# פרויקט בינה עסקית- הגשה 2

| | |
|---|---|
| עילי סולומון | 208934299 |
| גל נוברט | 208287953 |
| יובל אמסלם | 208998963 |

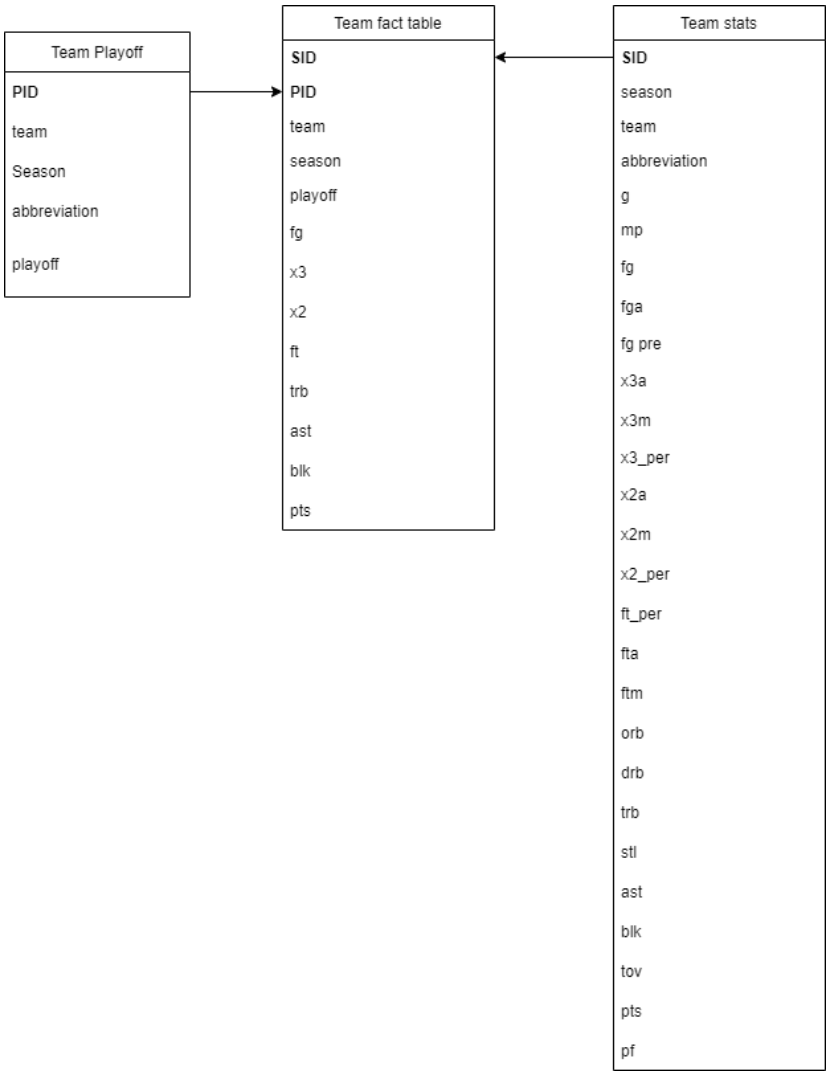
## מאגרי הנתונים

- קובץ 1 – 'teamstats' - מכיל נתונים על סטטיסטיקות מתקדמות עבור כל קבוצת כדורסל בליגת ה- NBA לכל עונה.
- קובץ 2 –'teamsplayoff' - מכיל נתונים על עלייה לשלב ה-Playoff עבור כל קבוצה לכל עונה.

## חלק 1- הגדרת Data Warehouse

1. נבחר להשתמש **בסכמת Star** מכיוון שיש לנו רק שיש לנו 2 טבלאות, אחת שמכילה נתונים רבים יותר (כ10 עמודות) ואחת שמכילה פחות נתונים.
בכל טבלה ניצור שדה חדש של מספר סידורי בעמודות חדשות- PID,SID שיהוו בתור ה-keys של אותה של כל טבלה בהתאמה.

