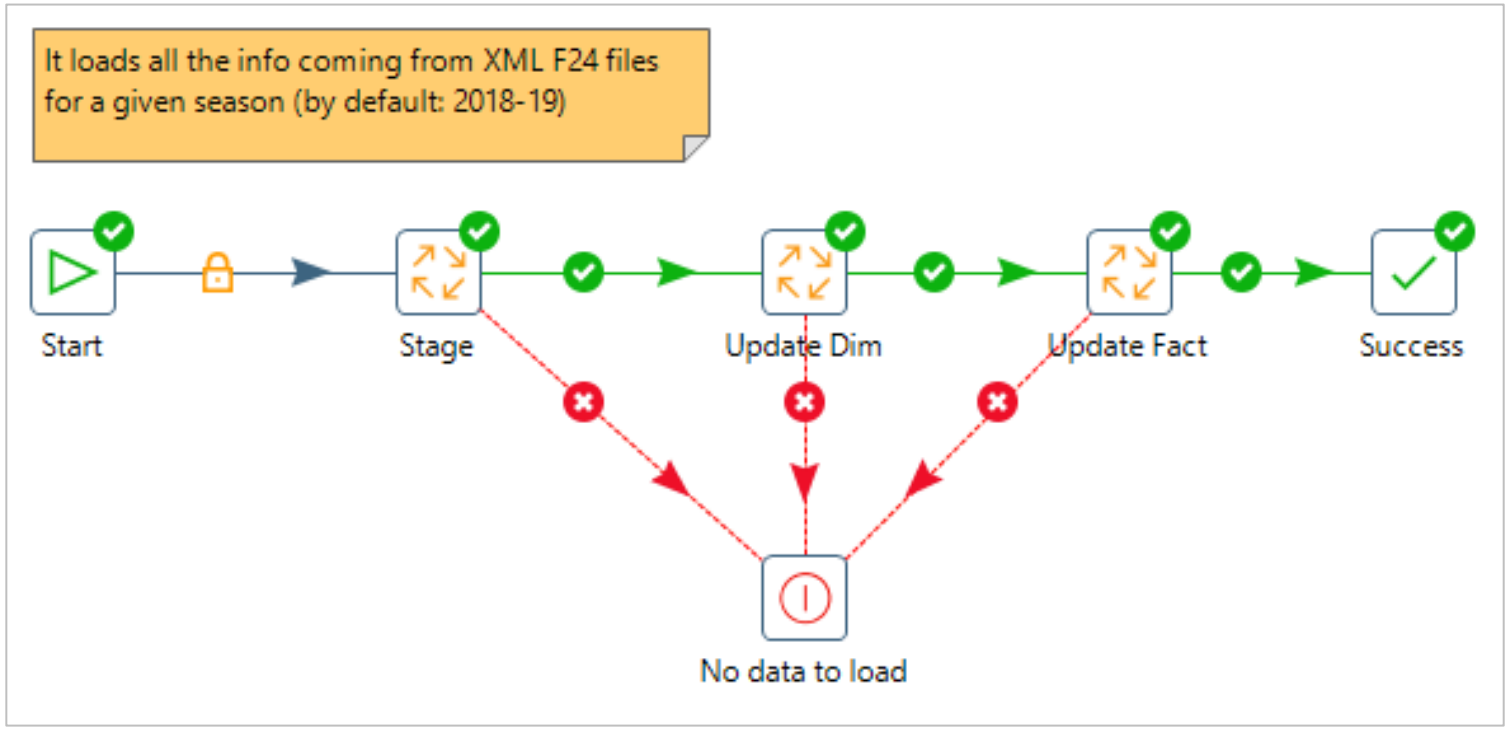
LIMPIEZA, INGESTA Y PROCESAMIENTO DE DATOS

PROCESOS ETL

Job
Transformation

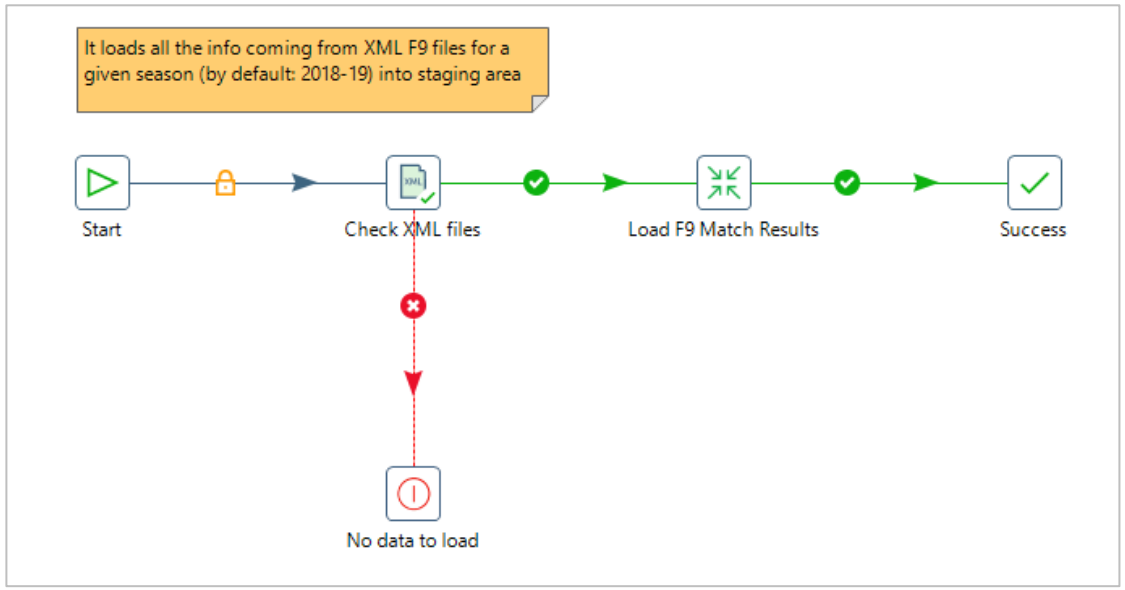
_Metodología común para todos los XML de OPTA con 3 trabajos (.kjb) y varias transformaciones (.ktr):

