PRA 1 VANESA NAVARRO ORONOZ

I. ANÁLISIS DE REQUERIMIENTOS

El análisis de requerimientos se basa en identificar las necesidades que tiene una organización particular respecto al análisis de la información. La información recopilada en esta fase permite definir a posteiori los procesos y perspectivas que hay que analizar.

Normalmente en esta fase, se debe ser previsor y pensar más allá de las necesidades actuales para poder cubrir las futuras. Es necesario comentar que, aunque para esta actividad el análisis de requisitos se fundamenta en el enunciado proporcionado y se complementa con las fuentes de datos, esta fase se debe hacer entrevistando al cliente y analizando las necesidades de informacion de la organización que se cubren en la actualidad y las que debe cubrir el proyecto a futuro. Por este motivo, como consultores del proyecto es conveniente hacer un diagnóstico y una conceptualización adecuada del proyecto. Es decir, identificar claramente los objetivos y el contenido del mismo.

La necesidad principal del caso es disponer de información integrada para su análisis y difusión. Esta información debe ayudar a los usuarios potenciales en la toma de decisiones y a alcanzar los objetivos deseados. Para ello, se diseñará un almacén de datos que permita la gestion de informacion para su analisis y difusion, a través de un proyecto que incluye la creación del propio almacén, el diseño e implementación de procesos ETL, el diseño e implementación de un modelo multidimensional OLAP y, por último, la construcción de las consultas establecidas posteriormente en este apartado.

Según el contexto del caso práctico, identificamos las siguientes necesidades que nuestro sistema deberá cubrir teniendo en cuenta la descripición que se ha hecho en el enunciado de sus usuarios potenciales:

- Conocer los datos estadísticos de las jugadoras de la WNBA y los jugadores de la NBA para analizar y mejorar su rendimiento.
- Conocer la evolución temporal de las jugadoras de la WNBA y los jugadores de la NBA. Esta evolución debe poderse analizar desde diferentes perspectivas:
  - Edad promedio por temporada
  - Puntos promedio por temporada
  - Partidos jugados por temporada
  - Porcentaje de tiros libres
  - Porcentaje de tiros de campo
- Determinar las universidades y estados de los cuáles salen más jugadores de la NBA para que seleccionadores y ojeadores prioricen esos lugares en sus viajes.
- Disponer datos de los mejores jugadores y jugadoras actuales para ser promocionados en nuevas campañas de marketing con los que el público pueda sentirse identificado.
- Realizar la comparativa entre los mejores y peores jugadores y sus posiciones en el juego.
- Garantizar la correcta gobernanza del dato.

Si se tiene en cuenta toda esta información, el sistema podrá responder a múltiples preguntas y de esta manera, cubrir las necesidades de los usuarios potenciales. De forma específica, se pide que el sistema sea como mínimo capaz de dar respuesta a las siguientes preguntas:

- Top 10 de los mejores tiradores de tiros libres de la historia.
- ¿Cuántos jugadores actuales de la NBA existen para cada universidad? ¿Y para cada isntituto?
- Top 3 estados que producen la mayoría de los jugadores de la NBA.
- Top 5 de jugadores de la NBA y top 5 de jugadoras de la WNBA que tienen el mejor porcentaje de tiros de campo.
- Evolución de las jugadoras de la WNBA y los jugadores de la NBA por temporada según diferentes métricas.
- Resumen de los mejores y peores jugadores según su ratio de eficacia y posición en el campo.
- ¿Hay mayoría de jugadores extranjeros o de EEUU en la NBA?
- ¿Cuándo se anotan más tiros libres, playoffs o partidos regulares?
- ¿En qué período del partido se anotan más tiros libres?

II. ANÁLISIS DE FUENTES DE DATOS

En este apartado se deben revisar las fuentes de datos proporcionadas, qué tipo de información contienen, cuál es su formato, y qué cantidad representan para la carga inicial (volumetría).

En los proyectos de almacenes de datos, es muy relevante analizar las fuentes de datos. Del análisis de las mismas puede desprenderse información clave para el éxito y la evolución de proyecto, así como la identificación de riesgos vinculados.

En un proyecto real, consensuariamos con las partes interesadas del proyecto para identificar si hay un gobierno de datos, calidad de éstos, en qué procesos participan…Como nuestro caso de estudio particular es un ejemplo, y no tenemos acceso a las partes interesadas, utilizaremos el sentido común para el análisis de fuentes de datos.

Para esta actividad, nos encontramos con varios ficheros obtenidos de diferentes repositorios de datos abiertos. En concreto, disponemos de información de equipos, jugadoras y jugadores o estadísticas de juego extraídas de las páginas oficiales de la NBA (stats.nba.com) y WNBA (stats.wnba.com), así como otros datos obtenidos de fuentes de datos abiertas (open data) disponibles en internet.