2. **תיאור ה-Data Warehouse ע"י ERD :**



### 3. מימוש באמצעות SQL :

```sql
CREATE TABLE dsPlayoff (
 PID INT NOT NULL,
 season INT NOT NULL,
 lg VARCHAR (3),
 team VARCHAR (20),
 playoff BIT (20),
 abbreviation VARCHAR (3),
 PRIMARY KEY (PID)
);

```

```sql
CREATE TABLE dsStats (
 SID INT NOT NULL,
 season INT NOT NULL,
 team VARCHAR (20),
 abbreviation VARCHAR (3),
 g VARCHAR (20),
 mp VARCHAR (20),
 fg VARCHAR (20),
 fga VARCHAR (20),
 fg_per DECIMAL (1,4),
 x3a VARCHAR (20),
 x3m VARCHAR (20),
 x3_per DECIMAL (1,4),
 x2a VARCHAR (20),
 x2m VARCHAR (20),
 x2_per DECIMAL (1,4),
 fta VARCHAR (20),
 ftm VARCHAR (20),
 ft_per DECIMAL (1,4),
 orb VARCHAR (20),
 drb VARCHAR (20),
 trb VARCHAR (20),
 ast VARCHAR (20),
 stl VARCHAR (20),
 blk VARCHAR (20),
 tov VARCHAR (20),
 pf VARCHAR (20),
 pts VARCHAR (20),
 PRIMARY KEY (SID)
);
```

```sql
CREATE TABLE Fact (
 PID INT NOT NULL,
 SID INT NOT NULL,
 team VARCHAR (20),
 season INT NOT NULL,
 playoff BIT (5),
 fg VARCHAR (20),
 x3_per DECIMAL(1,4),
 x2_per DECIMAL (1,4),
 ft_per DECIMAL (1,4),
 trb VARCHAR (20),
 ast VARCHAR (20),
 blk VARCHAR (20),
 pts VARCHAR (20),
 FOREIGN KEY (PID) REFERENCES dsPlayoff(PID),
 FOREIGN KEY (SID) REFERENCES dsStats(SID)
);
```

### 4. תיאור Usecase

אתר הימורים בארץ רוצה להבין האם קיים קשר בין הנתונים הסטטיסטיים של הקבוצות לבין ההצלחה שלכם להגיע למשחקי הפלייאוף. אתר ההימורים מעוניין לבצע ניתוח של הנתונים על מנת להציע יחסי הימורים שתואמים את החיזוי. אתר ההימורים פנה אלינו על מנת לדעת מה הם סיכויי העפלה למשחקי הפלייאוף. בחרנו לבצע איחוד טבלאות לטבלת FACT אחת לטובת עבודה יעילה ומהירה יותר בשליפת הנתונים. בנוסף, הורדת הפרמטרים שנראים לנו פחות רלוונטיים מבחינה מקצועית חסכה עבורנו מקום אחסון ועומס לסט הנתונים, על כן השליפה מתבצעת בצורה מהירה יותר מאחר ויותר מהיר לעבד נתונים ל-2 טבלאות קטנות יותר מאשר על טבלה אחת גדולה.

## חלק 2- הגדרה ומימוש ETL

### 1. תהליך ה-ETL עבור אוסף הנתונים

**Extraction** – בשלב זה נחלץ את הנתונים והעמודות הרלוונטיות מתוך 2 הקבצים שיש לנו- teamplayoff, teamstats. הנתונים שנשלוף יהיו תואמים לסכמת כוכב שבחרנו על מנת לממש את ה- Data Warehouse .

**Transformation** – בשלב זה נסנן את העמודות והנתונים הלא רלוונטים לנו (עמודות חסרות או שאין צורך במידע על פרמטרים אלו מבחינה מקצועית) , נבצע אחידות בין העמודות שאנו מייבאים מתוך שני הקבצים וננקה את הטבלה כך שנוכל לעבוד עליה (תווים לא מזוהים).

**Loading** – בשלב זה נטען את הנתונים שלנו לאחר ניקוי וסינון של עמודות לא רלוונטיות אל ה Data warehouse שלנו.

## 2. תהליך ה-ETL Pipeline

**שלב 1 Data Reference –**בשלב זה נגדיר את סט הנתונים, כלומר ניצור סכמת כוכב אחת שתשלב 2 טבלאות: טבלת playoff וטבלת team stats אשר לכל אחד נגדיר מפתח ייחודי(PIS וSID).

**שלב 2 Reference Data from Extract –** בשלב זה נבצע חילוץ של הנתונים ע"י קובץ CSV.

**שלב 3 Validation Data –** בשלב זה נוודא כי הנתונים שלנו אמינים, כלומר אימות הנתונים של הקבוצות משתי הנקודות אכן תואמים למדדי ה KPI-שהגדרנו בתחילת הפרויקט.