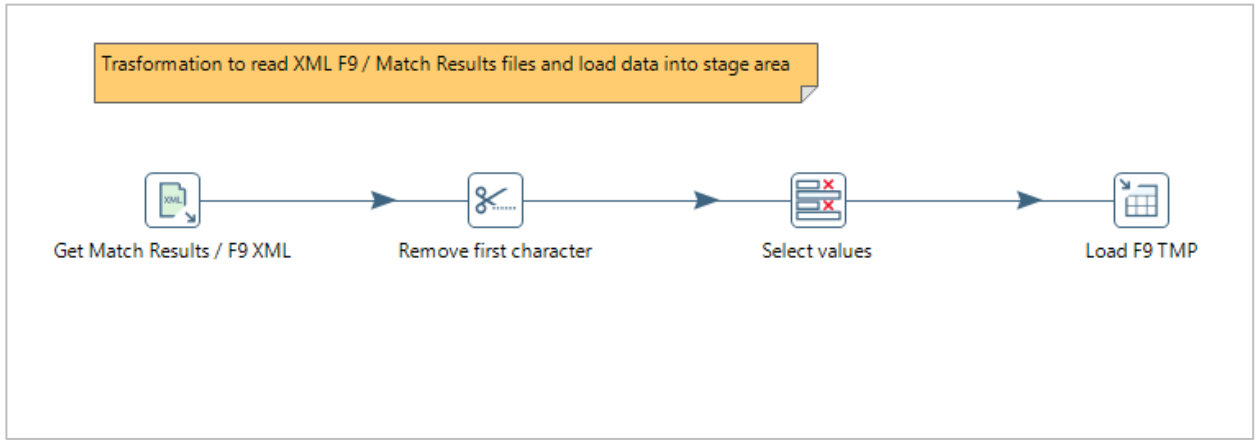
- Stage area, carga XML en tablas temporales
- Actualización de dimensiones (dim)
- Actualización de tablas de hechos (fact)

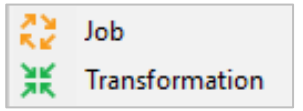


LIMPIEZA, INGESTA Y PROCESAMIENTO DE DATOS

PROCESOS ETL (Stage)





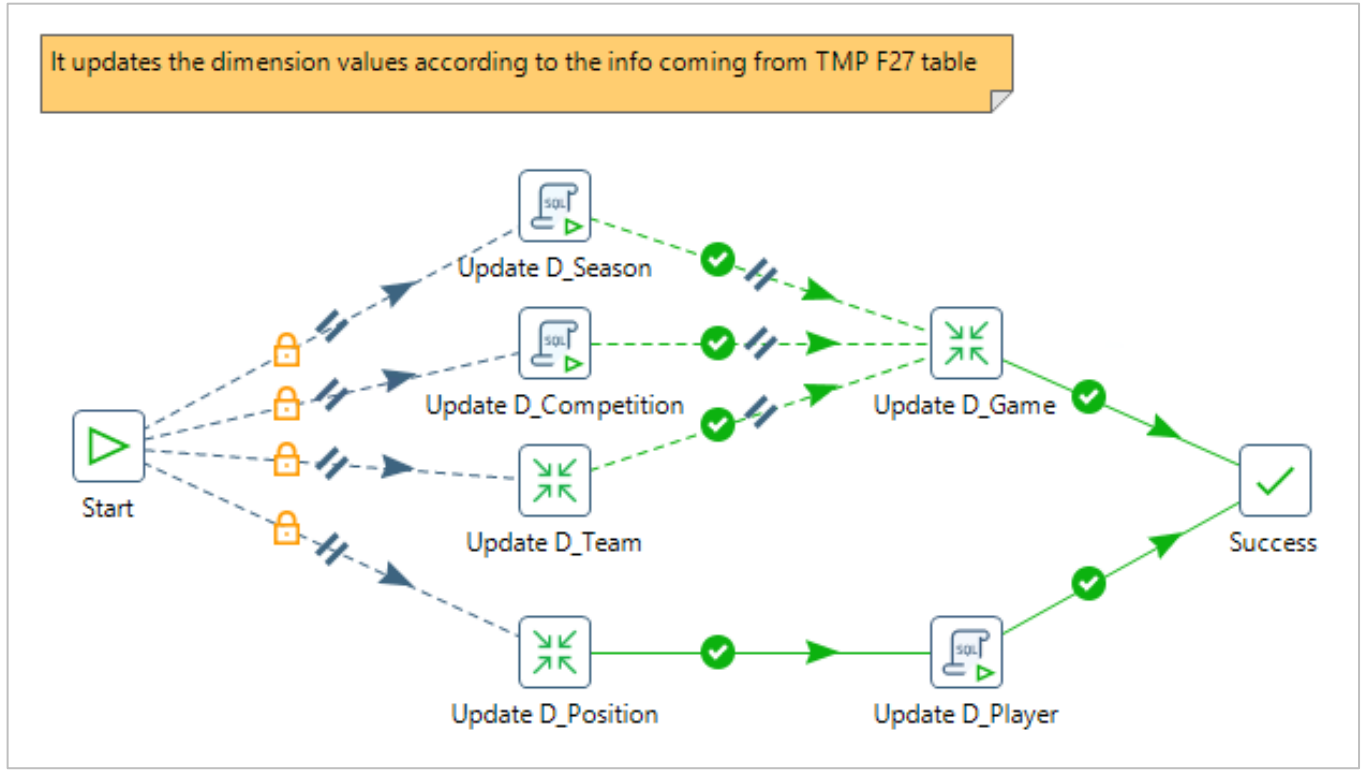


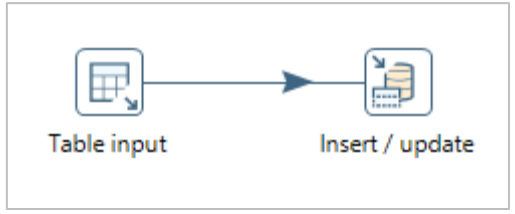
Jobs STAGE

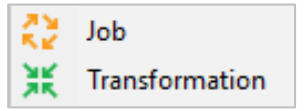
1. Comprobación de la existencia de ficheros XML
   - Error en caso de no haber datos
2. Transformación para cargar el contenido en el data stage de la base de datos
   - Lectura del fichero XML
   - Tratamiento de datos, si procede (creación de nuevas columnas, substring, pivotado…)
   - Inserción en tabla temporal

LIMPIEZA, INGESTA Y PROCESAMIENTO DE DATOS

PROCESOS ETL (Dimensiones)







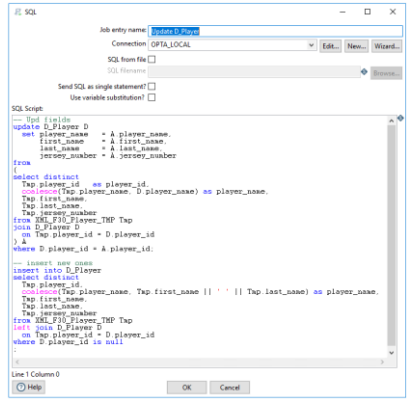
Update DIMs

_ Actualización de atributos

_ Inserción de nuevos elementos que aparecen en los ficheros a cargar

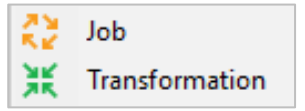
mediante

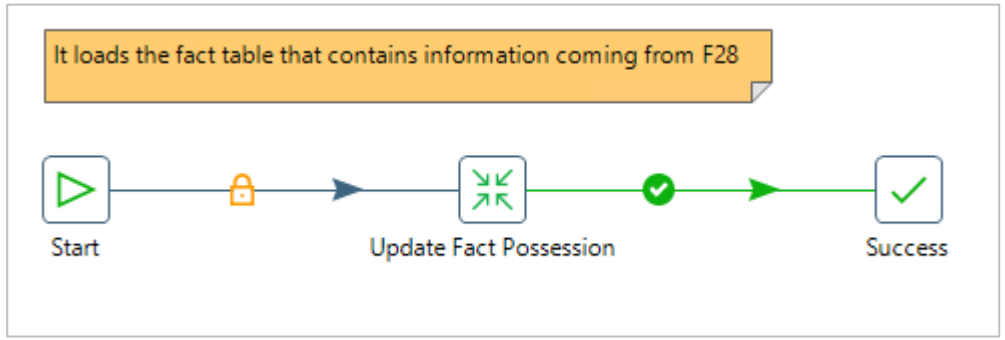
- Transformaciones específicas ofrecidas por PDI
- Queries SQL ad-hoc

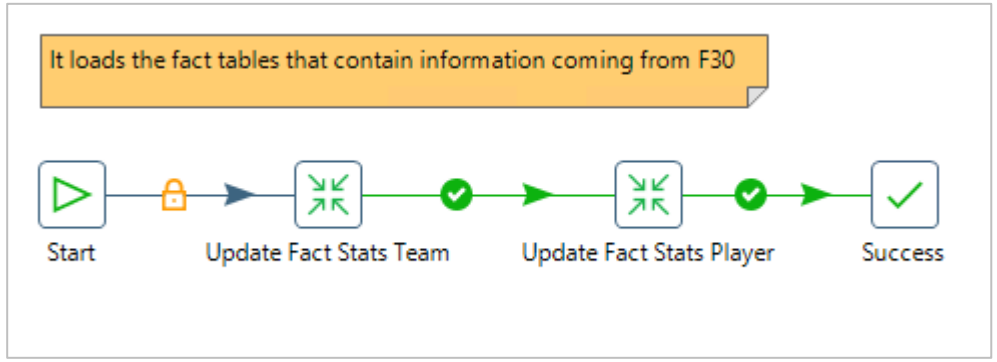


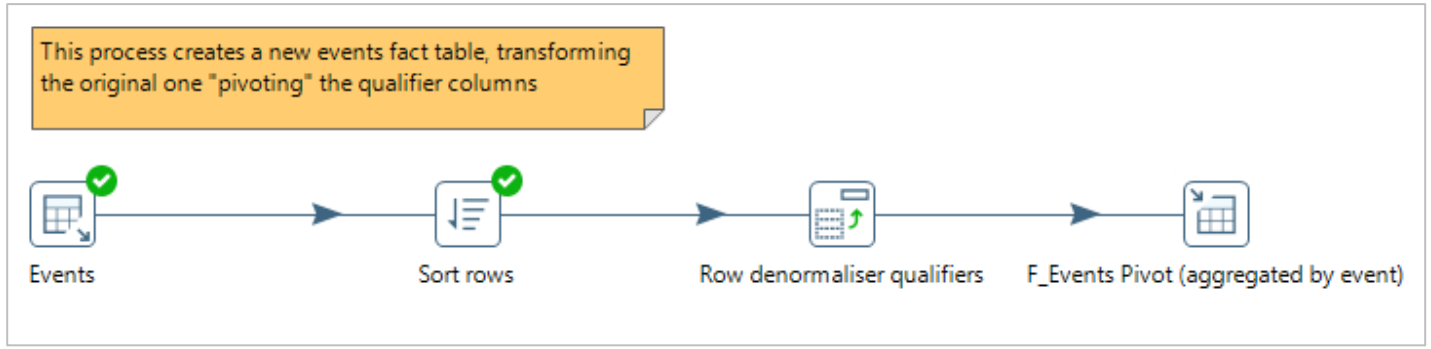
LIMPIEZA, INGESTA Y PROCESAMIENTO DE DATOS

PROCESOS ETL (Hechos)



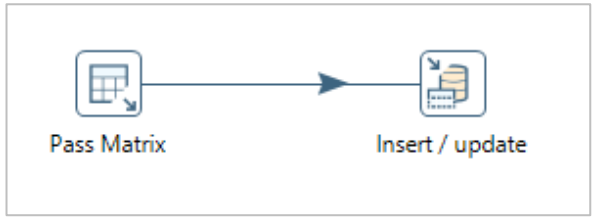


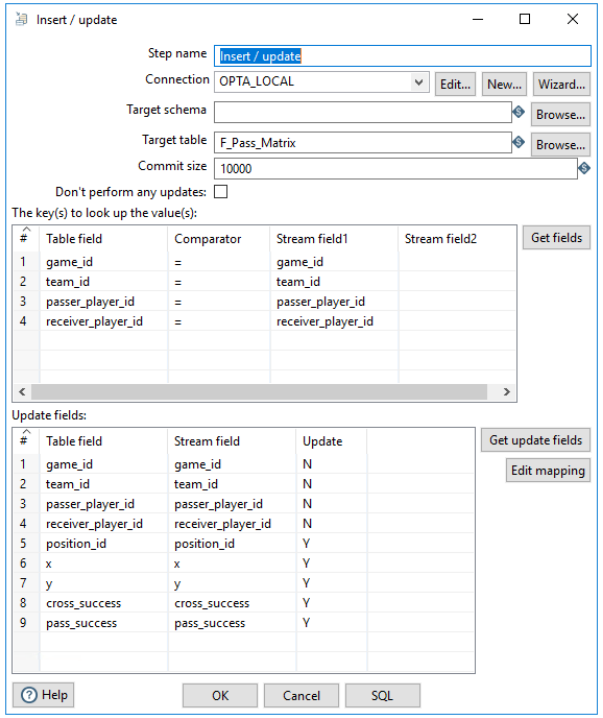




Update FACTs

Sin borrados, excepto en agregados/f24 pivotado, solo inserción de nuevos datos o actualización de los existentes en caso de recargas



