Does Attendance Matter?

Gathering Insights on the NBA

By:

Iris Cheng, Sebastian Clavijo, Emily MacQuarrie, Vidhath Raghavan

Executive Summary

Sports is a profit-driven industry that facilitates competition, entertainment, and community. Data analytics can be leveraged in this industry to maximize these benefits and drive value for consumers. In particular, our projects would analyze the trends and factors that influence a fan's willingness and ability to attend an NBA game event from game-level, season-level and team-level perspectives. Based on these insights, we would make recommendations to increase the number of attendees and better inform the business decisions of different stakeholders across the organization.

Business Case

The NBA has massive amounts of data that can be leveraged to better meet the demands of team owners and consumers. From a team owner perspective, increasing attendance is a means of maximizing revenue and improving the team's reputation. From a consumer perspective, high attendance may be considered a signal for consumer satisfaction and team pride. By finding ways to increase attendance, we can maximize the utility of both team owners and consumers.

To this end, we are investigating the following research question: *Which aspects of the NBA (teams, players, player behavior, seasons etc.) attract the highest attendance and why?*

Data Sources

The following publicly available data will be used:

- *http://www.espn.com/nba/attendance*
- *https://www.bigdataball.com/*

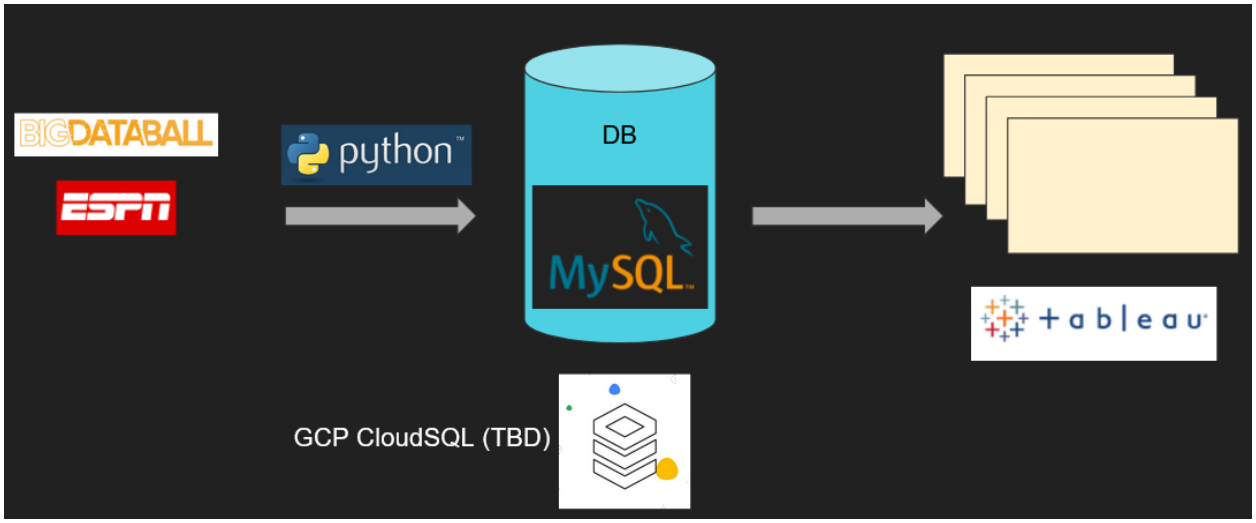
These datasets contain information on game stats per season and team, player stats per game by team and season, and attendance rates.

Data Preparation

To begin, the 2021-2022 NBA season consists of 30 NBA teams playing in 82 regular season games. Those games will be played in 29 arenas across a roughly 5 month span. As our data extends back in time, we cannot assume that all teams, locations and arenas are the same size. For instance, franchises can change cities and areas, therefore we will also need to track those as well. There are some considerations on data availability as we go back in time. That being said, most of your data sources are fairly surface level and should be easily normalized in our database platforms.

With this in mind, we will need to take two major steps to clean and process the data before loading it into the database. The first transformation is creating a consistent naming convention for teams across data sources. The second transformation requires us to match home and away teams based on a unique game identifier to collect game-level statistics.

Data Pipeline Model:



Database Platform Considerations

For the purposes of this project, we have chosen to use a **relational database** to support transactions and future-proof our database.

Our data is well normalized to support unique franchises over time over different cities and can track player information over a set period of time. That allows us to support various sets of transactions be it specific conditional queries or larger data summaries of our table.

As mentioned earlier, NBA franchises can change and grow. Therefore, we wanted to have our database support updates over time. This includes ‘momentous’ updates such as new franchises in new cities to updating player and team level statistics in future seasons. With that in mind, we believe that our relational database addresses our concerns better than a dimensional database.

Current Build:

6 Tables:

- *address*
- *nba_games*
- *nba_season*
- *nba_season_attendance*
- *nba_team*
- *nba_team_stats*

Next Steps:

- Franchise Information
- Player Information
- Update *address* to include arena attributes

EER Document: