

UE20CS581 – Advanced Big Data Analytics

Project-1

Title: IPL Analysis using Spark

Professor: Dr. K.V. Subramaniam

Submitted by: Suraj S (PES1PG20CS046)

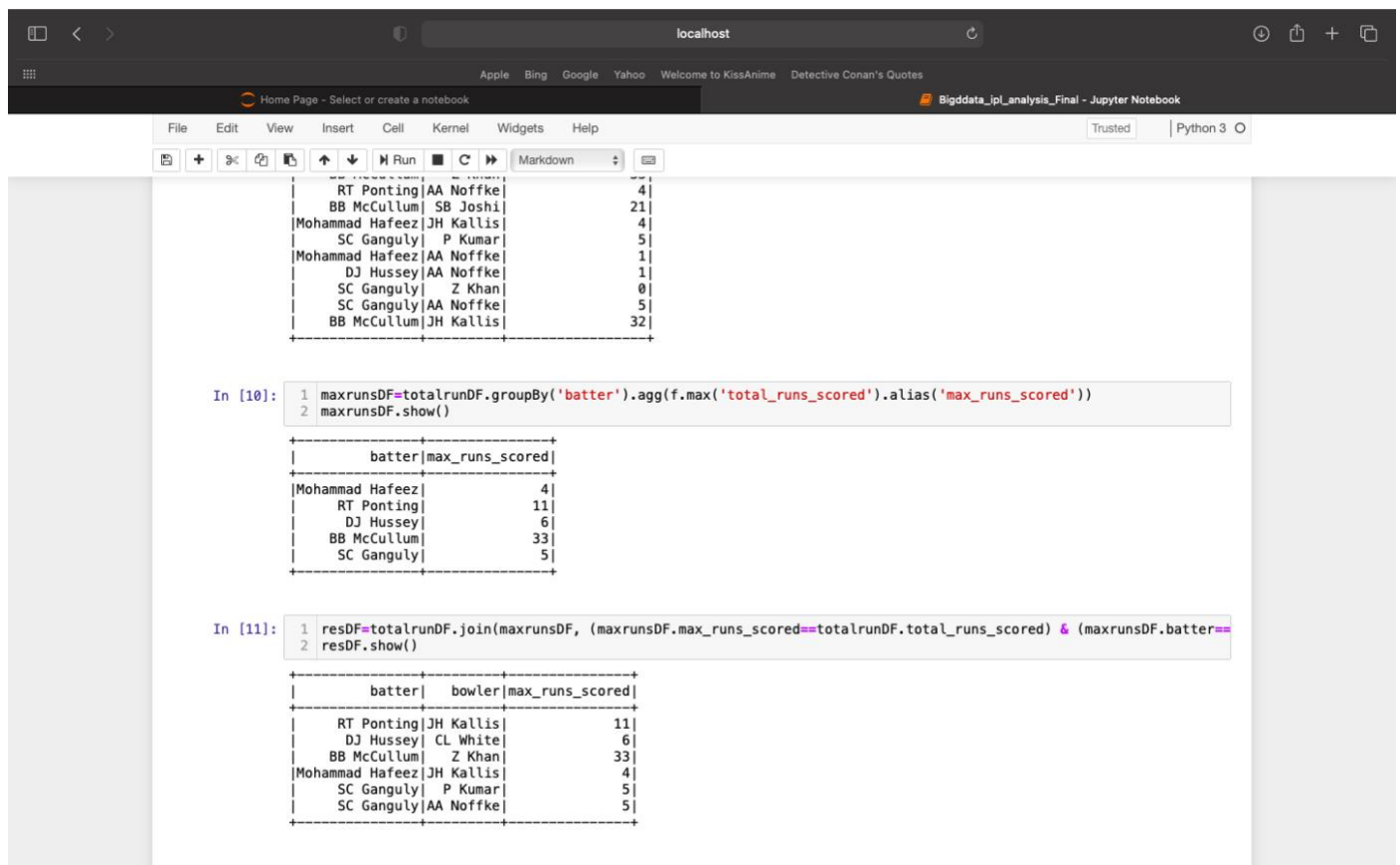
Divya M S (PES1PG20CS053)

GitHub Repository: <https://github.com/surajs757/IPL-Analysis>

Summary:

- Learning how to install spark.
- Run sample application.
- Exploring Json file based data and understanding spark.
- Evaluation of data.
- Understanding the schema.
- IPL analysis.

Output: Maximum Runs Score and Maximum Runs Scored per Bowler



The screenshot shows a Jupyter Notebook interface with a browser window at the top displaying 'localhost'. The notebook has a dark theme and shows the following content:

File Edit View Insert Cell Kernel Widgets Help

Home Page - Select or create a notebook | Bigdata_ipl_analysis_Final - Jupyter Notebook

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Cell 1:

```
RT Ponting|AA Noffke|4|
BB McCullum|SB Joshi|21|
Mohammad Hafeez|JH Kallis|4|
SC Ganguly|P Kumar|5|
Mohammad Hafeez|AA Noffke|1|
DJ Hussey|AA Noffke|1|
SC Ganguly|Z Khan|0|
SC Ganguly|AA Noffke|5|
BB McCullum|JH Kallis|32|
```

In [10]:

```
1 maxrunsDF=totalrunDF.groupBy('batter').agg(f.max('total_runs_scored').alias('max_runs_scored'))
2 maxrunsDF.show()
```

Output:

```
batter|max_runs_scored|
Mohammad Hafeez|4|
RT Ponting|11|
DJ Hussey|6|
BB McCullum|33|
SC Ganguly|5|
```

In [11]:

```
1 resDF=totalrunDF.join(maxrunsDF, (maxrunsDF.max_runs_scored==totalrunDF.total_runs_scored) & (maxrunsDF.batter==
2 resDF.show())
```

Output:

```
batter|bowler|max_runs_scored|
RT Ponting|JH Kallis|11|
DJ Hussey|CL White|6|
BB McCullum|Z Khan|33|
Mohammad Hafeez|JH Kallis|4|
SC Ganguly|P Kumar|5|
SC Ganguly|AA Noffke|5|
```