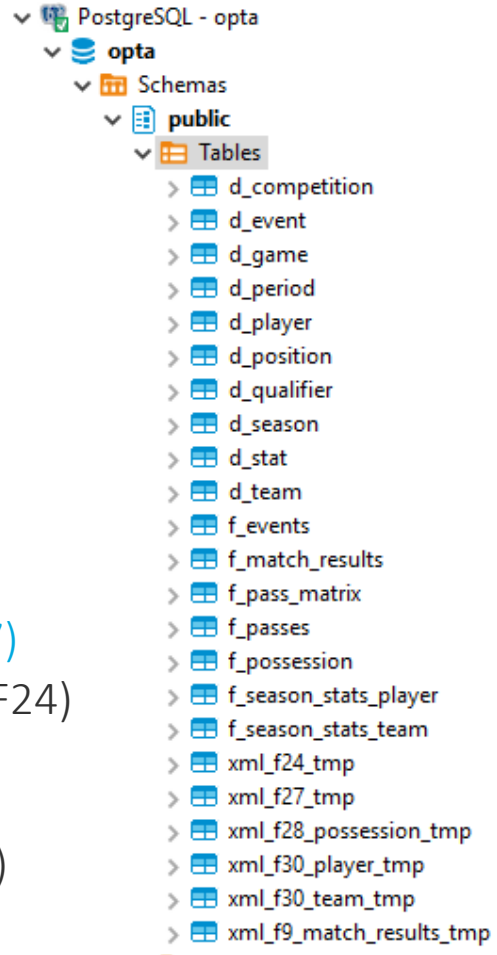


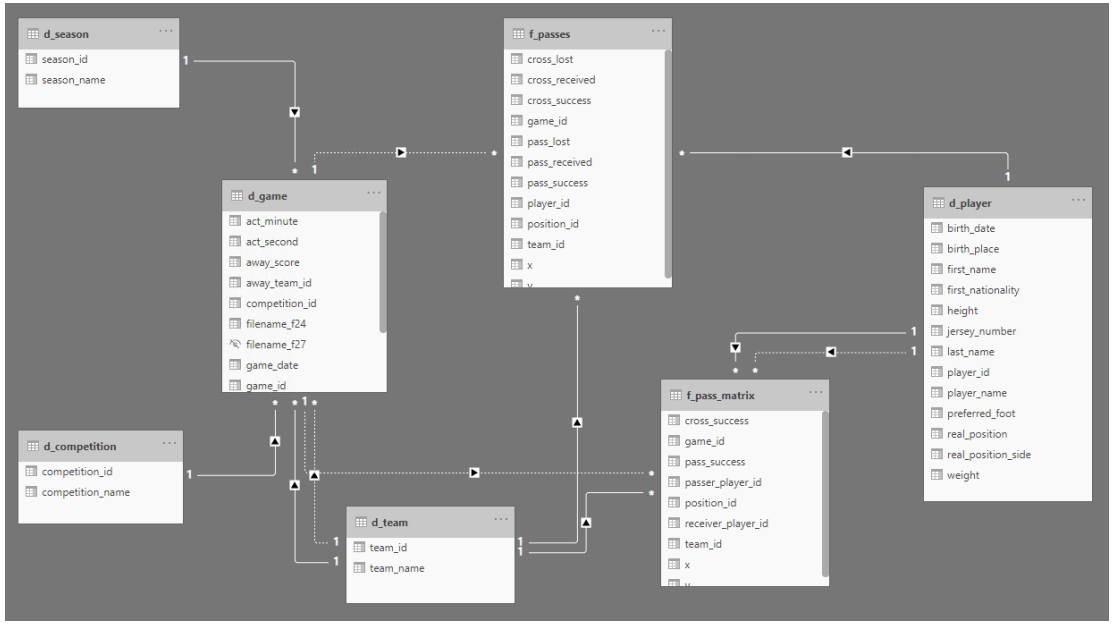
LIMPIEZA, INGESTA Y PROCESAMIENTO DE DATOS

MODELO DE DATOS

_Tablas de Dimensiones

- D_Competition
- D_Season
- D_Game
- D_Team
- D_Player
- D_Position
- D_Stat
- D_Event
- D_Qualifier





_Tablas de Hechos

- F_Pass y F_Pass_Matrix (F27)
- F_Events y F_Events_Pivot (F24)
- F_Possession (F28)
- F_Season_Stats_Team y F_Season_Stats_Player (F30)
- F_Match_Results (F9)

