Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

NBA

# NBA 16-17 Regular Season Analysis

# Enterprise Architectures for Big Data

Oksana Hrytsiv
Laman Mammadova
Fabian Petschke
Adrian Villegas

# Project Framework

1. Introduction
2. Data Preparation
3. Use Cases
    a. Scenario 1 (Google BigQuery + Tableau)
    b. Scenario 2 (Apache Hive + Power BI)
    c. Scenario 3 (PySpark + Zeppelin)
4. Data Visualization
5. Technologies Evaluation

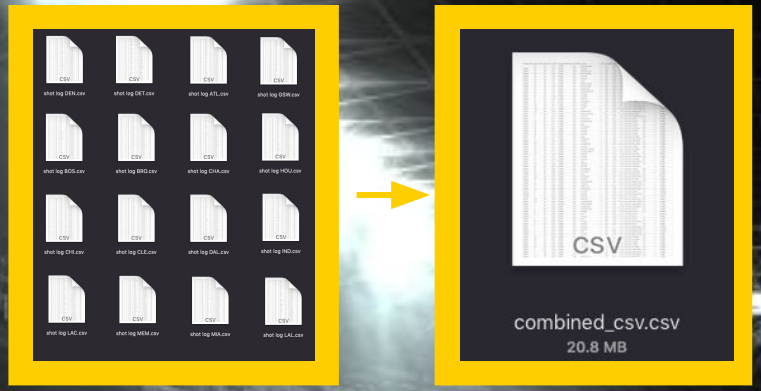
# 1. Introduction

Help basketball analysts to evaluate the performance of players and each team

- Game Schedule 16-17-Reg
- Player Regular 16-17 Stats
- Shot log data for each team

Apply big data tools to process and visualize data

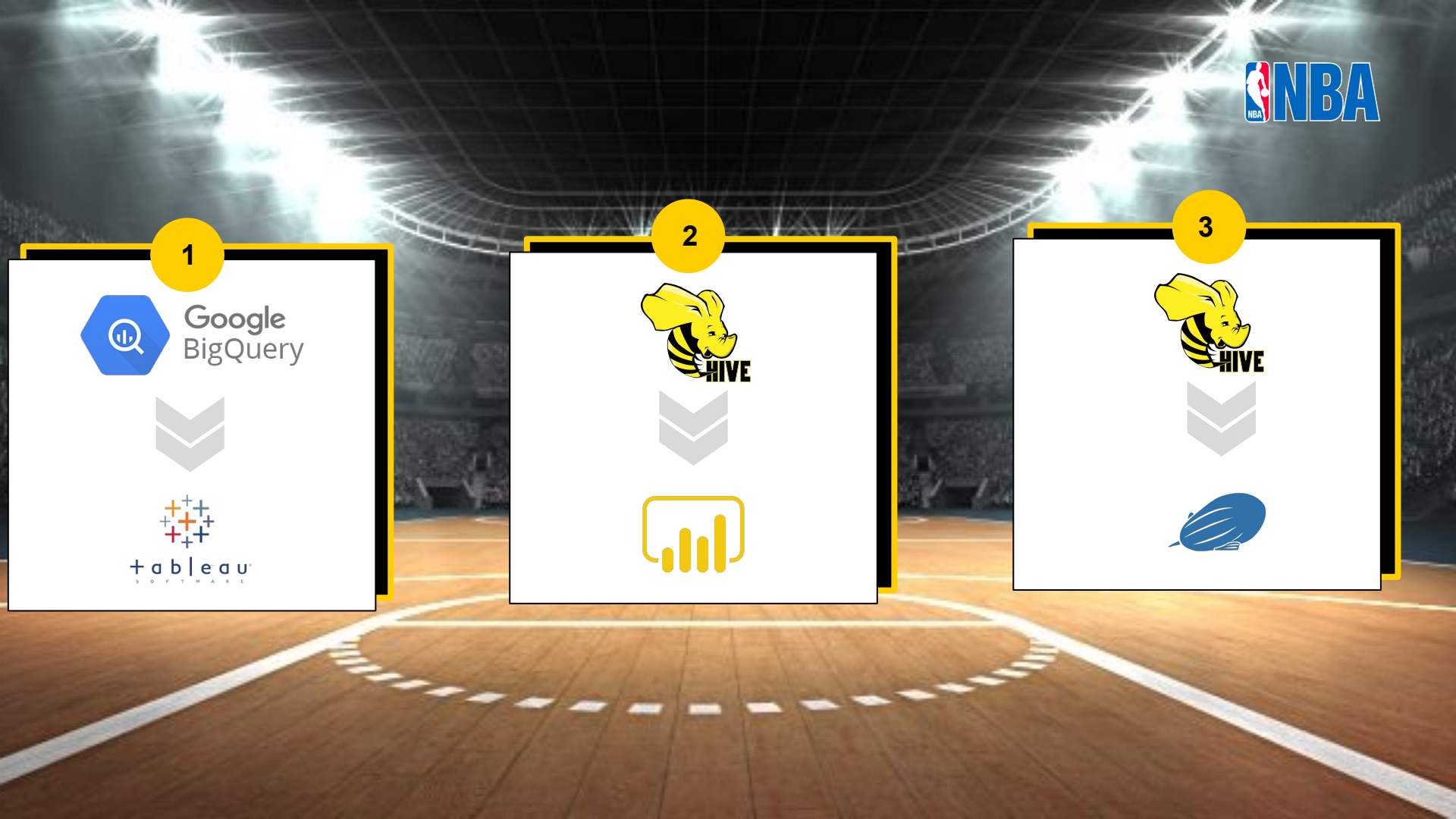

NBA
2. Data Preparation



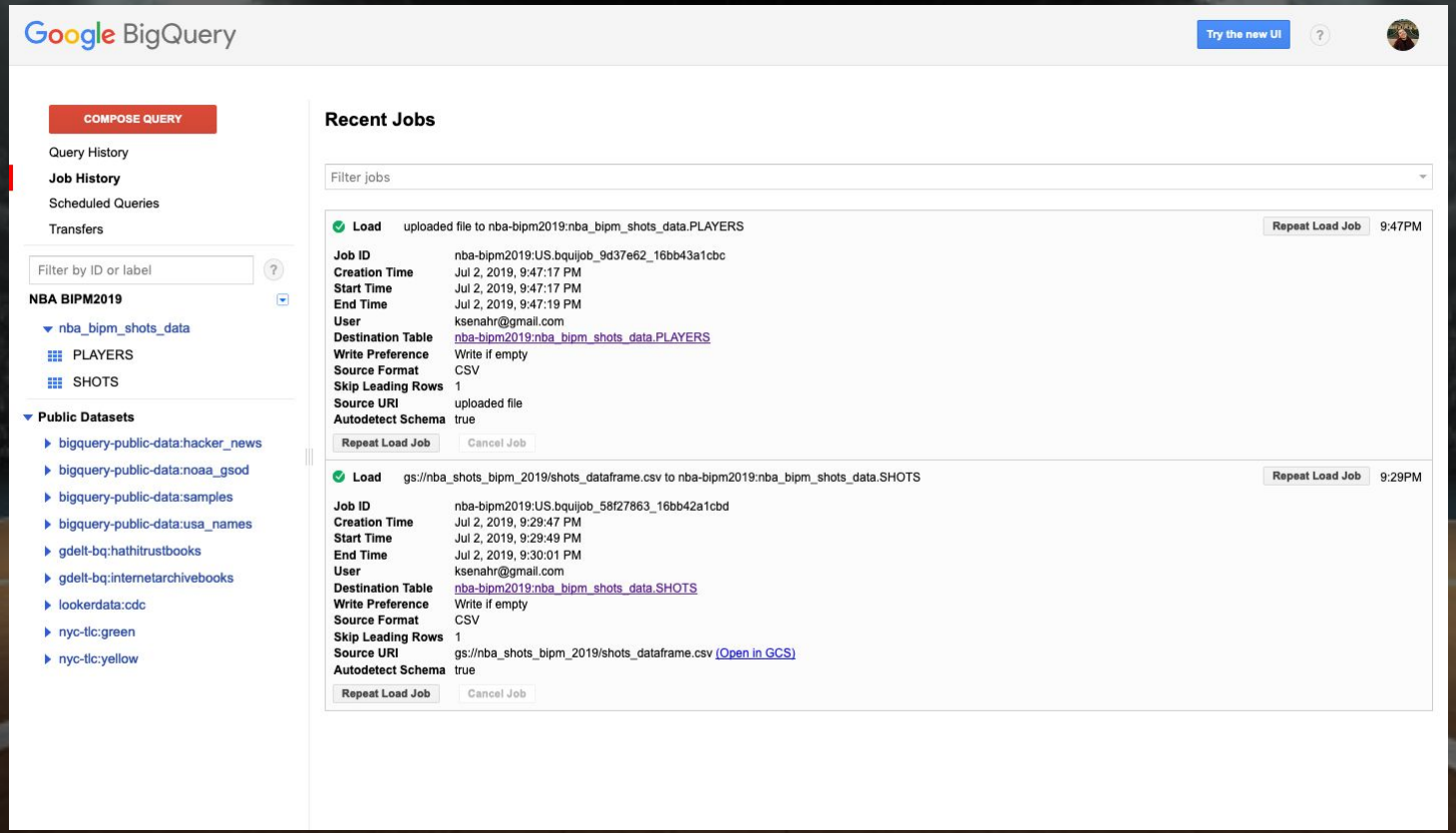

Rows: 209176
Columns: 13

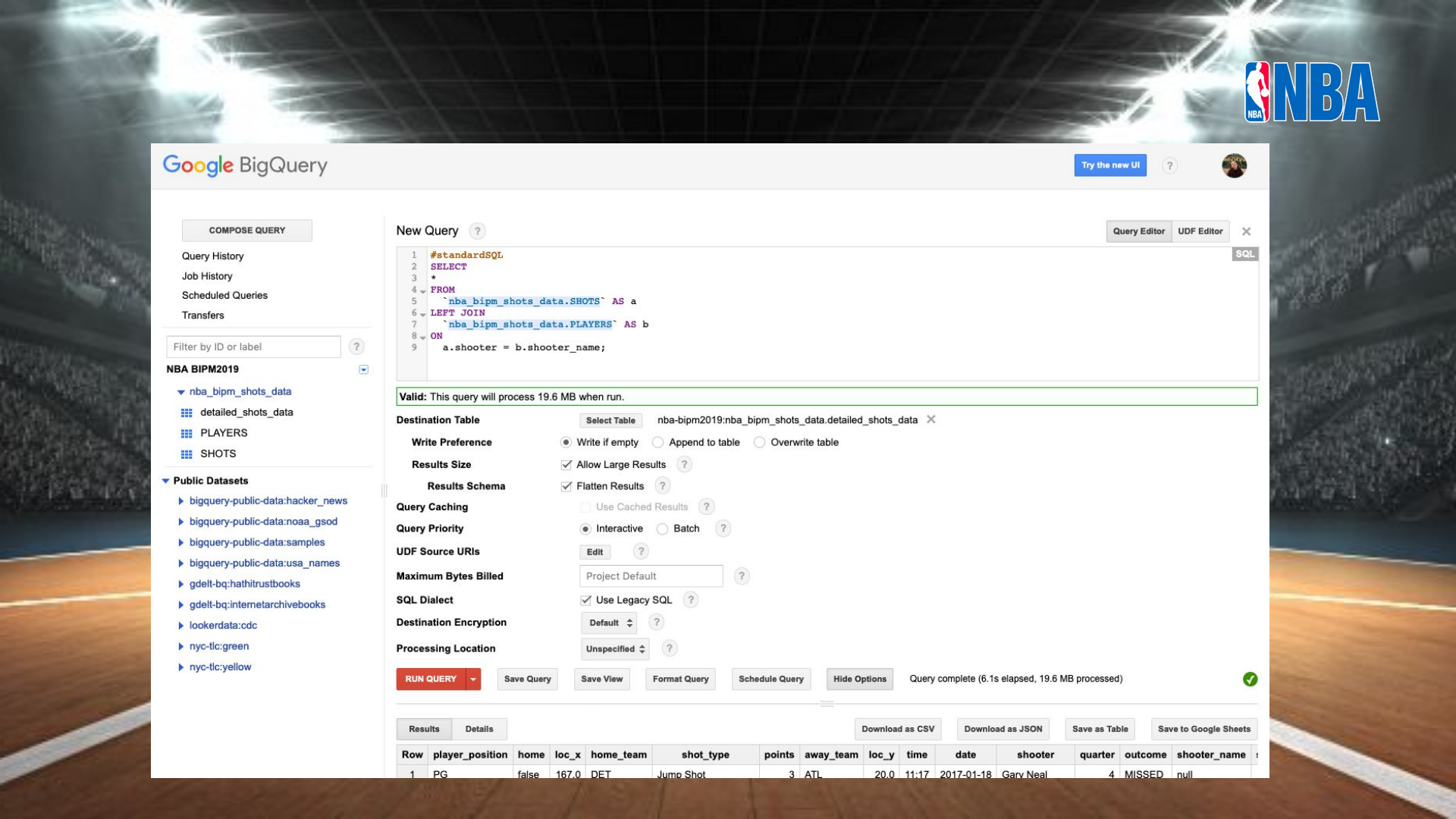
| | player_position | home | loc_x | home_team | shot_type | points | away_team | loc_y | time | date | shooter | quarter | outcome |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | PF | Yes | 107.0 | PHX | Floating Jump Shot | 2 | SAC | 252.0 | 0:51 | 2016-10-26 | Jared Dudley | 1 | SCORED |
| 1 | SG | Yes | 254.0 | PHX | Jump Shot | 3 | SAC | 56.0 | 1:14 | 2016-10-26 | Devin Booker | 1 | MISSED |
| 2 | C | Yes | 52.0 | PHX | Cutting Dunk Shot | 2 | SAC | 250.0 | 1:44 | 2016-10-26 | Tyson Chandler | 1 | SCORED |
| 3 | PG | Yes | 241.0 | PHX | Pullup Jump Shot | 2 | SAC | 359.0 | 2:16 | 2016-10-26 | Eric Bledsoe | 1 | MISSED |
| 4 | SG | Yes | 225.0 | PHX | Jump Shot | 3 | SAC | 447.0 | 2:40 | 2016-10-26 | Devin Booker | 1 | MISSED |