A continuación, se detalla la relación de ficheros que utilizaremos para la carga inicial del data warehouse agrupados según categoría (“***geo***”*-archivo geográfico, “****teams****”-archivos referentes a equipos de NBA y WNBA, “****players****”-archivos referentes a los jugadores de NBA y WNBA, “****stats****”-archivos con estadisticas de NBA y WNBA, “****only nba****”-archivos que solo tienen datos de la NBA*):

CATEGORIA	NOMBRE ARCHIVO	DESCRIPCIÓN DEL ARCHIVO	FUENTE
GEO	Estados_Unidos.json	Información de los estados de los EE. UU. donde se juegan partidos de la NBA y WNBA.	https://es.wikipedia.org/wiki/Estado_de_los_Estados_Unidos
TEAMS	TeamCodes.txt	Códigos de los equipos de la NBA y WNBA	
	NBA_Teams.xml	Equipos de la NBA por conferencia y división.	https://en.wikipedia.org/wiki/National_Basketball_Association
	WNBA_Teams.xls	Equipos de la WNBA por conferencia	https://en.wikipedia.org/wiki/Women%27s_National_Basketball_Association
PLAYERS	NBA_Players_List.xls	Lista de jugadores (NBA) en activo y retirados	https://www.nba.com/players
	WNBA_Players_List.txt	Lista de jugadoras (WNBA) en activo y retiradas	https://stats.wnba.com/players/list/?Historic=Y
STATS	WNBA_Seasons_Stats_2005_2017.xls	Estadísticas de juego de las jugadoras de la WNBA. Datos desde la temporada 2005 hasta 2017	https://stats.wnba.com/players/traditional/?sort=PTS&dir=-1
	NBA_Seasons_Stats_1950_2017.csv	Estadísticas de juego de los jugadores de la NBA. Datos desde la temporada 1950 hasta 2017	https://www.kaggle.com/drgilermo/nba-players-stats
ONLY NBA	NBA_players_data.json	Información personal de los jugadores y estadísticas globales de su carrera deportiva.	https://data.world/search?context=community&page=5&q=NBA&type=all
	nba_free_throws.csv	Información detallada sobre tiros libres (2006-2016).	https://www.kaggle.com/sebastianmantey/nba-free-throws

Otra aproximación a los datos es clasificarlos según su el formato inicial de los archivos, eso nos da una idea de los procesos que tendremos que implementar durante la carga al almacén:

FORMATO	DESCRIPCION FORMATO	ARCHIVOS INCLUIDOS
CSV	Valores separados por comas	nba_free_throws.csv
		NBA_Seasons_Stats_1950_2017.csv
JSON	JavaScript Object Notation	Estados_Unidos.json
		NBA_players_data.json
TXT	Texto simple, plano o sin formato	TeamCodes.txt
		WNBA_Players_List.txt
XML	Lenguaje de Marcas Extensibles	NBA_Teams.xml
XLS	Documento nativo Microsoft Excel	NBA_Players_List.xls
		WNBA_Seasons_Stats_2005_2017.xls
		WNBA_Teams.xls

Tenemos un total de diez ficheros de los cuales dos son archivos CSV, dos son JSON, dos archivos de texto TXT, uno es XML y tres son documentos excel XML.

Seguidamente, pasaremos a analizar el contenido de cada fichero individualmente. Los objetivos a analizar de cada uno de los ficheros son:

- Identificar los campos de datos contenidos (número de atributos).
- Identificar el separador de campos (si procede).
- Identificar el tipo de campo.
- Detallar el número total de registros en el fichero.
- Señalar si el fichero posee una cabecera con las etiquetas de los campos.
- Identificar los campos de datos no relevantes y que se no se cargarán en el almacén.
- Describir brevemente el campo y mostrar un ejemplo.

nba_free_throws.csv

Información detallada sobre tiros libres: cuándo se lanzó el tiro libre durante el juego, tirador, acierto/fallo, etc. (Temporadas 2006 a 2016).

- Tipo de fichero: CSV (Archivo de valores separado por comas)
- Primera línea con etiquetas de los campos.
- Separador de campos: “,“
- Total de registros: 618019
- Total de atributos: 11

Campo	Descripción	Tipo	Ejemplo	Carga al DW
end_result	Resultado final	Texto	106 - 114	SI
game	Partido	Texto	PHX - LAL	SI
game_id	ID Partido	Numérico	261031013	SI
period	Periodo	Numérico	1	SI
play	Comentario tiro	Texto	Andrew Bynum makes free throw 1 of 2	NO
player	Jugador	Texto	Andrew Bynum	SI
playoffs	Regular o Playoff	Texto	regular	SI
score	Marcador	Texto	0 - 1	SI
season	Temporada	Texto	2006 - 2007	SI
shot_made	Si mete 1, si falla 0	Numérico	1	SI
time	Tiempo	Time	11:45	SI

No se cargará el comentario porque ya esta si se hace el tiro libre o si se falla en el campo “shot_made”.

NBA_Seasons_Stats_1950_2017.csv

Estadísticas de juego (totales de la temporada) de los jugadores de la NBA. Datos desde la temporada 1950 hasta 2017. El conjunto de datos contiene estadísticas individuales agregadas para 67 temporadas de la NBA. desde atributos básicos de puntuación de caja, como puntos, asistencias, rebotes, etc., hasta funciones más avanzadas similares a bolas de dinero, como Valor sobre reemplazo.