LIMPIEZA, INGESTA Y PROCESAMIENTO DE DATOS

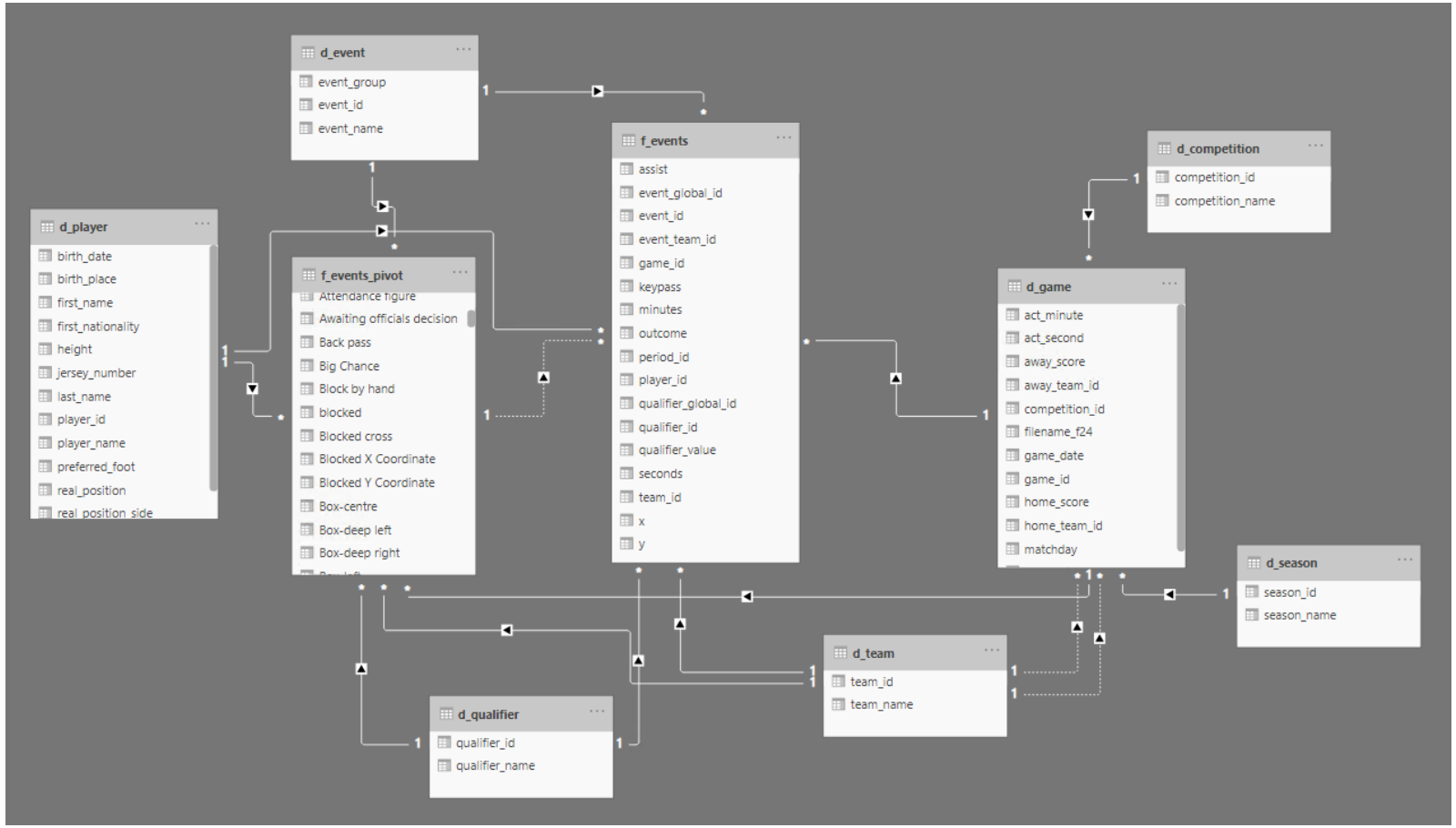
MODELO DE DATOS

_Tablas de Dimensiones

- D_Competition
- D_Season
- D_Game
- D_Team
- D_Player
- D_Position
- D_Stat
- D_Event
- D_Qualifier

_Tablas de Hechos

- F_Pass y F_Pass_Matrix (F27)
- F_Events y F_Events_Pivot (F24)
- F_Possession (F28)
- F_Season_Stats_Team y F_Season_Stats_Player (F30)
- F_Match_Results (F9)



LIMPIEZA, INGESTA Y PROCESAMIENTO DE DATOS

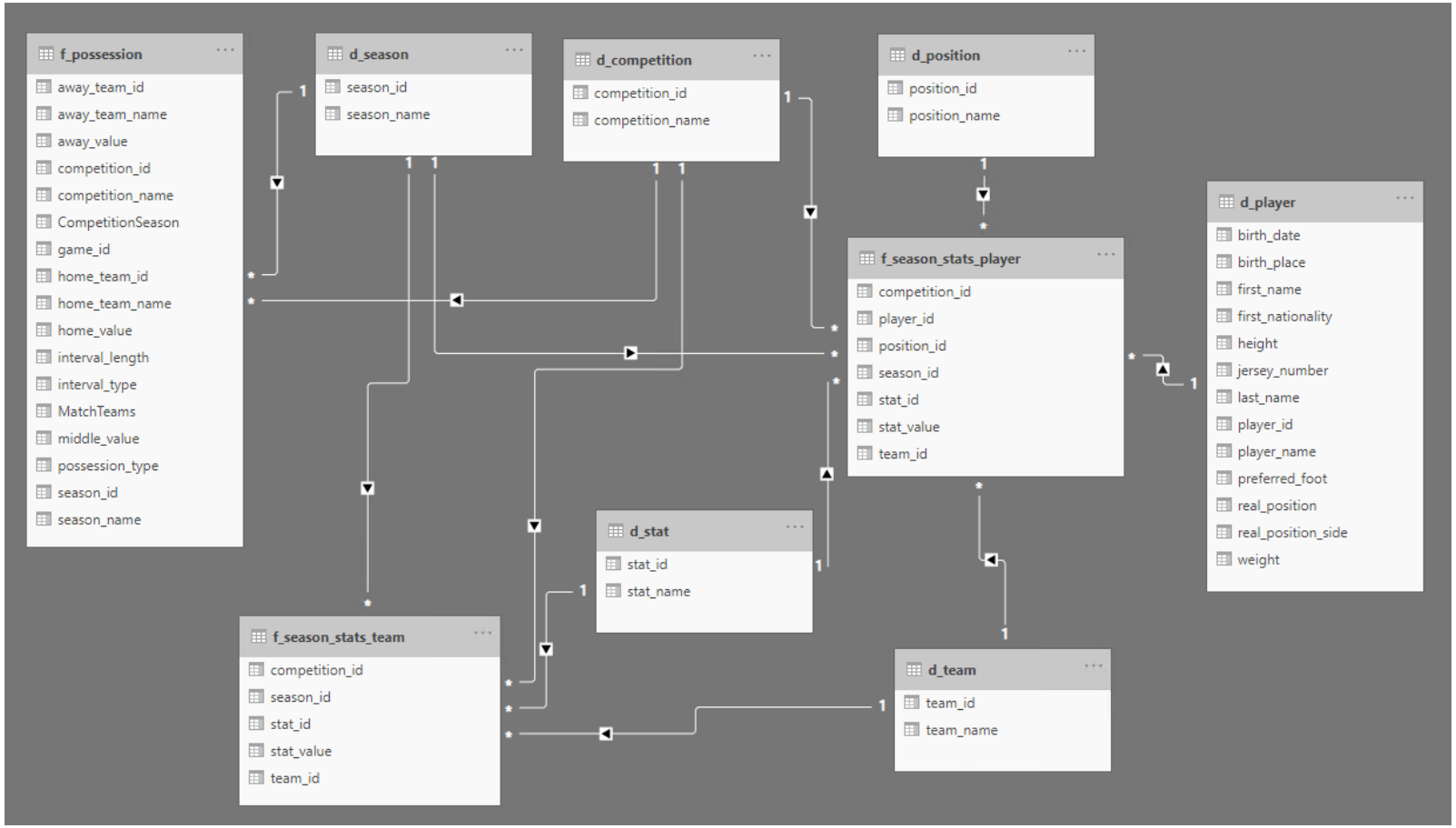
MODELO DE DATOS

_Tablas de Dimensiones

- D_Competition
- D_Season
- D_Game
- D_Team
- D_Player
- D_Position
- D_Stat
- D_Event
- D_Qualifier