# 3. Use Cases

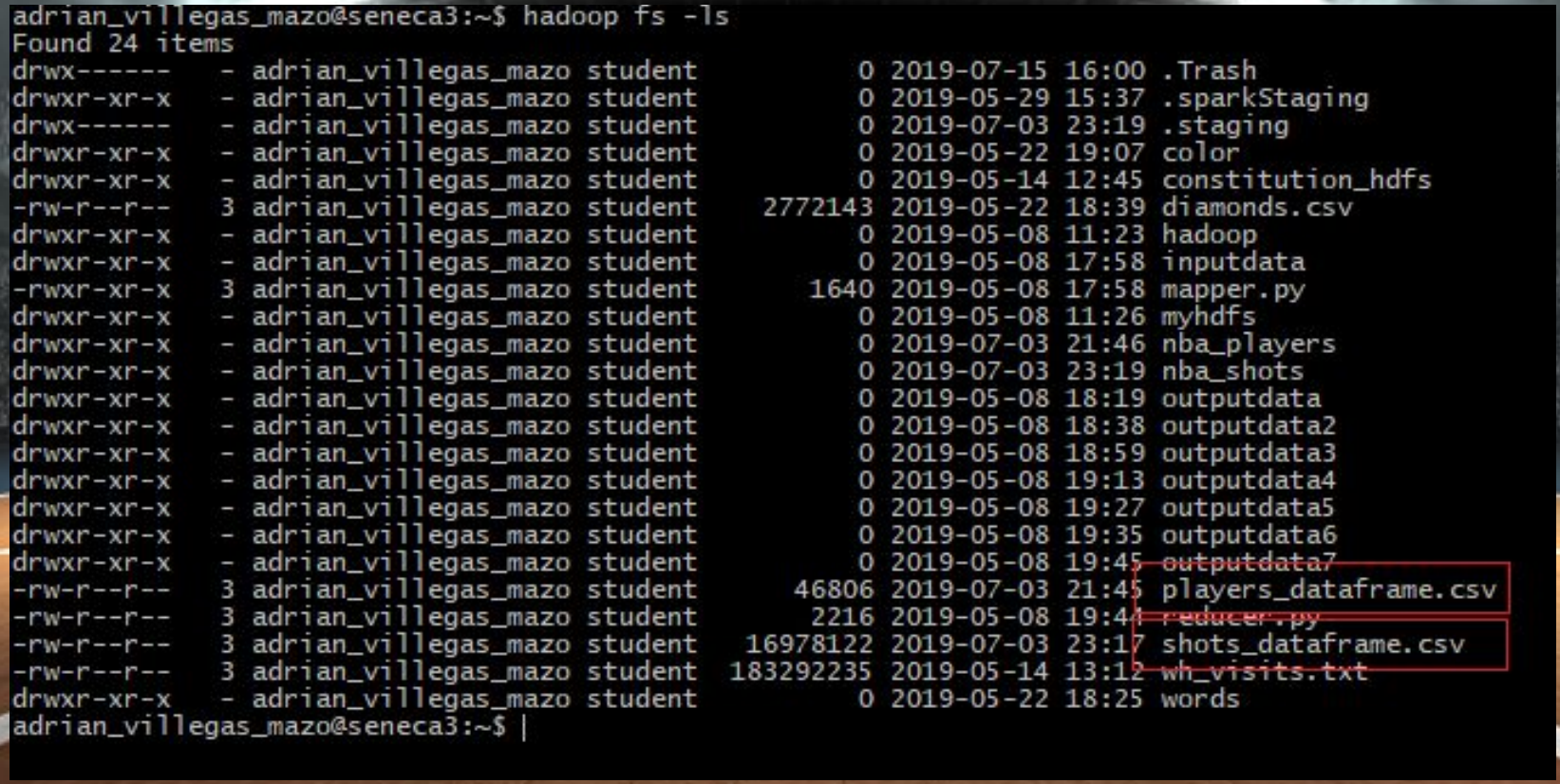

NBA
1
Google BigQuery
tableau
SOFTWARE
2
HIVE
3
HIVE

# Scenario 1 (Google BigQuery + Tableau)



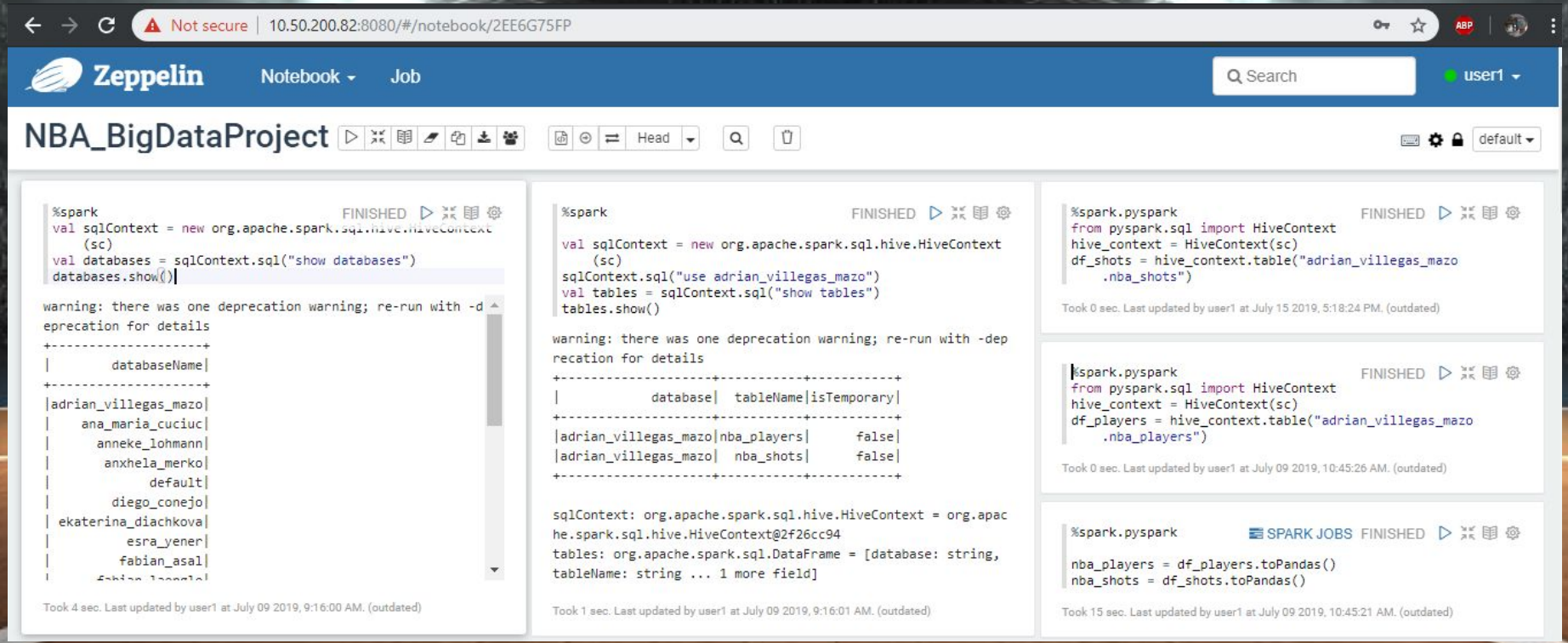

Google BigQuery
Try the new UI
COMPOSE QUERY
Query History
Job History
Scheduled Queries
Transfers
Filter by ID or label
NBA BIPM2019
nba_bipm_shots_data
detailed_shots_data
PLAYERS
SHOTS
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
nyc-tlc:green
nyc-tlc:yellow
New Query
Query Editor
UDF Editor
SQL
#standardSQL
SELECT
*
FROM
`nba_bipm_shots_data.SHOTS` AS a
LEFT JOIN
`nba_bipm_shots_data.PLAYERS` AS b
ON
a.shooter = b.shooter_name;
Valid: This query will process 19.6 MB when run.
Destination Table
Select Table
nba-bipm2019:nba_bipm_shots_data.detailed_shots_data
Write Preference
Write if empty
Append to table
Overwrite table
Results Size
Allow Large Results
Results Schema
Flatten Results
Query Caching
Use Cached Results
Query Priority
Interactive
Batch
UDF Source URIs
Edit
Maximum Bytes Billed
Project Default
SQL Dialect
Use Legacy SQL
Destination Encryption
Default
Processing Location
Unspecified
RUN QUERY
Save Query
Save View
Format Query
Schedule Query
Hide Options
Query complete (6.1s elapsed, 19.6 MB processed)
Results
Details
Download as CSV
Download as JSON
Save as Table
Save to Google Sheets
Row
player_position
home
loc_x
home_team
shot_type
points
away_team
loc_y
time
date
shooter
quarter
outcome
shooter_name
1
PG
false
167.0
DET
Jump Shot
3
ATL
20.0
11:17
2017-01-18
Gary Neal
4
MISSED
null