**שלב 4 Data Transformation –** בשלב זה נבצע מספר פעולות: תחילה עלינו להגדיר חוקים עסקיים שיתאימו לשאלות המחקר שלנו,לבצע ניקוי יסודי של הנתונים וכן לוודא כי אינטגרציית המידע תקינה.

- בבדיקת סט הנתונים ראינו כי בשנת 2022 נתוני המשחקים לא תואמים את שאר השנים מכיוון שהעונה עדיין משוחקת ולכן עדיין אין לנו את עמודת המטרה והנתונים אינם של עונה שלמה כמו כל שאר העונות.
- כמו כן החלטנו למחוק את הרשומות של השנים 1947-1970 מכיוון שבשנים אלו איסוף נתוני הסטטיסטקה היה בחיתוליו ולכן ישנם הרבה ערכים ריקים בשנים אלו,לכן הוחלט למחוק בכדי שלא תהיה השפעה והטעיה של הנתונים.
- בחרנו להכניס את העמודות הרלוונטיות למדדי הKPI שלנו בכדי לעמוד חוקים עסקיים שקבענו.

**שלב 5 Stage –** בשלב זה ביצענו את העיבוד על הנתונים .

**שלב 6 Publish to Data warehouse –** בשלב זה העברנו את הדאטה למחסן הנתונים המיועד .כלומר, לטבלה המרכזית של הניקוד ובה יהיו המפתחות הרלוונטיים של השחקנים והמשחקים.