_Tablas de Hechos

- F_Pass y F_Pass_Matrix (F27)
- F_Events y F_Events_Pivot (F24)
- F_Possession (F28)
- F_Season_Stats_Team y F_Season_Stats_Player (F30)
- F_Match_Results (F9)



VISUALIZACIÓN DE RESULTADOS

POWER BI (*Events.pbix*)

_Campograma con eventos

_Áreas convexas

Área convexa
con eventos seleccionados
(integración de R con Power BI)

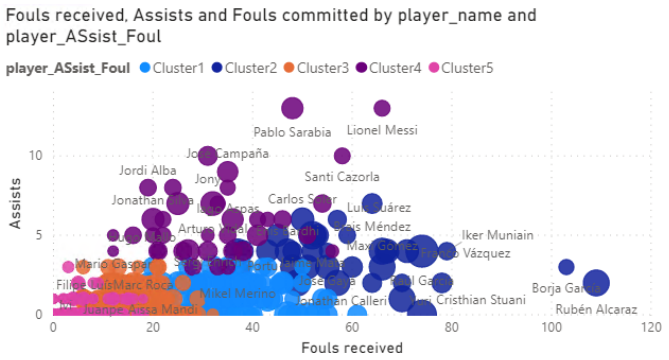
VISUALIZACIÓN DE RESULTADOS

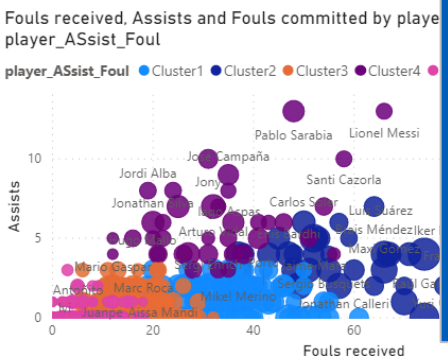
POWER BI (*Events.pbix*)