- Tipo de fichero: CSV (Archivo de valores separado por comas)
- Primera línea con etiquetas de los campos.
- Separador de campos: “,“
- Total de registros: 24691
- Total de atributos: 52

Campo	Descripción	Tipo	Ejemplo	Carga al DW
Year	Temporada	Numérico	2017	SI
Player	Jugador	Texto	Ivica Zubac	SI
Pos	Posición jugador	Texto	C	SI
Age	Edad	Numérico	19	SI
Tm	Equipo	Texto	LAL	SI
G	Games (Played)	Numérico	38	SI
GS	Games Started	Numérico	11	SI
MP	Minutes Played	Numérico	609	SI
PER	Player Efficiency Rating	Numérico	17	SI
TS%	True Shooting %	Numérico	0.547	SI
3PAr	3-Point Attempt Rate	Numérico	0.013	SI
FTr	Free Throw Rate	Numérico	0.206	SI
ORB%	Offensive Rebound Percentage	Numérico	7.1	SI
DRB%	Defensive Rebound Percentage	Numérico	21.9	SI
TRB%	Total Rebound Percentage	Numérico	14.3	SI
AST%	Assist Percentage	Numérico	8.1	SI
STL%	Steal Percentage	Numérico	1.1	SI
BLK%	Block Percentage	Numérico	4.4	SI
TOV%	Turnover Percentage	Numérico	10.4	SI
USG%	Usage Percentage	Numérico	20.3	SI
blanl	En blanco			NO
OWS	Offensive Win Shares	Numérico	0.6	SI
DWS	Defensive Win Shares	Numérico	0.5	SI
WS	Win Shares	Numérico	1.1	SI
WS/48	Win Shares Per 48 Minutes	Numérico	0.086	SI
blank2	En blanco			NO
OBPM	Offensive Box Plus/Minus	Numérico	-2.7	SI
DBPM	Defensive Box Plus/Minus	Numérico	0.3	SI
BPM	Box Plus/Minus	Numérico	-2.5	SI
VORP	Value Over Replacement	Numérico	-0.1	SI
FG	Field Goals (Made)	Numérico	126	SI

FGA	Field Goal Attempts	Numérico	238	SI
FG%	Field Goal Percentage	Numérico	0.529	SI
3P	3-Point Field Goals (Made)	Numérico	0	SI
3PA	3-Point Field Goal Attempts	Numérico	3	SI
3P%	3-Point Field Goal Percentage	Numérico	0	SI
2P	2-Point Field Goals (Made)	Numérico	126	SI
2PA	2-Point Field Goal Attempts	Numérico	235	SI
2P%	2-Point Field Goal Percentage	Numérico	0.536	SI
eFG%	Effective Field Goal Percentage	Numérico	0.529	SI
FT	Free Throws (Made)	Numérico	32	SI
FTA	Free Throw Attempts	Numérico	49	SI
FT%	Free Throw Percentage	Numérico	0.653	SI
ORB	Offensive Rebounds	Numérico	41	SI
DRB	Defensive Rebounds	Numérico	118	SI
TRB	Total Rebounds	Numérico	159	SI
AST	Assists	Numérico	30	SI
STL	Steals	Numérico	14	SI
BLK	Blocks	Numérico	33	SI
TOV	Turnovers	Numérico	30	SI
PF	Personal Fouls	Numérico	66	SI
PTS	Points	Numérico	284	SI

Hay dos campos en blanco que no se cargarán.

TeamCodes.txt

Códigos de los equipos de la NBA y WNBA.

- Tipo de fichero: TXT (Texto plano)
- Primera línea con etiquetas de los campos.
- Separador de campos: “#“
- Total de registros: 42
- Total de atributos: 3

Campo	Descripción	Tipo	Ejemplo	Carga al DW
League	Liga	Texto	WNBA	SI
Team	Nombre del equipo	Texto	Atlanta Dream	SI
Code	Código del equipo	Texto	ADR	SI

WNBA_Players_List.txt

Lista de jugadoras (WNBA) en activo y retiradas.

- Tipo de fichero: TXT (Texto plano)
- Primera línea con etiquetas de los campos.
- Separador de campos: TAB
- Total de registros: 963
- Total de atributos: 2

Campo	Descripción	Tipo	Ejemplo	Carga al DW
Player	Nombre jugadora	Texto	Abdi, Farhiya	SI
Active	Si está activa o no	Texto	NO	SI

NBA_Teams.xml

Equipos de la NBA por conferencia y división.

- Tipo de fichero: XML (Extensible Markup Language)
- Total de registros: 30
- Total de atributos: 8

Campo	Descripción	Tipo	Ejemplo	Carga al DW
Conferencia	Conferencia	Texto	Conferencia Oeste	SI
División	División	Texto	Noroeste	SI
Equipo	Nombre Equipo	Texto	Denver Nuggets	SI
Ciudad	Ciudad	Texto	Denver	SI
Estado	Estado	Texto	CO	SI
Pabellón	Pabellón	Texto	Pepsi Center	SI
Fundado	Fundado	Numérico	1967	SI
Patrocinio	Patrocinio	Texto	Western Union	SI

WNBA_Teams.xls

Equipos de la WNBA por conferencia.

- Tipo de fichero: XLS (Hoja de cálculo de Microsoft Excel)
- Total de registros: 12
- Total de atributos: 10

Campo	Descripción	Tipo	Ejemplo	Carga al DW
Conferencia	Conferencia	Texto	Conferencia Este	SI
Equipo	Equipo	Texto	Indiana Fever	SI
Ptos	Puntos clasificación	Numérico	6	NO
J	Partidos jugados	Numérico	3	NO
G	Ganados	Numérico	3	NO
P	Perdidos	Numérico	0	NO
p.	Puntos a favor	Numérico	216	NO
c.	Puntos en contra	Numérico	189	NO
dif	Diferencia de puntos	Numérico	27	NO
Ciudad, Estado	Ciudad, Estado	Texto	Indianapolis, Indiana	SI

Solo se cargará el nombre del equipo, conferencia y ciudad, estado.

NBA_Players_List.xls

Lista de jugadores (NBA) en activo y retirados. El fichero agrupa los nombres por la inicial y contiene una fila (subcabecera) con cada una de las letras.