# Scenario 2 (Apache Hive + Power BI)



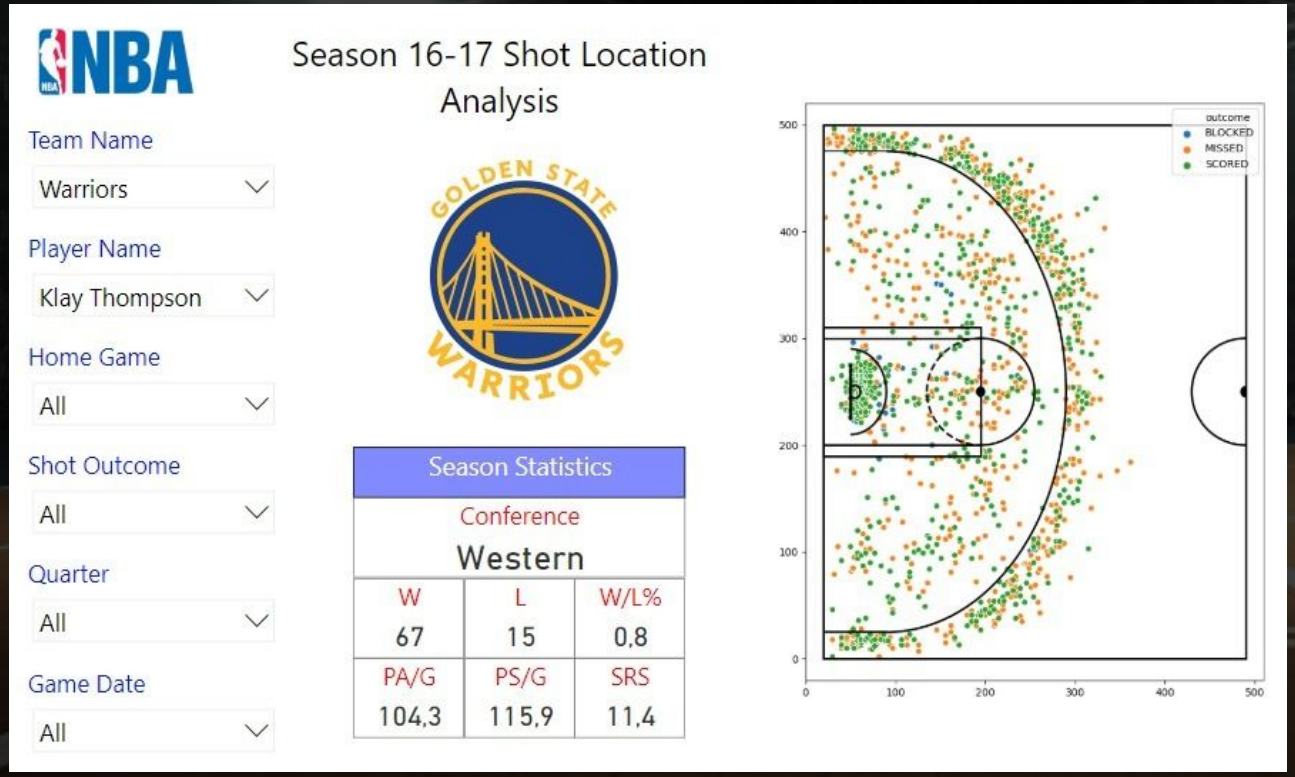
```
adrian_villegas_mazo@seneca3:~$ hadoop fs -ls
Found 24 items
drwx------   - adrian_villegas_mazo student          0 2019-07-15 16:00 .Trash
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-05-29 15:37 .sparkStaging
drwx------   - adrian_villegas_mazo student          0 2019-07-03 23:19 .staging
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-05-22 19:07 color
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-05-14 12:45 constitution_hdfs
-rw-r--r--   3 adrian_villegas_mazo student    2772143 2019-05-22 18:39 diamonds.csv
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-05-08 11:23 hadoop
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-05-08 17:58 inputdata
-rwxr-xr-x   3 adrian_villegas_mazo student       1640 2019-05-08 17:58 mapper.py
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-05-08 11:26 myhdfs
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-07-03 21:46 nba_players
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-07-03 23:19 nba_shots
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-05-08 18:19 outputdata
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-05-08 18:38 outputdata2
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-05-08 18:59 outputdata3
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-05-08 19:13 outputdata4
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-05-08 19:27 outputdata5
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-05-08 19:35 outputdata6
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-05-08 19:45 outputdata7
-rw-r--r--   3 adrian_villegas_mazo student      46806 2019-07-03 21:45 players_dataframe.csv
-rw-r--r--   3 adrian_villegas_mazo student       2216 2019-05-08 19:41 reducer.py
-rw-r--r--   3 adrian_villegas_mazo student   16978122 2019-07-03 23:17 shots_dataframe.csv
-rw-r--r--   3 adrian_villegas_mazo student  183292235 2019-05-14 13:12 wh_visits.txt
drwxr-xr-x   - adrian_villegas_mazo student          0 2019-05-22 18:25 words
adrian_villegas_mazo@seneca3:~$
```

