Case to be solved

Estudiante: Marcel Andrés Palma Céspedes

Curso: Asignatura 4

Profesora: Ing. Layla Cheli

Fecha: 24 de Julio de 2025

Enlace Kaggel: https://www.kaggle.com/datasets/ratin21/nba-player-stats-2024-25-per-game

Enlace Medium: https://medium.com/@mapace22/apache-spark-vs-fc3fca1a58a4

Archivos adjuntos: Power BI (.pbix) y tres datasets originales CSV para su revisión.

Caso práctico | Impacto y valor del Big Data

Parte 1

Resumen Ejecutivo: Desarrollo de un Dashboard Interactivo de Estadísticas NBA en Power BI

Introducción

El presente informe detalla el proceso integral de construcción de un modelo de datos y la creación de visualizaciones interactivas en Microsoft Power BI, utilizando tres conjuntos de datos relacionados con estadísticas de jugadores de la NBA para la temporada 2024-25. El objetivo principal fue transformar datos crudos en información valiosa y accionable, superando los desafíos inherentes al modelado de datos y al diseño de visualizaciones.

1. Proceso de Adquisición y Transformación de Datos (Power Query)

El proyecto se inició importando tres archivos CSV, cada uno conteniendo diferentes tipos de estadísticas de jugadores de la NBA:

1. **NBA Player Advanced Stats_2024-25.csv**: Estadísticas avanzadas (PER, TS%, USG%, WS, etc.).
2. **NBA Player Stats_2024-25_2.csv**: Estadísticas promedio por partido (PTS, REB, AST, etc.).
3. **NBA Player Stats_2024-25_Total.csv**: Estadísticas totales acumuladas por temporada.

La fase de transformación en Power Query Editor fue crucial para preparar los datos para el análisis:

- **Promoción de Encabezados:** Se identificó y promovió la primera fila de cada dataset como encabezados de columna para asegurar un esquema de datos legible.
- **Detección y Conversión de Tipos de Datos:** Se revisaron y ajustaron manualmente los tipos de datos para garantizar la precisión en los cálculos.

Por ejemplo, columnas numéricas como PTS, REB, AST, TS%, USG%, y WS se convirtieron a **números decimales (decimal number)**. Esto fue vital para permitir cálculos correctos como sumas y promedios.

- **Eliminación de Columnas Redundantes/No Necesarias:** Se identificaron y eliminaron columnas que no eran relevantes para el análisis final o que duplicaban información sin aportar valor adicional.
- **Limpieza General:** Se realizaron verificaciones para asegurar que no hubiera valores inconsistentes o errores evidentes en las columnas clave.

2. Modelado de Datos (Vista de Modelo)

El modelado de datos representó el desafío más significativo y la piedra angular para la interconectividad de la información: **Creación de Tabla de Dimensión Dim_Players:** Se identificó la columna Player (Nombre del Jugador) como una dimensión clave compartida entre los tres datasets. Para optimizar el modelo y evitar la duplicación, se creó una nueva tabla de dimensión (Dim_Players) extrayendo la columna Player de una de las tablas de hechos y eliminando duplicados para asegurar que cada jugador apareciera una única vez. Esta tabla actuaría como el "centro" de nuestro modelo en estrella.

- **Establecimiento de Relaciones Uno a Varios (1:*) y Bidireccionales:**
  - Se establecieron **tres relaciones activas, sólidas y bidireccionales ("Ambos")** desde la tabla Dim_Players (lado "uno") hacia cada una de las tres tablas de hechos (lado "varios").
  - **Reto Superado:** Inicialmente, Power BI tendía a crear relaciones indirectas inactivas o unidireccionales, lo que impedía el flujo de filtros deseado. Tuvimos que intervenir para **eliminar manualmente todas las relaciones iniciales** y **recrearlas una a una**, asegurándonos de que fueran:
    - **Activas (línea sólida, sin 'X').**
    - **Cardinalidad "Uno a varios (1:*)"** (Dim_Players 1 <--> * Tabla de Hechos).
    - **Dirección del filtro cruzado "Ambos"**, lo cual permite que los filtros aplicados en cualquier tabla de hechos se propaguen a la dimensión y luego a las otras tablas de hechos. Este punto fue crucial para la interactividad del dashboard.