- Tipo de fichero: XLS (Hoja de cálculo de Microsoft Excel)
- Total de registros: 4995
  - Hoja “Players” = 525
  - Hoja “Historic” = 4470

Campo	Descripción	Tipo	Ejemplo	Carga al DW
Jugador	Nombre y apellido	Texto	Adams, Steven	SI

WNBA_Season_Stats_2005_2017.xls

Estadísticas de juego (totales de la temporada), de las jugadoras de la WNBA. Datos desde la temporada 2005 hasta 2017.

- Tipo de fichero: XLS (Hoja de cálculo de Microsoft Excel)
- Primera línea con etiquetas de los campos.
- Total de registros: 64635
- Total de atributos: 4

Estos datos habrá que transponer posteriormente porque todas las estadísticas se encuentran bajo el atributo “stat” y su valores en “data”, todo agrupado por ID.

Campo	Descripción	Tipo	Ejemplo	Carga al DW
ID	ID	Numérico	20171	SI
Data	Datos	Texto	2017	SI
Type	Tipo de datos	Texto	Number	SI
Stat	Nombre Estadisticas	Texto	SEASON	SI

Estados_Unidos.json

Información de los estados de los EE. UU. donde se juegan partidos de la NBA y WNBA.

- Tipo de fichero: JSON (JavaScript Object Notation)
- Total de registros: 50
- Total de atributos: 9

Campo	Descripción	Tipo	Ejemplo	Carga al DW
Estado	Estado	Texto	Alabama	SI
Nombre oficial	Nombre oficial	Texto	State of Alabama	NO
Superficie (km2)	Superficie (km2)	Texto	135756	NO
Abrev.	Abrev.	Texto	AL	SI
Ingreso en la Unión	Ingreso en la Unión	Texto	14-12-1819	NO
Población (2010)	Población (2010)	Texto	4779736	NO
Densidad de Población (hab/km2)	Densidad de Población (hab/km2)	Texto	35.21	NO

Capital	Capital	Texto	Montgomery	NO
Ciudad con mayor Población (2006)	Ciudad con mayor Población (2006)	Texto	Birmingham	NO

Solo se cargará el nombre del estado y su abreviatura.

NBA_players_data.json

Información personal de los jugadores (fecha y lugar de nacimiento, peso, altura, etc.) y estadísticas globales de su carrera deportiva.

- Tipo de fichero: JSON (JavaScript Object Notation)
- Total de registros: 4639
- Total de atributos: 26

Campo	Descripción	Tipo	Ejemplo	Carga al DW
id	ID	Texto	"abdelal01"	SI
birthDate	Fecha Nacimiento	Texto	"06/24/1968"	NO
birthDate_str	Fecha Nacimiento	Texto	"24/06/1968"	SI
birthPlace	Lugar Nacimiento	Texto	"Cairo"	SI
State_Country	Pais Nacimiento	Texto	"Egypt"	SI
career_AST	Career Assists	Texto	"3"	SI
career_FG%	Career Field Goal Percentage	Texto	"502"	SI
career_FG3%	Career 3-Point Field Goal Percentage	Texto	"0.0"	SI
career_FT%	Career Free Throw Percentage	Texto	"70.1"	SI
career_G	Career Games (Played)	Texto	"256"	SI
career_PER	Career Player Efficiency Rating	Texto	"13.0"	SI
career_PTS	Career Points	Texto	"57"	SI
career_TRB	Career Total Rebounds	Texto	"3.3"	SI
career_WS	Career Win Shares	Texto	"48"	SI
career_eFG%	Career Effective Field Goal Percentage	Texto	"502"	SI
college	Universidad	Texto	"Duke University"	SI
draft_pick	Puesto Draft	Texto	"25th overall"	NO
draft_round	Ronda Draft	Texto	"1st round"	NO
draft_team	Equipo Draft	Texto	"Portland Trail Blazers"	NO
draft_year	Año Draft	Texto	"1990"	NO
height	Altura	Texto	"6-10"	SI
highSchool	Instituto	Texto	"Bloomfield in Bloomfield, New Jersey"	SI
name	Nombre jugador	Texto	"Alaa Abdelnaby"	SI
position	Posicion	Texto	"Power Forward"	SI
shoots	Tirador	Texto	"Right"	SI
weight	Peso	Texto	"240lb"	SI

No se cargarán los datos referentes al “draft” ni uno de los campos que tiene la fecha de nacimiento, porque se repite con otro formato.

Estimación de volumetría

En los proyectos de diseño de factoría de información corporativa existe una primera fase en la que se realiza una carga inicial, y a posteriori, una segunda fase para realizar las cargas incrementales de los datos nuevos que nos van llegado.

Una estimación del volumen de datos de nuestro almacén para la carga de los datos que disponemos sería:

Fuente de datos	Datos
NBA_Teams.xml 1 fichero anual	30 registros x 8 valores = **240 datos**
WNBA_Players_List.txt 1 fichero anual	963 registros x 2 valores = **1926 datos**
nba_free_throws.csv 1 fichero anual	618019 registros x 10 valores = **6180190 datos**
NBA_Seasons_Stats_1950_2017.csv 1 fichero anual	24691 registros x 50 valores = **1234550 datos**
TeamCodes.txt 1 fichero anual	42 registros x 3 valores = **126 datos**
WNBA_Teams.xls 1 fichero anual	12 registros x 3 valores = **36 datos**
NBA_Players_List.xls 1 fichero anual	4995 registros x 1 valores **= 4995 datos**
WNBA_Season_Stats_2005_2017.xls 1 fichero anual	64635 registros x 4 valores = **258540 datos**
Estados_Unidos.json 1 fichero anual	50 registros x 2 valores = **100 datos**
NBA_players_data.json 1 fichero anual	4639 registros x 21 valores = **97419 datos**
TOTAL DATOS	**7.778.122**

III. ANÁLISIS FUNCIONAL

En el análisis funcional, se debe proponer el tipo de arquitectura para la factoría de información que mejor se adecue al caso de estudio. En el momento de considerar los requisitos funcionales, es necesario tener encuenta que cada requisito tendrá una prioridad asociada y podrá ser exigible (E) o deseable (D).

