**EXTRACT/LOAD**

**Introduction**:

→ We looked at NBA data, specifically focusing on players who got player of the week from 1985-86 to 2017-18 season and seeing their stats. We got players stats & salaries info from data.world and player of the week info from Kaggle. Initially, we wanted to look at all of the stats for players who got player of the week in one comprehensive table but ran into issues regarding duplicates so we decided to focus on two fields (salaries & draft year) for simplicity’s sake for this project.

**Set-up:** Reasoning for using GCP and how we set up our Database

Storage, Folder, Load into the database by using google cloud storage, “Create table” function

This process was done for multiple different data files.

**TRANSFORM**

We needed to consolidate several files to transform our dataset. We chose to create views for the first two joins (Player\_join & POTW) because it hides the complexity of the code from the end-user (simulates real-life scenarios for requests from other departments at work) and is suitable for layering multiple queries while saving space.

**Player\_join (View)**

SELECT p.name, p.career\_FG\_, p.birthPlace, p.position, p.college, p.career\_FG3\_, p.career\_PTS, p.career\_WS, p.career\_PER, p.career\_FT\_, p.career\_AST, p.career\_TRB, p.highSchool, s.team, s.season, s.salary,

FROM `all-about-that-database.nba.sal\_players` as p

JOIN `all-about-that-database.nba.salaries` as s

ON p.\_id = s.player\_id

→ We joined the player’s stats dataset (sal\_players) and salaries dataset together by their common ids, mainly to put all the players’ information in one table which we used for the final table to look at players of the week, their respective draft year and salaries.

**POTW\_VIEW**

SELECT Player, Date, Age, Weight, Height, Pre\_draft\_Team, Draft\_Year, Conference,

FROM `all-about-that-database.nba.player\_otw`

→ We wanted to isolate the fields we needed from NBA player of the week to prepare for joining

**NOTE: For the join - we needed to redo it because it was multiplying the data values and creating extra rows**

We were trying to join player stats with the salaries info table and player of the week table together to see all the stats for player of the week in one table. But, we realized it led to duplicate rows and we needed to join on a unique key. So, we found the average salary from the player stats table and draft year from the player of the week table and joined them (which was possible because both tables are unique) to look at players (the players that got player of the week) draft year and salary.

**FINAL Table**

SELECT v.\*, FORMAT("%.f", salary)

FROM (SELECT name, avg(salary) salary FROM `all-about-that-database.nba.player\_join` group by name) as j

JOIN (SELECT Player, min(Draft\_Year) draft\_year FROM `all-about-that-database.nba.potw\_view` group by Player) as v

ON j.name = v.Player

→ We changed the format of salary, and joined the average salary from the player table to the player of the week view, we used the name of the player to join both tables

**What we would’ve done differently:**

- Used another database tool to learn from the manual process

- Presented in a Jupyter notebook OR used Google Colaborators since it works seamlessly with GCP

- Spent more time inspecting the columns with quick database and assigning primary keys to avoid dropping relevant data.

- Distribute workload better so we could strengthen the project (extract data from an API, create a flask page, etc.)

Data Sources:

Kaggle:

* <https://www.kaggle.com/jacobbaruch/nba-player-of-the-week>

Dataworld:

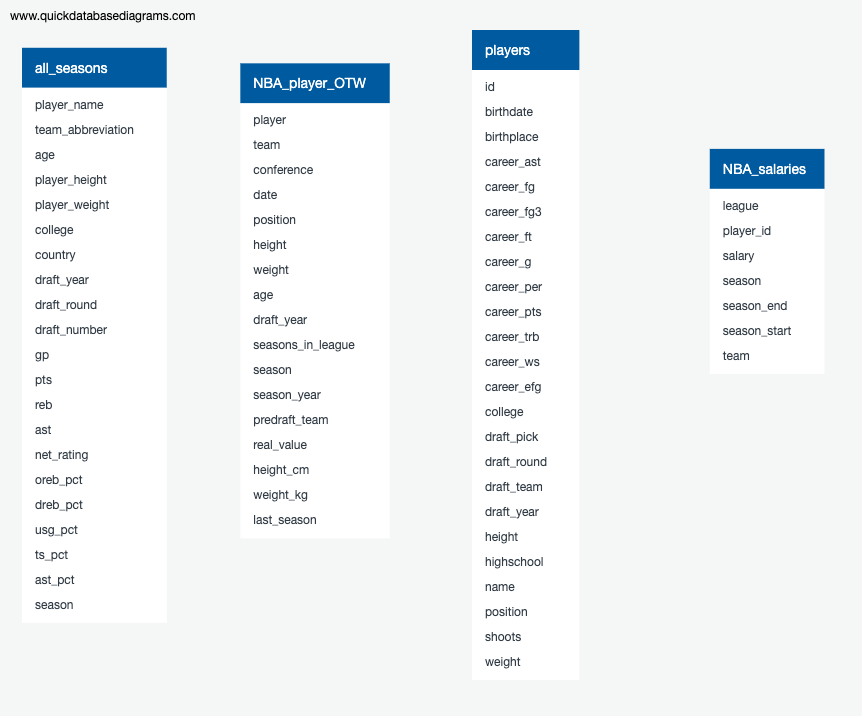
* <https://data.world/datadavis/nba-salaries>

Google Cloud Platform:

* <https://console.cloud.google.com/bigquery?project=all-about-that-database&p=all-about-that-database&d=nba&t=player_otw&page=table>

Project\_GitHub:

* <https://github.com/Rotas20/Data-Bass>

****

**Player\_join**

SELECT p.name, p.career\_FG\_, p.birthPlace, p.position, p.college, p.career\_FG3\_, p.career\_PTS, p.career\_WS, p.career\_PER, p.career\_FT\_, p.career\_AST, p.career\_TRB, p.highSchool, s.team, s.season, s.salary,

FROM `all-about-that-database.nba.sal\_players` as p

JOIN `all-about-that-database.nba.salaries` as s

ON p.\_id = s.player\_id

**POTW\_VIEW**

SELECT Player, Date, Age, Weight, Height, Pre\_draft\_Team, Draft\_Year, Conference,

FROM `all-about-that-database.nba.player\_otw`

**FINAL**

SELECT v.\*, FORMAT("%.f", salary)

FROM (SELECT name, avg(salary) salary FROM `all-about-that-database.nba.player\_join` group by name) as j

JOIN (SELECT Player, min(Draft\_Year) draft\_year FROM `all-about-that-database.nba.potw\_view` group by Player) as v

ON j.name = v.Player