# Scenario 3 (PySpark + Zeppelin)



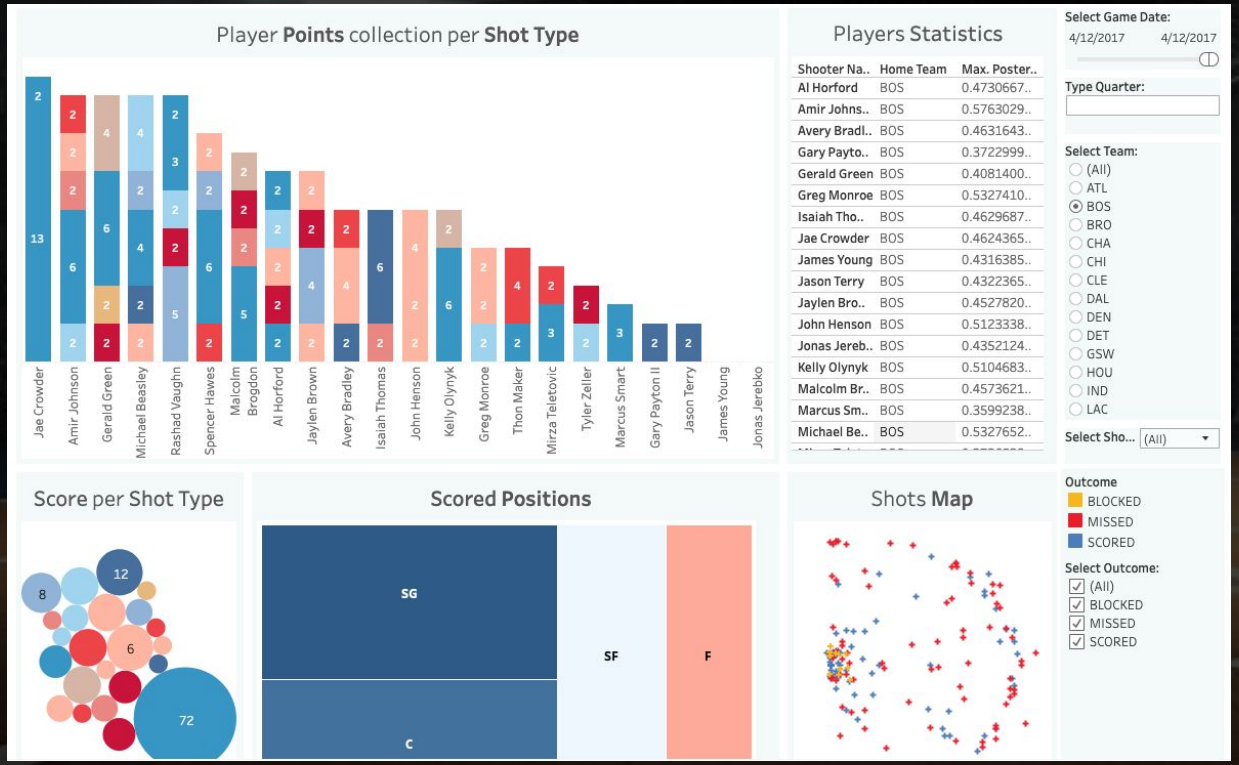

# 4. Data Visualization and Analysis

# Power BI Dashboard




[Click here to see an interactive dashboard]

# Tableau Dashboard




[click here to see an interactive dashboard]

# Zeppelin Notebook

Zeppelin Notebook Job Search user1

NBA_BigDataProject Head default

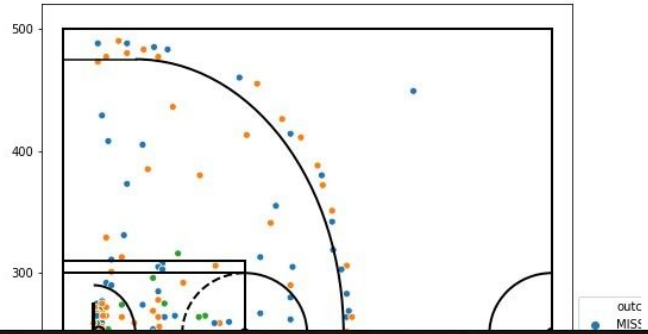
NBA_BigDataProject

FINISHED

```
%spark.pyspark
from matplotlib.patches import Circle, Rectangle, Arc
import matplotlib.pyplot as plt

def draw_half_court(ax=None, color='black', lw=2, outer_lines=False):
    # If an axes object isn't provided to plot onto, just get current one
    if ax is None:
        ax = plt.gca()

    backboard1 = Rectangle((50, 225), -1, 50, linewidth=lw, color=color)

    hoop1 = Circle((55, 250), radius=6, linewidth=lw, color=color, fill=False)

    restricted1 = Arc((50, 250), 80, 80, angle=270, theta1=0, theta2=180, linewidth=lw,
                      color=color)

    freethrow1_outer = Arc((195, 250), 100, 120, angle=270, theta1=0, theta2=180,
        linewidth=lw,
                      color=color)

    freethrow1_inner = Arc((195, 250), 100, 120, angle=270, theta1=180, theta2=0,
        linewidth=lw,
                      color=color, linestyle='dashed')

    freethrow1_point = Circle((195, 250), radius=4, linewidth=lw, color=color, fill=True
        )

    threepoint1 = Arc((90, 250), 450, 400, angle=270, theta1=0, theta2=180, linewidth=lw
        ,
                      color=color)

    innerbox1 = Rectangle((20, 200), 175, 100, linewidth=lw, color=color,
                          fill=False)

    outbox1 = Rectangle((20, 190), 175, 120, linewidth=lw, color=color,
                        fill=False)

    half_court_circle  = Arc((490, 250), 100, 120, angle=270, theta1=180, theta2=0,
```

FINISHED

```
%spark.pyspark
nba_shots_subset = nba_shots[:200]

plt.figure(figsize=(8,10))
sns.scatterplot(x='loc_x',y='loc_y',hue='outcome',data=nba_shots_subset)
draw_half_court(outer_lines=True)
plt.xlim(0,510)
plt.ylim(-20,520)
plt.xlabel('')
plt.ylabel('')
plt.legend(loc='center left', bbox_to_anchor=(1, 0.5), ncol=1)
plt.show()
```



# 5. Technologies Evaluation


NBA
tableau
SOFTWARE
HIVE
c:\>
Google
BigQuery
APACHE
Spark
hadoop
SQL

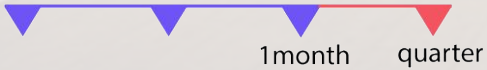
## Google Cloud Storage

### BigQuery

- ✓ Integration with Tableau
- ✓ On-demand pricing
- ✓ Free processing of small amount of data

- ✓ BigQuery is an asset to analyze billions of rows
- ✓ 3-5 seconds of response time
- ✓ Storing images in Google Storage to be processed by Google BigQuery is not possible

