# **Data Warehouse Project**

# **Project Title: NBA Player Performance Analysis**

# **Luyao Chu**

# 

# 

# 

# 

## 

## 

## 

## 

## 

## **The Business/Organization & Opportunity**

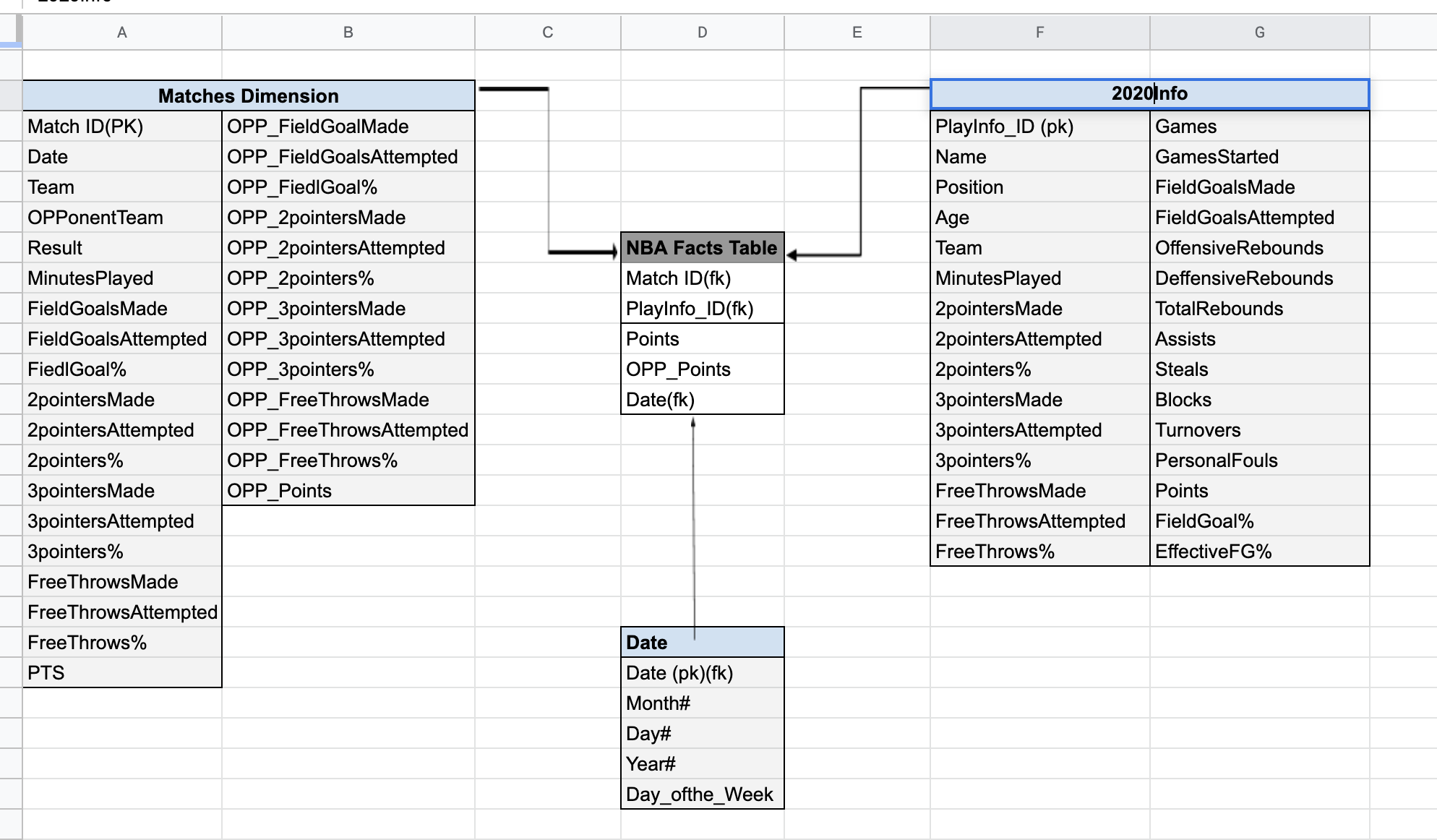
NBA is always an exciting world-class event. People from different countries and cultures will gather together to celebrate the joy of the victory from their favorite or just to join in the fun during every season. In order to help people who are passionate about the NBA to have more knowledge about their favorite team and players, we create this data warehouse to analyze each player’s performance based on the latest season. In this data warehouse, we use basketball references as the source of our dataset. We collect each athlete’s name, age, position, total points, total assists, total rebounds, total steals, and total blocks. Then we calculate the average of how they perform in each game. We will be using this information to provide deeper insights for each player. The teams also can use the metrics from this dataset to have a better understanding of each player. By knowing their strength and weakness, arrange them to a suitable position to play the game at their best.