En el contexto de esta actividad, los requerimientos exigibles son aquellos que demanda el enunciado, y los deseables son aquellos que complementan la actividad. Por otro lado, en términos de la escala de prioridades, asignamos una prioridad de 1 a 3, siendo 1 completamente prioritario para la actividad y 3 no prioritario para la actividad.

En la tabla, se describen los requerimientos funcionales para el diseño de una factoría de información teniendo en cuenta las consideraciones del caso práctico que estamos desarrollando:

#	**Requerimiento**	**Prioridad**	**Exigible / deseable**
1	Se extraerá de forma adecuada la información de las fuentes de datos (considerando solo la información relevante).	1	E
2	Se creará un almacén de datos con datos de la NBA y de la WNBA con fines de análisis estadísticos.	1	E
3	Se cargará la información en el almacén de datos mediante procesos ETL considerando una *staging area*.	1	E
4	Se creará un modelo OLAP para consultas multidimensionales pre-calculadas que satisfagan los requerimientos de los usuarios.	2	E
5	Se redactará un manual de carga de datos incremental.	3	D
6	Se crearán herramientas de visualización adecuadas a los requerimientos.	3	D

La estrategia de construcción que se propone para la resolución del caso es un enfoque basado en “FIC con Staging Area” por los siguientes motivos:

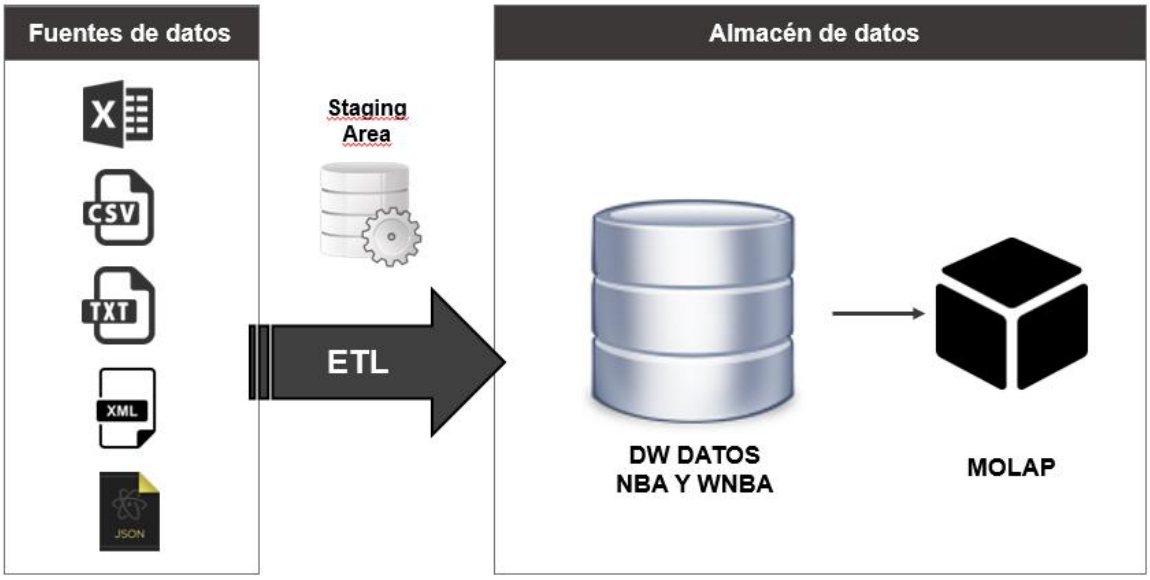
- El uso de una Staging Area simplifica el proceso de extracción del componente de integración de la FIC.
- Tenemos diferentes ficheros de entrada desde múltiples orígenes heterogéneos que es conveniente consolidar en una estructura de carga intermedia.
- Es necesario homogeneizar estos datos antes de su carga.
- Es necesatio realizar procesos de limpieza de los datos que mejoren su calidad.

Es habitual implementar la Staging Area con una base de datos, aunque también podría utilizarse un conjunto de archivos temporales como zona de trabajo. En este caso, usaremos los archivos temporales o intermedios.

En términos de la arquitectura funcional tenemos los siguientes elementos:

1. Las fuentes de datos están compuestas, como ya se ha analizado en el apartado anterior, por ficheros planos de texto TXT, ficheros de valores separados por comas CSV, ficheros XML, ficheros JSON y ficheros de Microsoft Excel, conteniedo información estadística y descriptiva sobre los jugadores y jugadoras de la NBA y WNBA respectivamente tanto en un contexto histórico como actual.
2. La arquitectura de la FIC está formada por los siguientes componentes que estarán alojados en la misma máquina:
   - **Staging area**: este componente y su función han sido explicado en las lineas anteriores.
   - **Almacén de datos**: será el repositorio donde se carguen los datos obtenidos de la NBA y WNBA.
   - OLAP: existen diferentes tipos de OLAP en función a la forma en que se almacenan los datos. En nuestro caso, el componente será **MOLAP (Multidimensional OLAP)**, que es la forma clásica de OLAP. MOLAP utiliza estructuras de bases de datos generalmente optimizadas para la recuperación de los mismos. Lo que se conoce como bases de datos multidimensionales (o más coloquialmente **cubos**). En definitiva, se crea un fichero que contiene todas las posibles consultas pre-calculadas. A diferencia de las bases de datos relacionales, estas formas de almacenaje están optimizadas para la velocidad de cálculo. A partir de la información cargada en el almacén de datos, se creará cubo multidimensional.