1 TB free querying
10 GB storage




## Apache Hive

### PySpark

- ✓ Python is slow compared to Scala for Spark Jobs
- ✓ Replaces the Map/Reduce function process with Python language code, which saves time and makes it easier to write

- ✓ Installation of ODBC driver to connect Power BI or Tableau
- ✓ Limited available number of functions built-in into the package

python

Spark

## Tableau

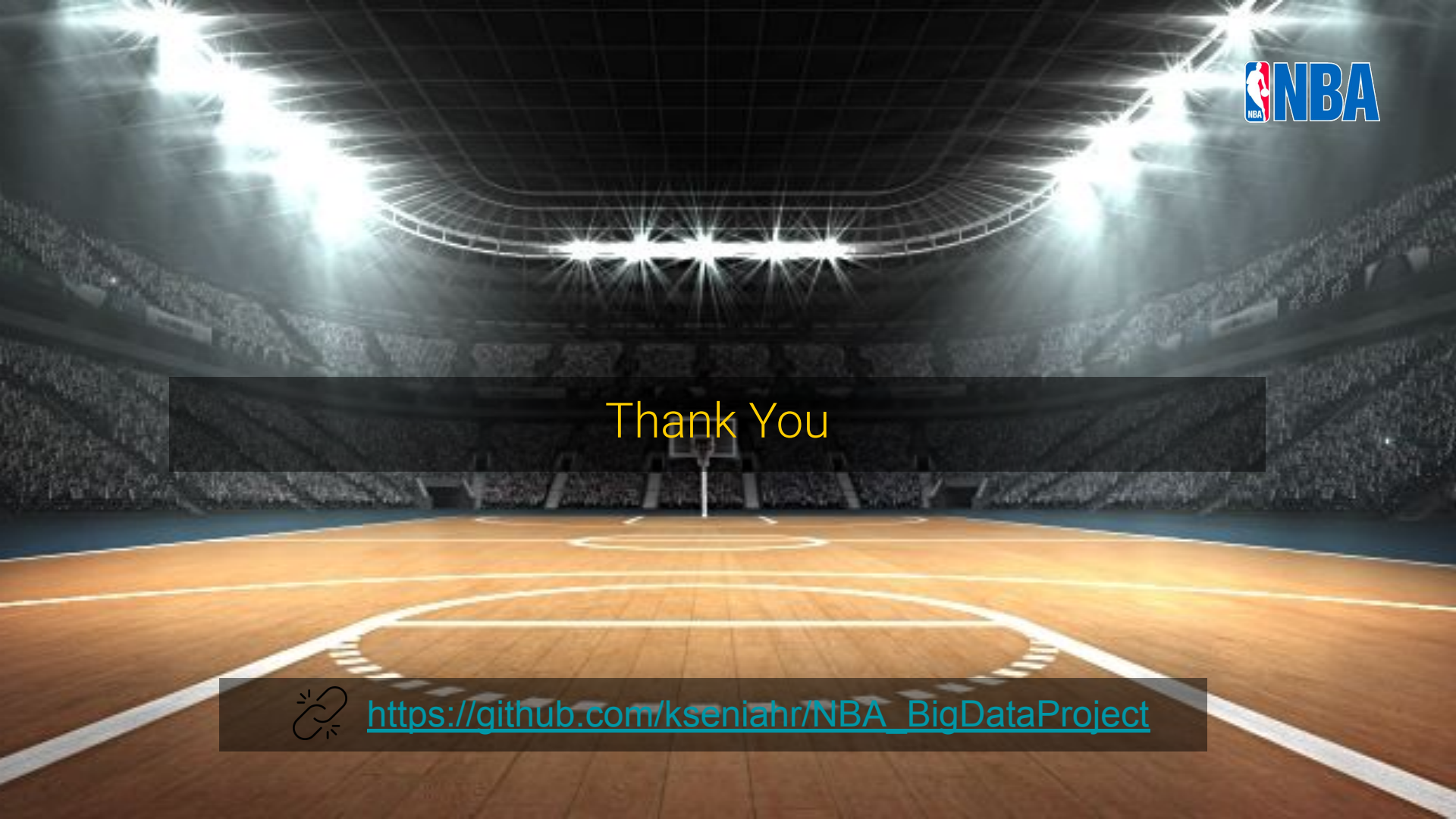
- ✓ Integration with BigQuery
- ✓ User-friendly interface
- ✗ Creating visuals using Python is not possible
- ✗ Integration with Hive only through ODBC connector

## Power BI

- ✓ Integration with Hive and with several other types of data sources
- ✓ User-friendly interface
- ✓ Allows schema creation regardless of the source with entity-relation models
- ✓ Allows custom Python visuals in the dashboards
- ✗ Free version does not support publishing dashboards with Python visualizations to Web

## Zeppelin

- ✓ Integration with Hive
- ✓ User-friendly interface
- ✓ Creating visuals using Python
- ✓ Huge amount of interpreters for different types of data and different types of querying/programming languages
- ✗ Free version does not support publishing dashboards with Python visualizations to Web


Thank You
https://github.com/kseniahr/NBA_BigDataProject


Q&A