## **Proposal**

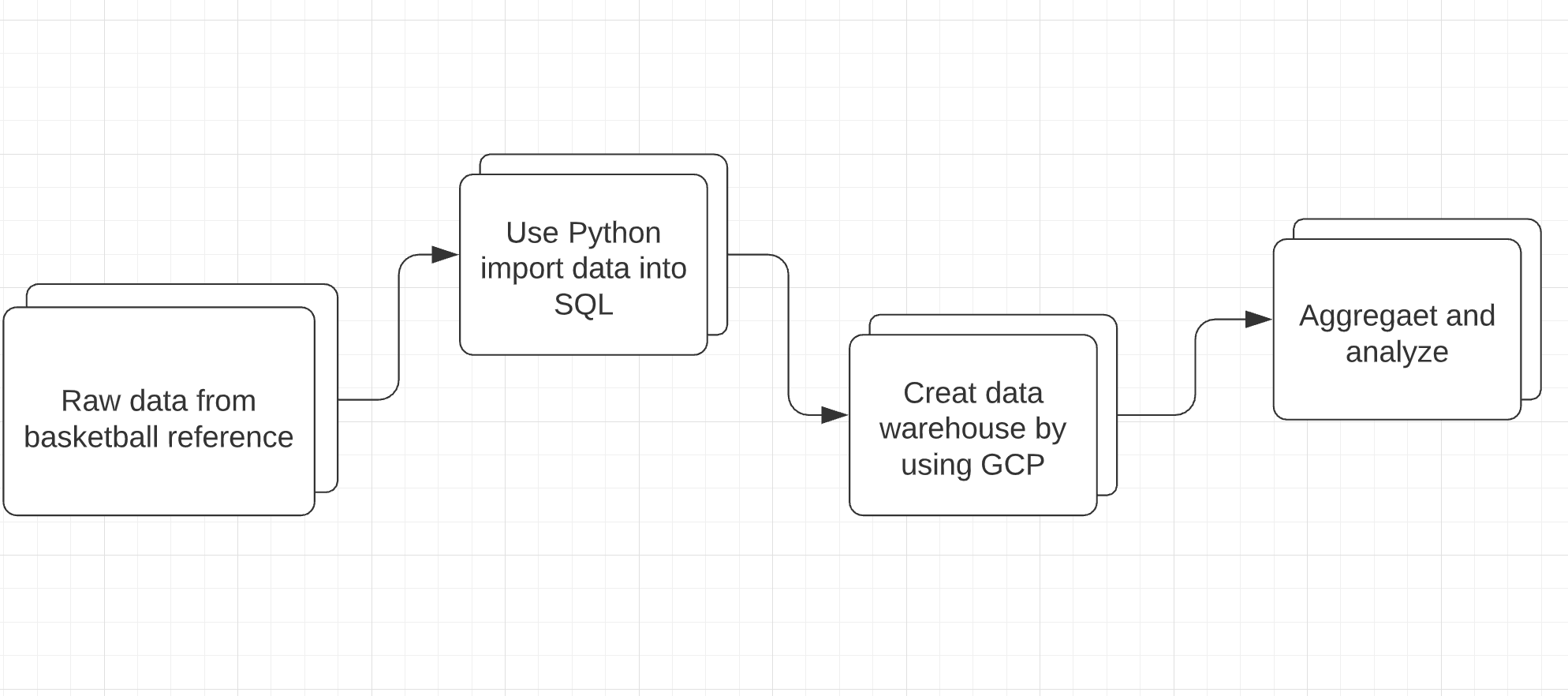
### **Problem**

As we address in the previous paragraph, the basketball reference webpage has generated each player’s and team’s performance into the computation of a systematic metric, specifically assists, rebounds, blocks, steals, 3-points, and 2-points. These are basic basketball metrics to measure a player's efficiency in a basketball game. By using these data, we are going to have deeper insights into each player in the current season and even forecast their position and performance in the upcoming season.