El siguiente gráfico resume los elementos de la arquitectura para esta actividad:



IV. DISEÑO DEL MODELO CONCEPTUAL, LÓGICO Y FÍSICO DEL ALMACÉN DE DATOS

Diseño Conceptual

Para el correcto desarrollo del DW es preciso definir los hechos (facts), dimensiones de análisis (dimensions), las métricas y los atributos que nos permitan tener el nivel de granularidad suficiente para presentación de los objetivos que se han definido en el análisis de requerimientos y de las fuentes de datos.

A partir de estos análisis se han identificado los siguientes tablas de hechos con sus dimensiones asociadas:

Tabla de hecho	Descripción
FACT_SEASONS_STATS	Estadísticas de juego, totales por temporada, de los jugadores de la NBA y jugadoras de la WNBA

Dimensiones	Descripción
Jugadores	Información personal de los jugadores cuyas estadísticas se quieren analizar.
Temporada	Temporada cuyas estadísticas se quieren analizar.
Liga	Definición de la liga a analizar NBA o WNBA.

A partir de las dimensiones y la tabla de hechos identificados, se construye el modelo conceptual, siendo tanto las dimensiones como los hechos, entidades independientes que forman parte de nuestro modelo de estrella.

El diseño conceptual para esta tabla de hechos y sus dimensiones es:

La siguiente tabla de hechos identificada y sus dimensiones:

Tabla de hecho	Descripción
FACT_PLAYER_CHAR	Caracterísitcas educativas, culturales y demográficas de los jugadores de la NBA.

Dimensiones	Descripción
Jugadores	Información personal de los jugadores cuyas estadísticas se quieren analizar.
Educacion	Información sobre los institutos y universidades de los jugadores.
Lugar de origen	Datos identificativos de la población, estado y país del jugador.

El diseño conceptual para esta tabla de hechos y sus dimensiones es:

El último hecho hecho que se ha identificado es:

Tabla de hecho	Descripción
FACT_FREE_THROWS_ANALYSIS	Estadísticas del análisis de los tiros libres, totales por temporada, de los jugadores de la NBA

Y sus dimensiones:

Dimensiones	Descripción
Temporada	Información sobre la temporada donde se realiza el tiro.
Jugadores	Información personal sobre los jugadores.
Game	Carcaterísticas del partido en el que se anota el tiro.

Veamos a continuación el diagrama del modelo conceptual para este caso:

Diseño Lógico

Para cada dimensión se han determinado sus atributos y para las tablas de hechos, las principales métricas. De nuevo, se debe tener en cuenta la consideración anterior que detalla las dimensiones de cada tabla de hecho. En este apartado vamos a detallar los atributos de las dimensiones y las métricas de las tablas de hechos con el objetivo de diseñar el modelo lógico.

A continuación, se muestra un resumen de las tablas de hechos y las métricas identificadas para ellos:

Tablas de Hechos	Métricas	Descripción
FACT_SEASONS_STATS	GP	Partidos jugados
	AGE	Edad
	PTS	Puntos
	FT%	Porcentaje Tiros Libres
	FG%	Porcentaje Tiros Campo
	PER	Player Efficiency Rating
FACT_PLAYER_CHAR	N_JUGADORES	Número de jugadores
	PERCENT_EXTRANJ	Porcentaje de jugadores extranjeros
FACT_FREE_THROWS_ANALYSIS	N_TF	Número de tiros libres anotados
	PERCENT_TF	Porcentaje de tiros libres anotados

A continuación, se procederá a detallar los atributos que contiene cada tabla de dimensiones. Los atributos junto con las métricas permitirán realizar los diferentes análisis de los requerimientos planteados.

Para FACT_SEASONS_STATS:

Dimensiones	Atributos descriptores
DIM_Jugadores	Código, nombre, posición juego, equipo
DIM_Temporada	Temporada
DIM_Liga	Nombre de la liga

Para FACT_PLAYER_CHAR:

Dimensiones	Atributos descriptores
DIM_Jugadores	Código, nombre, posición juego, equipo
DIM_Educacion	High school, College
DIM_Lugar_Origen	Birth place, State, Country

Para FACT_FREE_THROW_ANALYSIS:

Dimensiones	Atributos descriptores
DIM_Temporada	Temporada
DIM_Jugadores	Código, nombre, posición juego, equipo
DIM_Game	ID_Game, playoffs, period

Diseño Físico

Para el correcto diseño físico del almacén debemos tener en cuenta diversos aspectos:

- El tipo de base de datos con el que trabajemos, puesto que cada una de ellas tiene su particularidad.
- El diseño físico debe estar orientado a generar un buen rendimiento en el procesamiento de consultas.
- La definición de los procesos de administración del DW.
- La revisión periódica del diseño físico inicial para validar que continúa dando respuesta a las necesidades del cliente.

Una vez determinados qué tablas de hechos, dimensiones, métricas y atributos existen en nuestro modelo, podemos determinar las claves foráneas que debe definirse en el modelo físico.