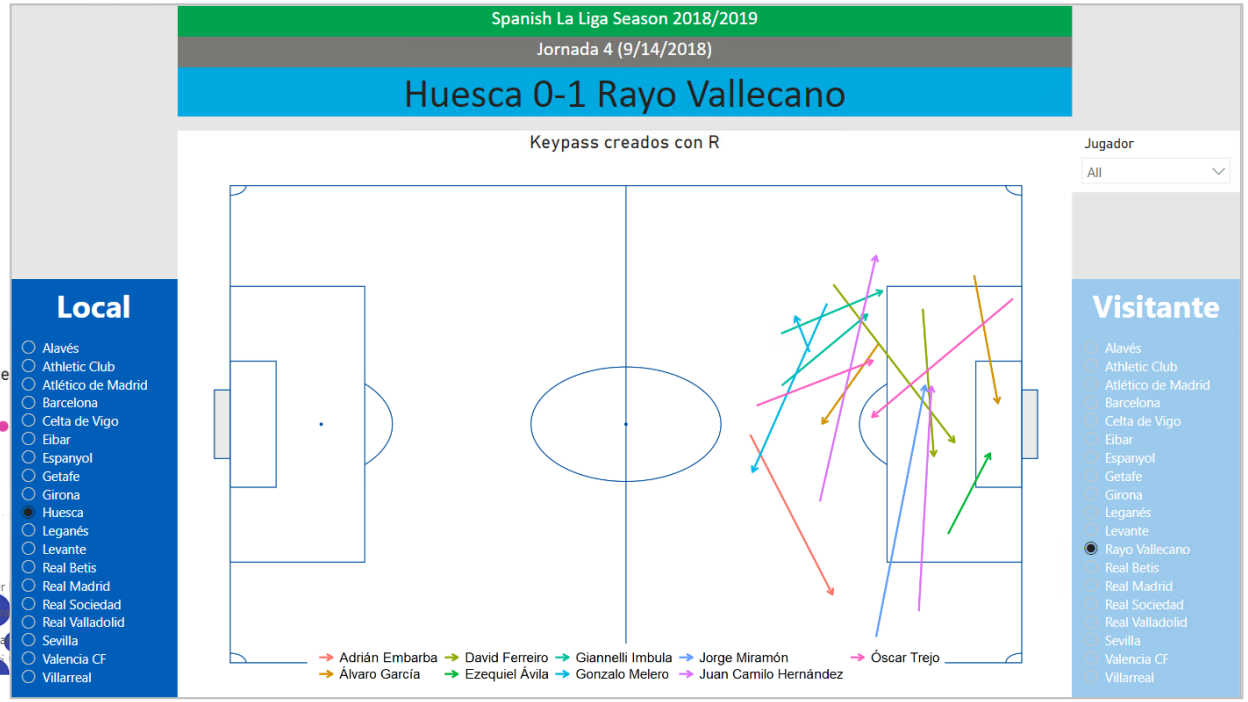
_ Pases

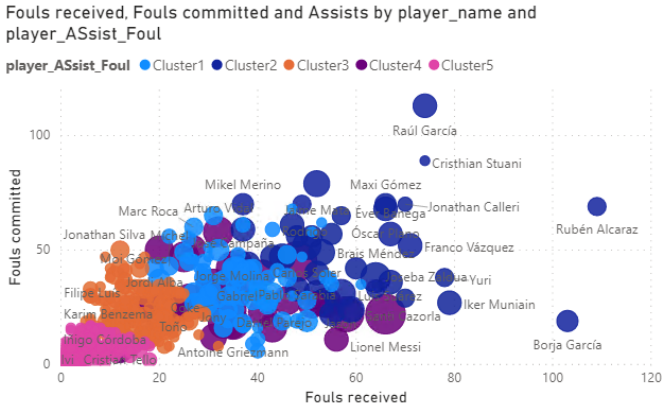
_ Duelos aéreos

_ Cluster de jugadores

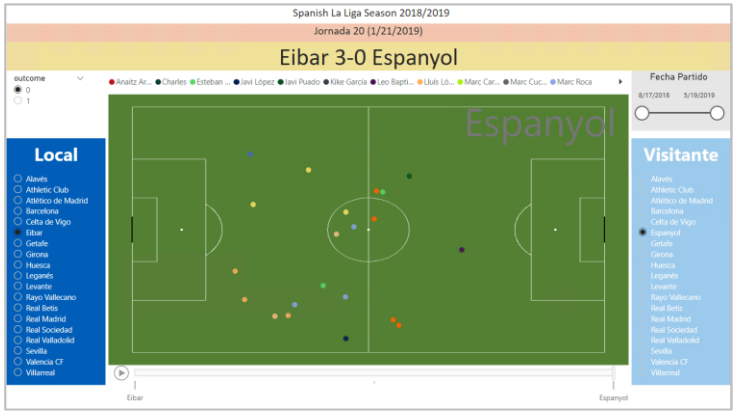








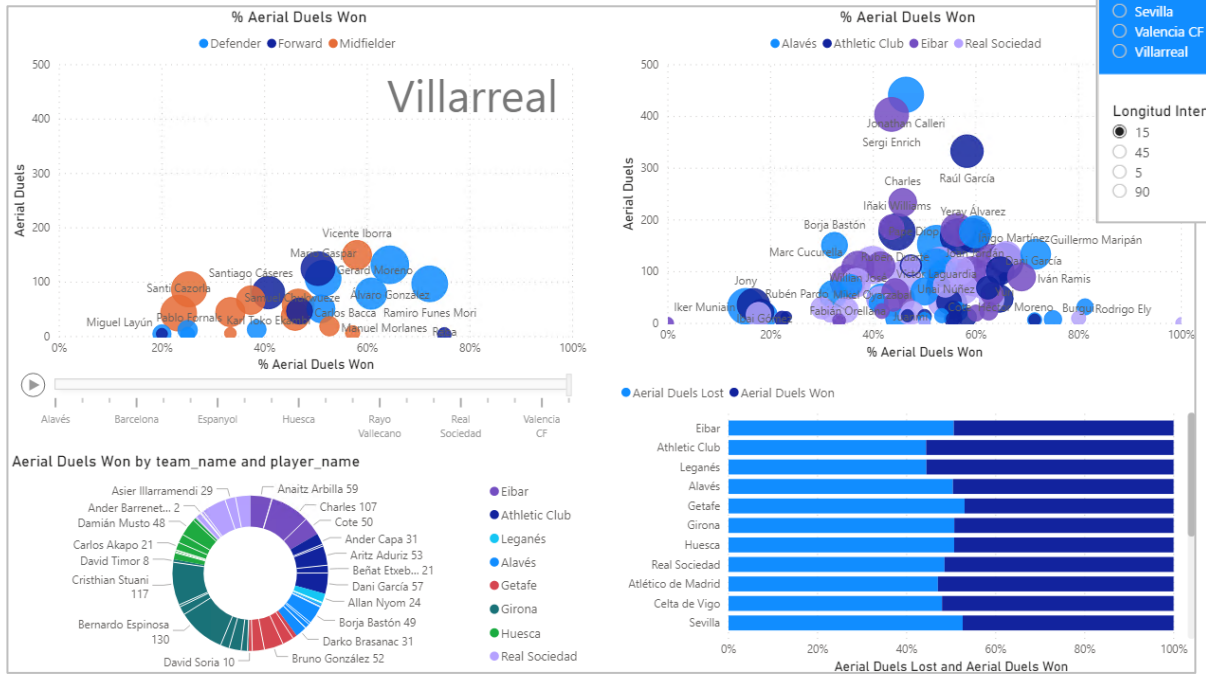
Player	Assists	Fouls committed	Fouls received	player_ASsist_Foul
Lionel Messi	13	22	66	Cluster4
Pablo Sarabia	13	45	48	Cluster4
Jony	10	34	31	Cluster4
Santi Cazorla	10	22	58	Cluster4
José Campaña	9	42	35	Cluster4
Antoine Griezmann	8	18	35	Cluster4
Jordi Alba	8	25	19	Cluster4
Wissam Ben Yedder	8	26	24	Cluster4
Arturo Vidal	7	58	32	Cluster4
Brais Méndez	7	38	64	Cluster2
Daniel Parejo	7	27	54	Cluster4
Moi Gómez	7	47	25	Cluster4
Sergi Roberto	7	18	33	Cluster4
Total	**615**	**10306**	**9734**	

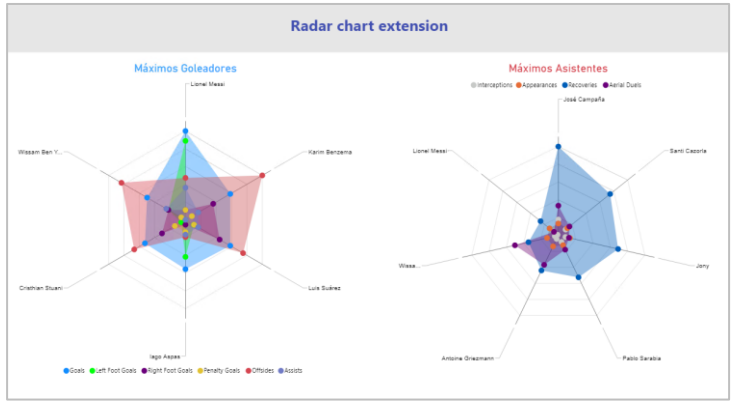


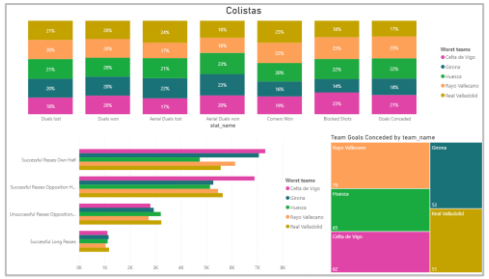
VISUALIZACIÓN DE RESULTADOS

POWER BI (*Stats & Possession.pbix*)

- Mapas de Posesiones
- Análisis de los duelos aéreos
- Gráficos de radar con estadísticas de los máximos goleadores y asistentes de La Liga
- Análisis de los últimos 5 clasificados