Tabla No. 1 Modelado de Datos

Normalizando con la tabla auxiliar de Dim_player

1

3. Desarrollo de Visualizaciones Clave

Una vez que el modelo de datos estuvo robusto y las relaciones correctamente establecidas, se procedió a la creación de los siguientes gráficos, seleccionando los campos de las tablas optimizadas:

1. **Gráfico de Barras Agrupadas: "Top 10 Anotadores por Puntos (PTS)"**
    - **Eje Y:** Player (de Dim_Players).
    - **Eje X:** Suma de PTS (de NBA Player Stats_2024-25_2).
    - **Filtro:** Se aplicó un filtro Top N para mostrar solo los 10 jugadores con mayor suma de puntos, ordenados descendentemente.
    - **Utilidad:** Permite una rápida identificación de los jugadores con mayor impacto ofensivo en términos de anotación por partido.
2. **Gráfico de Dispersión: "Eficiencia de Tiro (TS%) vs. Uso de Balón (USG%) por Jugador"**
    - **Eje X:** Promedio de USG% (de NBA Player Advanced Stats_2024-25).
    - **Eje Y:** Promedio de TS% (de NBA Player Advanced Stats_2024-25).
    - **Valores (Puntos):** Player (de Dim_Players).
    - **Leyenda (Opcional):** Pos (Posición, de NBA Player Advanced Stats_2024-25) para identificar patrones por rol.
    - **Tamaño (Opcional):** Promedio de PTS (de NBA Player Stats_2024-25_2) para dimensionar los puntos por su capacidad anotadora.
    - **Reto Superado (Tooltips):** Se experimentó un desafío con los "tooltips" (información sobre herramientas) que inicialmente aparecían en blanco. Esto se resolvió asegurando que los campos relevantes (Player, TS%, USG%, PTS) fueran arrastrados explícitamente al área de "Información sobre herramientas" en el panel de visualizaciones y que sus agregaciones fueran ajustadas a "Promedio" o "Primer/Último" según el tipo de dato. Además, se descubrió y corrigió un error visual donde el color del texto y el fondo de los elementos del gráfico coincidían, haciendo que no fueran visibles.
    - **Utilidad:** Proporciona una visión gráfica de la eficiencia de los jugadores en relación con su volumen de balón, permitiendo identificar jugadores altamente eficientes con alto o bajo uso de balón.

3. **Gráfico de Columnas Agrupadas: "Total de Win Shares (WS) por Equipo"**
    - **Eje X:** Team (de NBA Player Stats_2024-25_2 o NBA Player Stats_2024-25_Total).
    - **Eje Y:** Suma de WS (de NBA Player Advanced Stats_2024-25).
    - **Orden:** Ordenado descendentemente por la suma de WS.
    - **Reto Superado:** Se corrigió la asignación inicial del campo Team, que por error se había indicado de la tabla de estadísticas avanzadas. Se confirmó que Team estaba disponible en las tablas NBA Player Stats_2024-25_2 y NBA Player Stats_2024-25_Total, y que el modelo manejaría la relación correctamente a través de Dim_Players.
    - **Utilidad:** Evalúa la contribución colectiva de los jugadores de un equipo a las victorias, proporcionando una métrica de rendimiento global del equipo.

4. Conclusiones

Este proyecto demostró la capacidad de Power BI para integrar, transformar y visualizar complejos conjuntos de datos de forma efectiva.

La creación de la tabla de dimensión Dim_Players y la configuración cuidadosa de las relaciones fueron fundamentales para la solidez y la interactividad del modelo, permitiendo el flujo de filtros en ambas direcciones y el cruce de datos entre diferentes tablas de hechos.

Los desafíos encontrados, particularmente en el modelado de relaciones y la configuración de tooltips, resaltan la importancia de comprender la arquitectura de datos subyacente y los detalles de configuración de Power BI. La superación de estos retos fue clave para garantizar la precisión y la legibilidad de las visualizaciones.

Con este modelo base y los gráficos esenciales construidos, se abre la puerta a análisis más profundos, como la incorporación de segmentadores interactivos por jugador, equipo o posición, y la creación de nuevas métricas calculadas (DAX) para obtener insights aún más específicos del rendimiento de los jugadores y equipos en la NBA.