En este paso también es necesario tener en cuenta el tamaño adecuado de los atributos (por ejemplo, qué longitud tiene una cadena o si los numéricos contienen decimales). También es relevante acordarse de crear correctamente las claves primarias en las dimensiones.

Dado que nuestro modelo de almacén está compuesto de más de una tabla de hechos, también debemos revisar las dimensiones que hemos definido en el diseño conceptual y lógico de cada hecho aplicando una visión conjunta del modelo. Esto nos permitirá definir dimensiones comunes, como año o tipo de universidad y así simplificar el modelo final y conseguir un rendimiento óptimo en la ejecución de los análisis.

Como es lógico, primero se crean las tablas de dimensiones y posteriormente las tablas de hechos. De esta forma creamos cada una de las tablas de nuestro almacén de datos.

Las dimensiones del modelo podrán ser referenciadas en las tablas de hechos utilizando sus claves primarias. El modelo físico de las tablas de hechos consistirá en la creación de las tablas cuyos campos serán claves foráneas a las dimensiones del modelo estrella del diseño lógico y de las métricas.

PRIMER CASO HECHO-DIMENSIONES

Dimensiones:

DIM_Jugadores

Nombre campo	Tipo	Tamaño
id_jugador	Número	4
nombre_jugador	Texto	255
posicion	Texto	100
equipo	Texto	255

DIM_Liga

Nombre campo	Tipo	Tamaño
id_liga	Número	4
desc_liga	Texto	20

DIM_Temporada

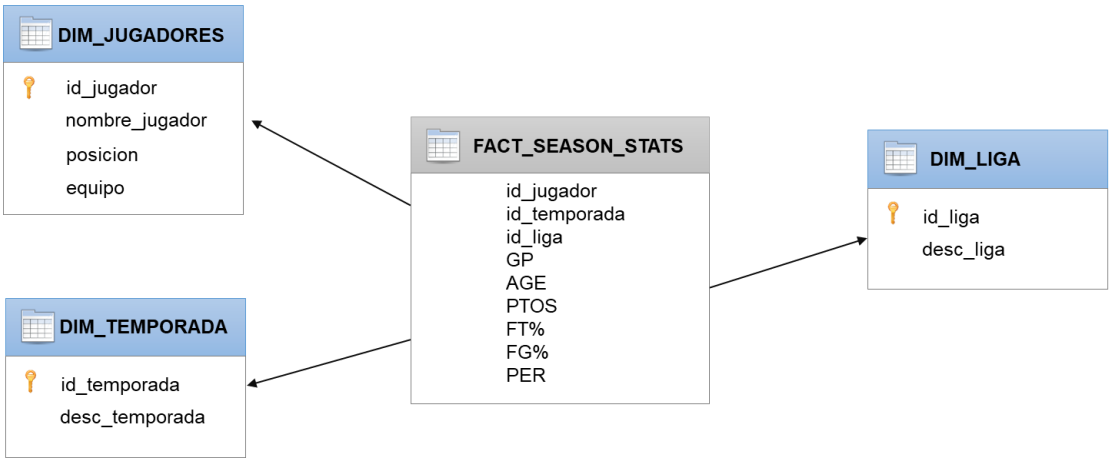
Nombre campo	Tipo	Tamaño
id_temporada	Número	4
desc_temporada	Texto	20

Hecho:

FACT_SEASON_STATS

Nombre campo	Tipo	Tamaño
id_jugador	Número	4
id_temporada	Número	4
id_liga	Número	4
GP	Número decimal	8
AGE	Número	4
PTS	Número decimal	8
FT%	Número decimal	8
FG%	Número decimal	8
PER	Número decimal	8

El esquema resultante es el siguiente:



SEGUNDO CASO HECHO-DIMENSIONES

Dimensiones:

DIM_Educacion

Nombre campo	Tipo	Tamaño
id_educacion	Número	4
high_school	Texto	255
high_school_state	Texto	255
college	Texto	255

DIM_Lugar_Origen

Nombre campo	Tipo	Tamaño
id_lugar	Número	4
birth_place	Texto	255
state	Texto	255
country	Texto	255

DIM_Jugadores

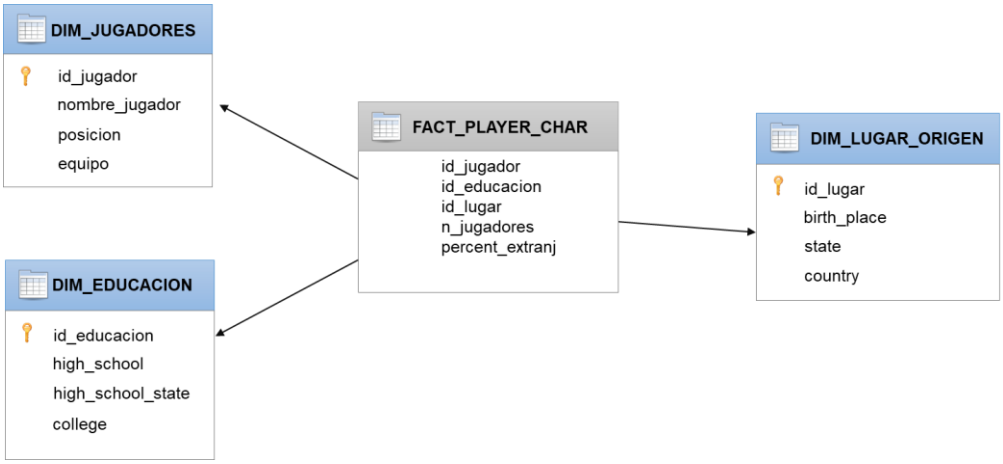
Nombre campo	Tipo	Tamaño
id_jugador	Número	4
nombre_jugador	Texto	255
posicion	Texto	100
equipo	Texto	255