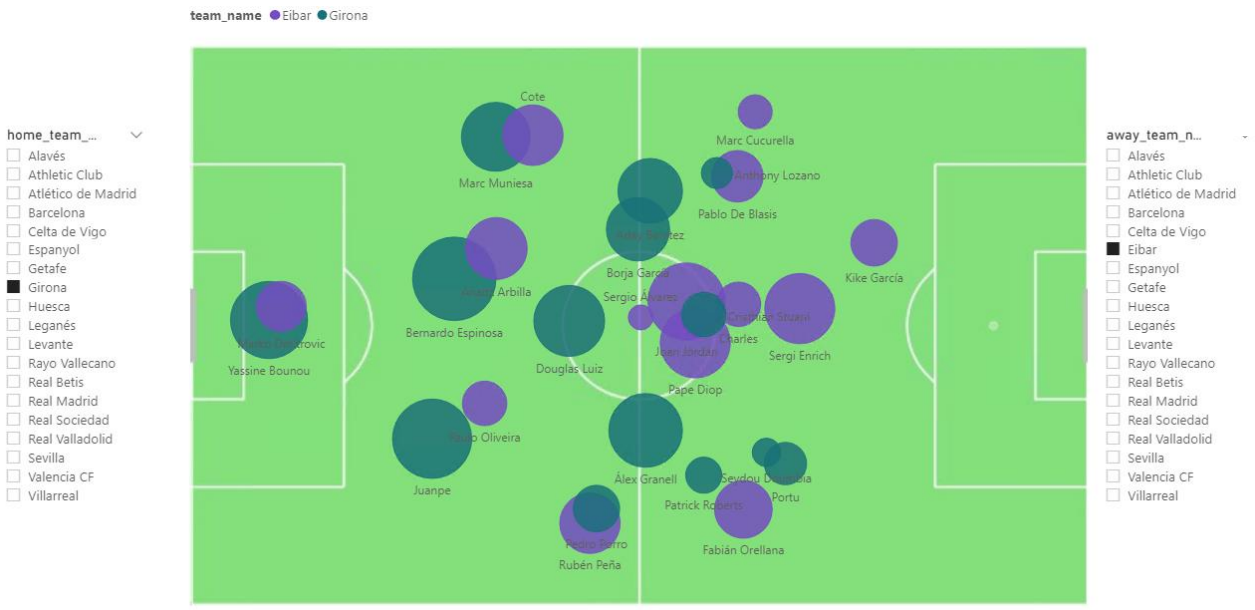


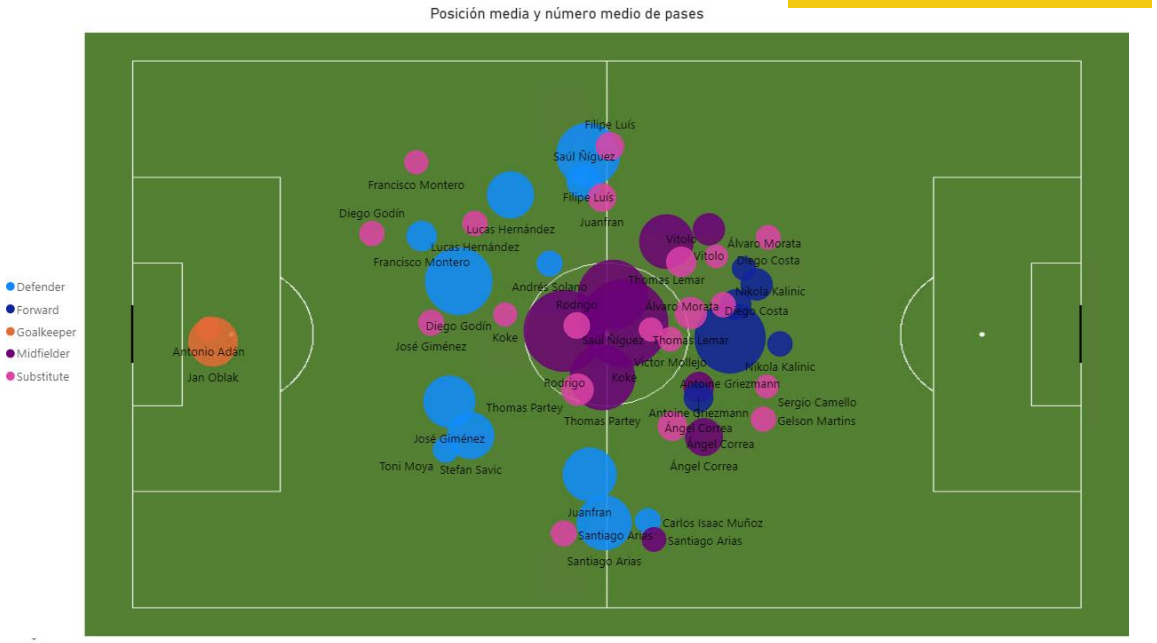
VISUALIZACIÓN DE RESULTADOS

POWER BI (*Pass Matrix.pbix*)

_ Matriz de pases

_ Posición media y número de pases





Pass Matrix

player_name	Arthur	Arturo Vidal	Coutinho	Gerard Piqué	Ivan Rakitic	Jordi Alba	Lionel Messi	Nélson Semedo	Sergi Roberto	Serg
Ivan Rakitic	100	83	115	223	2	255	241	139	277	
Sergio Busquets	169	153	133	217	253	165	283	92	138	
Gerard Piqué	115	115	39	2	274	62	86	210	265	
Jordi Alba	211	81	353	41	199		133	16	27	
Lionel Messi	85	116	147	42	167	135	2	67	127	
Sergi Roberto	42	132	37	161	248	18	218	54		
Arturo Vidal	49	1	31	139	78	77	178	127	80	
Clément Lenglet	63	34	68	137	141	217	11	24	16	
Coutinho	87	43		27	106	251	122	22	43	
Nélson Semedo	19	151	11	174	149	8	91		45	
Marc-André ter Stegen	41	35	21	208	60	68	23	37	69	
Luis Suárez	31	38	61	15	70	87	135	34	62	
Samuel Umtiti	23	39	67	97	81	112	10	12	37	
Ousmane Dembélé	23	45	50	15	70	93	99	42	26	
Malcom	11	29	5	3	8	13	14		24	

CONCLUSIONES Y LÍNEAS FUTURAS

CONCLUSIONES

_Optimización del modelo de datos

- Data mart
- Disponibilidad de cargar datos de otras temporadas y/o competiciones en un ‘único repositorio

_Procesos de carga

- Reducción de tareas manuales
- Pentaho Data Integration (PDI)
  - Fácil e intuitivo
  - Posibilidad de introducir mejoras

_Visualización

- Mejora notable al explotar la información estructurada procedente de la base de datos en vez de cargar ficheros XML directamente en Power BI

LÍNEAS FUTURAS

_Opción 1

- Extender y automatizar la interpretación de eventos del fichero F24
- Incorporar más OPTA feeds
  - F1-F75, especial hincapié en tracking (F53-55)

_Opción 2

- Integrar fuentes de datos procedentes de otros proveedores deportivos

¡Gracias!