Tabla No. 2 Siglas NBA

Data Type y significado de las siglas

Columna	Dataset(s)	Tipo de Dato Común	Significado en Español (Breve)
Rk	_2, _Total	Número entero	Rango del jugador en la lista.
Player	_Adv, _2, _Total	Texto	Nombre del jugador.
Age	_2, _Total	Número entero	Edad del jugador en la temporada.
Team	_2, _Total	Texto	Equipo al que pertenece el jugador.
Pos	_2, _Total	Texto	Posición principal del jugador (e.g., PG, SG, SF, PF, C).
G	_2, _Total	Número entero	Partidos jugados.
GS	_2, _Total	Número entero	Partidos iniciados como titular.
MP	_2, _Total	Número decimal	Minutos por partido (en _2) o Minutos totales jugados (en _Total).
FG	_2, _Total	Número decimal	Canastas de campo encestadas por partido o totales.
FGA	_2, _Total	Número decimal	Intentos de canasta de campo por partido o totales.
FG%	_2, _Total	Número decimal	Porcentaje de tiros de campo encestados.
3P	_2, _Total	Número decimal	Triples encestados por partido o totales.
3PA	_2, _Total	Número decimal	Intentos de triple por partido o totales.
3P%	_2, _Total	Número decimal	Porcentaje de triples encestados.
2P	_2, _Total	Número decimal	Canastas de 2 puntos encestadas por partido o totales.
2PA	_2, _Total	Número decimal	Intentos de canasta de 2 puntos por partido o totales.
2P%	_2, _Total	Número decimal	Porcentaje de tiros de 2 puntos encestados.
eFG%	_2, _Total	Número decimal	Porcentaje efectivo de tiros de campo (ajusta por valor del triple).
FT	_2, _Total	Número decimal	Tiros libres encestados por partido o totales.
FTA	_2, _Total	Número decimal	Intentos de tiro libre por partido o totales.
FT%	_2, _Total	Número decimal	Porcentaje de tiros libres encestados.
ORB	_2, _Total	Número decimal	Rebotes ofensivos por partido o totales.
DRB	_2, _Total	Número decimal	Rebotes defensivos por partido o totales.
TRB	_2, _Total	Número decimal	Rebotes totales por partido o totales.
AST	_2, _Total	Número decimal	Asistencias por partido o totales.
STL	_2, _Total	Número decimal	Robos por partido o totales.
BLK	_2, _Total	Número decimal	Tapones por partido o totales.
TOV	_2, _Total	Número decimal	Pérdidas de balón por partido o totales.
PF	_2, _Total	Número decimal	Faltas personales por partido o totales.
PTS	_2, _Total	Número decimal	Puntos por partido o totales.
PER	_Adv	Número decimal	Índice de Eficiencia del Jugador (Player Efficiency Rating).
TS%	_Adv	Número decimal	Porcentaje de Tiro Verdadero (True Shooting Percentage).
USG%	_Adv	Número decimal	Porcentaje de Uso (Usage Percentage).
OWS	_Adv	Número decimal	Win Shares Ofensivas (Offensive Win Shares).
DWS	_Adv	Número decimal	Win Shares Defensivas (Defensive Win Shares).
WS	_Adv	Número decimal	Win Shares (Total, OWS + DWS).
WS/48	_Adv	Número decimal	Win Shares por 48 minutos (ritmo de un partido completo).
OBPM	_Adv	Número decimal	Box Plus/Minus Ofensivo.
DBPM	_Adv	Número decimal	Box Plus/Minus Defensivo.
BPM	_Adv	Número decimal	Box Plus/Minus (Total, OBPM + DBPM).
VORP	_Adv	Número decimal	Valor sobre Jugador de Reemplazo (Value Over Replacement Player).

Leyenda de Datasets:

_Adv: NBA Player Advanced Stats_2024-25.csv

_2: NBA Player Stats_2024-25_2.csv

_Total: NBA Player Stats_2024-25_Total.csv

Artículo Medium | Apache Hadoop vs. Apache Spark

Parte 2

Se ha publicado un artículo en Medium titulado "Apache Spark vs. Apache Hadoop: Una Comparativa Esencial para la Elección de Herramientas Big Data".

El artículo ofrece un análisis comparativo fundamental entre estas dos herramientas clave del ecosistema Big Data, destacando sus arquitecturas, diferencias de rendimiento y casos de uso. Su objetivo es guiar en la toma de decisiones tecnológicas para el manejo eficiente de grandes volúmenes de datos.

Adjuntos y enlaces

Enlace Kaggel: https://www.kaggle.com/datasets/ratin21/nba-player-stats-2024-25-per-game

Enlace Medium: https://medium.com/@mapace22/apache-spark-vs-fc3fca1a58a4

Archivos adjuntos: Power BI (.pbix) y tres datasets originales CSV para su revisión.