Hecho:

FACT_PLAYER_CHAR

Nombre campo	Tipo	Tamaño
id_jugador	Número	4
id_educacion	Número	4
id_lugar	Número	4
n_jugadores	Número	4
percent_extranj	Número decimal	8

El esquema resultante es el siguiente:



TERCER CASO HECHO-DIMENSIONES

Dimensiones:

DIM_Game

Nombre campo	Tipo	Tamaño
id_game	Número	4
playoffs	Texto	20
period	Número	4
teams	Texto	20

DIM_Temporada

Nombre campo	Tipo	Tamaño
id_temporada	Número	4
desc_temporada	Texto	20

DIM_Jugadores

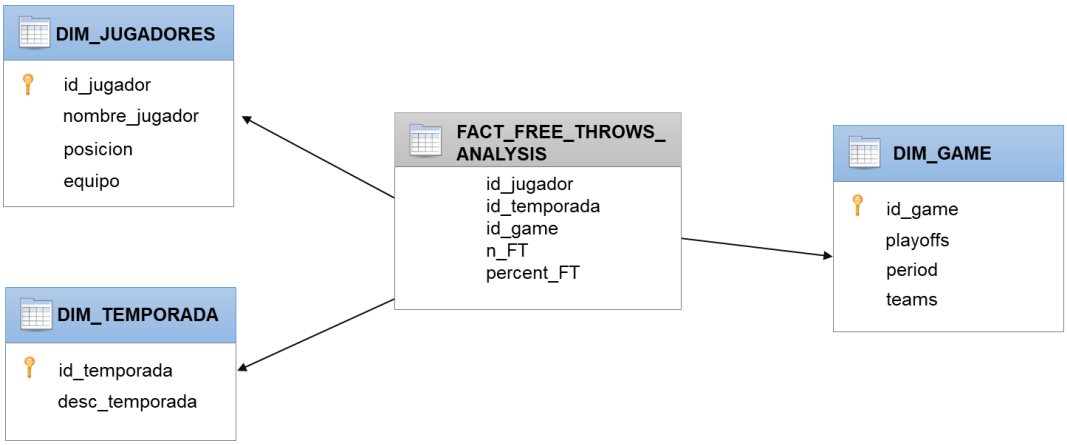
Nombre campo	Tipo	Tamaño
id_jugador	Número	4
nombre_jugador	Texto	255
posicion	Texto	100
equipo	Texto	255

Hecho:

FACT_FREE_THROWS_ANALYSIS

Nombre campo	Tipo	Tamaño
id_jugador	Número	4
id_temporada	Número	4
id_game	Número	4
n_FT	Número	4
percent_FT	Número decimal	8

El esquema resultante es el siguiente:


DIM_JUGADORES
id_jugador
nombre_jugador
posicion
equipo
FACT_FREE_THROWS_ANALYSIS
id_jugador
id_temporada
id_game
n_FT
percent_FT
DIM_GAME
id_game
playoffs
period
teams
DIM_TEMPORADA
id_temporada
desc_temporada

REFERENCIAS BIBLIOGRÁFICAS

Abelló Gamazo, A. , Curto Díaz, J. , Samos Jiménez, J. , Vidal Gil, J. y Díaz Arias, D. (2020). *La construcción de la factoria de información corporativa*. PID_00270641. FUOC

Abelló Gamazo, A. , Llorach Rius, C. , Ruiz Marqués, V. (2020). *Diseño multidimensional y explotación de datos*. PID_00270639. FUOC

J. Conesa, J. Curto. *El conocimiento imprescindible*. Disponible en: http://reader.digitalbooks.pro/content/preview/books/43005/book/OEBPS/chapter02.xhtml

Sevilla Marchena, N. (2019-2020). *Caso Práctico: Sistema Integrado de Egresados*. UOC

Díaz, D. (2020). *Caso práctico: Almacén de datos para el análisis de estadísticas deportivas de las ligas de baloncesto WNBA y NBA*. PID_00273959. UOC

Díaz, D. (2020). *Caso práctico: Almacén de datos para el análisis de estadísticas deportivas de las ligas de baloncesto WNBA y NBA*: *PRA1- Análisis y diseño del data warehouse*. PID_00277332. UOC

Llorach Rius, C. *Diseño de un almacén de datos para la gestión y reutilización de información meteorológica.* FUOC

Curto Díaz, J. *Solución: "El diseño de un almacén de datos para la gestión de la hospitalización de un hospital general básico".* PID_00211704. FUOC

Kimball, R. (2013). *The Data Warehouse Toolkit (3.ª ed.).* Nueva York: John Wiley & Sons Inc.

Inmon W. H. (1996). *Building the Data Warehouse (2.ª ed.)*. John Wiley & Sons Inc.

NBA Advanced Stats. Disponible en: https://www.nba.com/stats/

WNBA Advanced Stats. Disponible en: https://www.wnba.com/stats/

NBA Players stats since 1950. Disponible en: https://www.kaggle.com/drgilermo/nba-players-stats

NBA Free Throws. Disponible en: https://www.kaggle.com/sebastianmantey/nba-free-throws

NBA Players data. Disponible en: https://data.world/search?context=community&page=5&q=NBA&type=all