## 3. מימוש ה- באמצעות כלי Excel:

| Versi | Source Tal | Source Colun | Source Ty | Trans | Target Tab | Target Colur | Target Ty | Target Len | Default Val |
|---|---|---|---|---|---|---|---|---|---|
| 1 | playoff | PID | INT | | dsPlayoff | PID | INT | | NOT NULL |
| 1 | playoff | PID | INT | | Fact | PID | INT | | NOT NULL |
| 1 | playoff | season | INT | | dsPlayoff | season | INT | | NOT NULL |
| 1 | playoff | season | INT | | Fact | season | INT | | NOT NULL |
| 1 | playoff | lg | VARCHAR | | dsPlayoff | lg | VARCHAR | 3 | |
| 1 | playoff | lg | VARCHAR | | Fact | lg | VARCHAR | 3 | |
| 1 | playoff | team | VARCHAR | | dsPlayoff | team | VARCHAR | 20 | |
| 1 | playoff | team | VARCHAR | | Fact | team | VARCHAR | 20 | |
| 1 | playoff | playoff | VARCHAR | | dsPlayoff | playoff | VARCHAR | 20 | |
| 1 | playoff | playoff | VARCHAR | | Fact | playoff | VARCHAR | 20 | |
| 1 | playoff | abbreviation | VARCHAR | | dsPlayoff | abbreviation | VARCHAR | 3 | |
| 1 | playoff | abbreviation | VARCHAR | | Fact | abbreviation | VARCHAR | 3 | |
| 1 | Stats | SID | INT | | dsStats | SID | INT | | NOT NULL |
| 1 | Stats | SID | INT | | Fact | SID | INT | | NOT NULL |
| 1 | Stats | season | INT | | dsStats | season | INT | | NOT NULL |
| 1 | Stats | season | INT | | Fact | season | INT | | NOT NULL |
| 1 | Stats | team | VARCHAR | | dsStats | team | VARCHAR | 20 | |
| 1 | Stats | team | VARCHAR | | Fact | team | VARCHAR | 20 | |
| 1 | Stats | abbreviation | VARCHAR | | dsStats | abbreviation | VARCHAR | 3 | |
| 1 | Stats | abbreviation | VARCHAR | | Fact | abbreviation | VARCHAR | 3 | |
| 1 | Stats | g | VARCHAR | | dsStats | g | VARCHAR | 20 | |
| 1 | Stats | g | VARCHAR | | Fact | g | VARCHAR | 20 | |
| 1 | Stats | mp | VARCHAR | | dsStats | mp | VARCHAR | 20 | |
| 1 | Stats | mp | VARCHAR | | Fact | mp | VARCHAR | 20 | |
| 1 | Stats | fg | VARCHAR | | dsStats | fg | VARCHAR | 20 | |
| 1 | Stats | fg | VARCHAR | | Fact | fg | VARCHAR | 20 | |
| 1 | Stats | fga | VARCHAR | | dsStats | fga | VARCHAR | 20 | |
| 1 | Stats | fga | VARCHAR | | Fact | fga | VARCHAR | 20 | |
| 1 | Stats | fg_per | DECIMAL | | dsStats | fg_per | DECIMAL | (1,4) | |
| 1 | Stats | fg_per | DECIMAL | | Fact | fg_per | DECIMAL | (1,4) | |
| 1 | Stats | x3a | VARCHAR | | dsStats | x3a | VARCHAR | 20 | |
| 1 | Stats | x3a | VARCHAR | | Fact | x3a | VARCHAR | 20 | |
| 1 | Stats | x3m | VARCHAR | | dsStats | x3m | VARCHAR | 20 | |
| 1 | Stats | x3m | VARCHAR | | Fact | x3m | VARCHAR | 20 | |
| 1 | Stats | x3_per | DECIMAL | | dsStats | x3_per | DECIMAL | (1,4) | |
| 1 | Stats | x3_per | DECIMAL | | Fact | x3_per | DECIMAL | (1,4) | |
| 1 | Stats | x2a | VARCHAR | | dsStats | x2a | VARCHAR | 20 | |
| 1 | Stats | x2a | VARCHAR | | Fact | x2a | VARCHAR | 20 | |
| 1 | Stats | x2m | VARCHAR | | dsStats | x2m | VARCHAR | 20 | |
| 1 | Stats | x2m | VARCHAR | | Fact | x2m | VARCHAR | 20 | |
| 1 | Stats | x2_per | DECIMAL | | dsStats | x2_per | DECIMAL | (1,4) | |
| 1 | Stats | x2_per | DECIMAL | | Fact | x2_per | DECIMAL | (1,4) | |
| 1 | Stats | fta | VARCHAR | | dsStats | fta | VARCHAR | 20 | |
| 1 | Stats | fta | VARCHAR | | Fact | fta | VARCHAR | 20 | |
| 1 | Stats | ftm | VARCHAR | | dsStats | ftm | VARCHAR | 20 | |
| 1 | Stats | ftm | VARCHAR | | Fact | ftm | VARCHAR | 20 | |
| 1 | Stats | ft_per | DECIMAL | | dsStats | ft_per | DECIMAL | (1,4) | |
| 1 | Stats | ft_per | DECIMAL | | Fact | ft_per | DECIMAL | (1,4) | |
| 1 | Stats | orb | VARCHAR | | dsStats | orb | VARCHAR | 20 | |
| 1 | Stats | orb | VARCHAR | | Fact | orb | VARCHAR | 20 | |
| 1 | Stats | drb | VARCHAR | | dsStats | drb | VARCHAR | 20 | |
| 1 | Stats | drb | VARCHAR | | Fact | drb | VARCHAR | 20 | |
| 1 | Stats | trb | VARCHAR | | dsStats | trb | VARCHAR | 20 | |
| 1 | Stats | trb | VARCHAR | | Fact | trb | VARCHAR | 20 | |
| 1 | Stats | ast | VARCHAR | | dsStats | ast | VARCHAR | 20 | |
| 1 | Stats | ast | VARCHAR | | Fact | ast | VARCHAR | 20 | |
| 1 | Stats | stl | VARCHAR | | dsStats | stl | VARCHAR | 20 | |
| 1 | Stats | stl | VARCHAR | | Fact | stl | VARCHAR | 20 | |
| 1 | Stats | blk | VARCHAR | | dsStats | blk | VARCHAR | 20 | |
| 1 | Stats | blk | VARCHAR | | Fact | blk | VARCHAR | 20 | |
| 1 | Stats | tov | VARCHAR | | dsStats | tov | VARCHAR | 20 | |
| 1 | Stats | tov | VARCHAR | | Fact | tov | VARCHAR | 20 | |
| 1 | Stats | pf | VARCHAR | | dsStats | pf | VARCHAR | 20 | |
| 1 | Stats | pf | VARCHAR | | Fact | pf | VARCHAR | 20 | |
| 1 | Stats | pts | VARCHAR | | dsStats | pts | VARCHAR | 20 | |
| 1 | Stats | pts | VARCHAR | | Fact | pts | VARCHAR | 20 | |