**Dimensional Model**

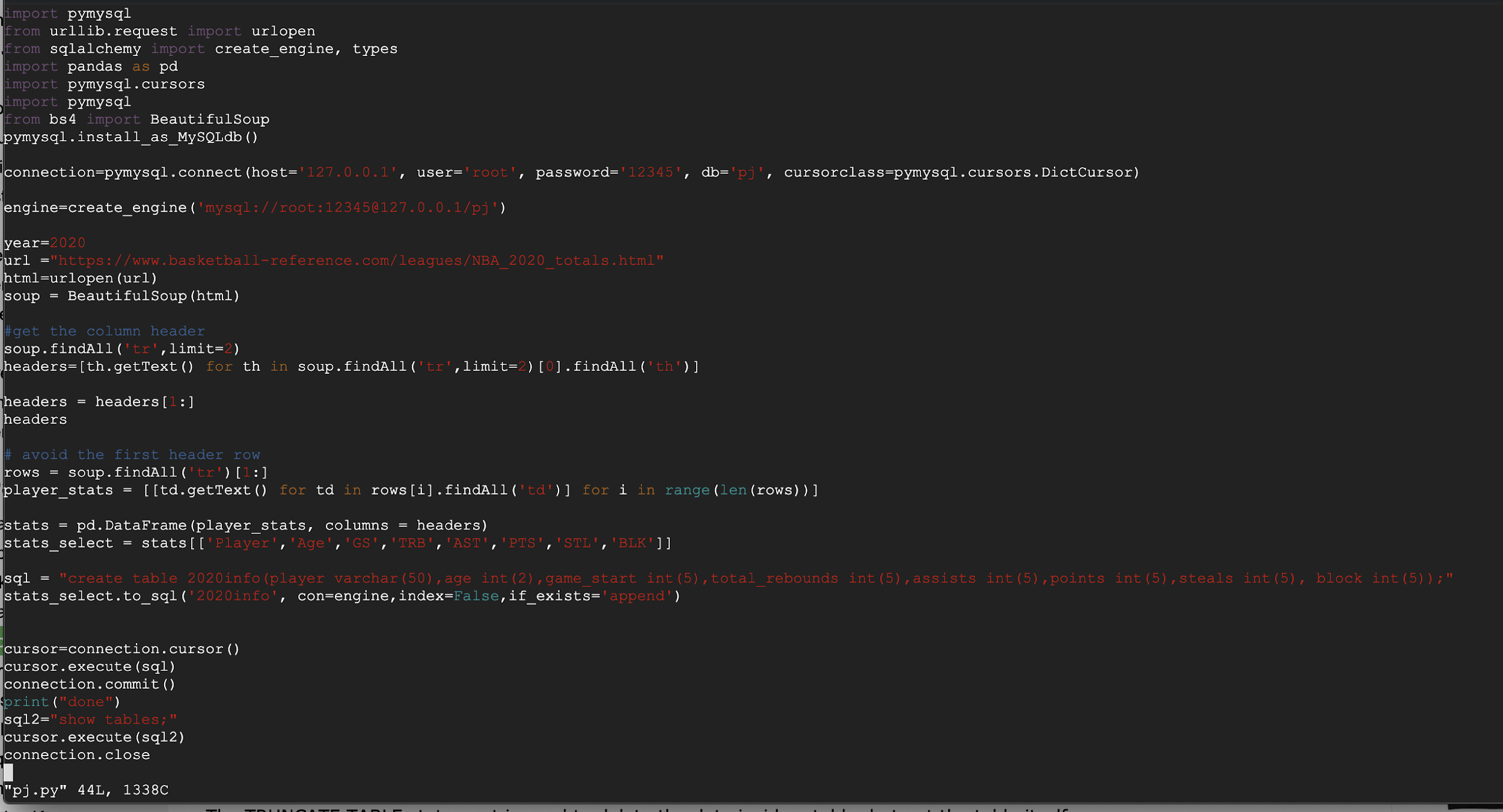


### **ETL processing**

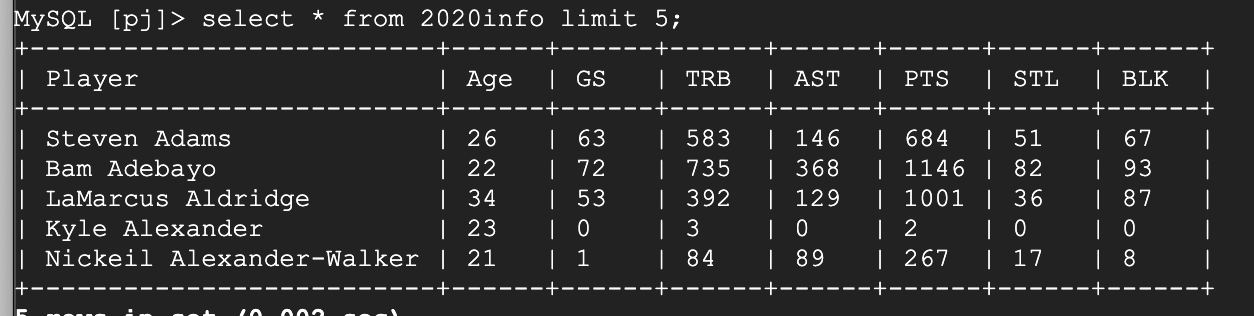


1. The data source is basketball-reference.com. At first, we only need to collect the raw data from the latest season.
2. Then we will use python to scrape and process the data.
3. Import data into Google Cloud Platform Mysql.
4. In the end, we will aggregate all the information we have into a more precise visualization format by using Tableau.

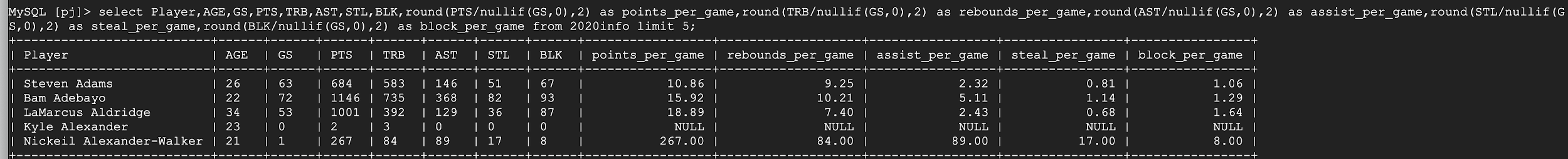
### **Detailed Design**

In this project, the method that will be used for the ETL process are Python, Google Cloud Platform, and Tableau. First, I will use Python to Scrap the player statistics from basketball references. The following screenshot is how I use beautiful soup to scrap and trim the player’s scores from the 2020 season game into MySQL. Then I extract the headers of the table and store them in a list to prepare for further use. After that, I extract the leftover data and combine them into a two-dimensional format. At the end of the scraping, I use pandas to create DataFrame to aggregate everything I have. After uploading them into MySQL by using sqlalchemy, I only select Player, Age, GS, TRB, AST, PTS, STL, and BLK, which are basic and important factors to consider a player’s strength, to create tables for my further analysis.

This is what the table looks like in MySQL database.



Then I calculate the average of points, assists, rebounds, steals, and blocks each player can make per game. After that, we will deeper insights into each player’s strengths. The audience will have more understanding of each player. On the other hand, this data warehouse can also be a great tool for those professional commentators or even NBA teams to have better metrics to make more advanced plans and advice for the upcoming season.



**Ethics**

This data warehouse is gathering information from a public reference website. There’s no personal information is collected in this data warehouse. Therefore, this data warehouse is completely ethical. However, the privacy statement from the basketball-reference website declares that the visitor’s home servers’ information might be aggregated for internal use or provided to a third party. Also, if a person signs up for the website, their personal information, such as their name, email, address, and even credit/debit card information will be collected for internal use. As they state on the website, the stats of the user’s information for analysis of the user’s behavior. However, nobody is going to know what exactly their information is going to be used for. Their information might be investigated or sold to a scam company. That’s why people always receive scam calls ten times a day. However, without collecting data, the company wouldn’t provide the accurate information the audience wants.