Notebook to run hub ingestion using lake tables

The notebook will load match flows, score cards etc

Assumptions

- **Ideally we should be doing incremental loads fetching new data since the last lake load using watermark, but in this example we will doing full load everytime.**
- **We should have versioning logic to support Type 2 SCD using metadata columns like records version and record sha, but in this since we will be doing full load everytime we skipping this.**

In [1]:

```
# %%configure
# {
#     "--datalake-formats":"delta",
#     "--conf":"spark.sql.extensions=io.delta.sql.DeltaSparkSessionExtension --conf spark
.sql.catalog.spark_catalog=org.apache.spark.sql.delta.catalog.DeltaCatalog"
# }
```

```
Welcome to the Glue Interactive Sessions Kernel
For more information on available magic commands, please type %help in any new cell.

Please view our Getting Started page to access the most up-to-date information on the Int
eractive Sessions kernel: https://docs.aws.amazon.com/glue/latest/dg/interactive-sessions
.html
Installed kernel version: 0.37.3
```

In [6]:

```
%idle_timeout 200
%glue_version 3.0
%worker_type G.1X
%number_of_workers 2
```

```
Current idle_timeout is 2800 minutes.
idle_timeout has been set to 200 minutes.
Setting Glue version to: 3.0
Previous worker type: G.1X
Setting new worker type to: G.1X
Previous number of workers: 5
Setting new number of workers to: 2
```

In [1]:

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

import boto3

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
```

```
Authenticating with environment variables and user-defined glue_role_arn: arn:aws:iam::68
7003041478:role/orka-glue-role
Trying to create a Glue session for the kernel.
Worker Type: G.1X
Number of Workers: 2
Session ID: faa9da1b-7918-4fa0-9a9d-0d4ef2da36ca
```

```
Job Type: glueetl
Applying the following default arguments:
--glue_kernel_version 0.37.3
--enable-glue-datacatalog true
Waiting for session faa9da1b-7918-4fa0-9a9d-0d4ef2da36ca to get into ready status...
Session faa9da1b-7918-4fa0-9a9d-0d4ef2da36ca has been created.
```

In [2]:

```
from pyspark.sql.functions import *
from pyspark.sql.window import Window
```

In [25]:

```
# Create a Glue client
glue_client = boto3.client('glue')
```

Variables

In [26]:

```
# Crawler IAM role ARN
crawler_role = 'arn:aws:iam::687003041478:role/orka-glue-role'  # Replace with your IAM r
ole ARN

# DB variables
bronze_database_name = "orka_warehouse_bronze" # Specify the lake/bronze glue catalog dat
abase name

# Target Player info table
player_info_table_name = "player_info" # Specify player info target table name in lake
# Target Deliveries table
deliveries_table_name = "deliveries"
# Target matches table
matches_table_name = "matches"

# Specify the hub target database name
hub_database_name = 'orka_warehouse_silver'
```

Utility functions

In [29]:

```
def crawl_tables(glue_client, df_dic, data_loc, database_name):
    """Crawler to crawl tables stored in s3 after processing
    @param: glue_client Our glue client
    @param: df_dic dictionary containing mapping of dataframe name and dataframe object
    @param: data_loc Database location
    @param: database_name Name of database in glue catalog where data is to be loaded
    """
    for table, data_df in df_dic.items():
        crawler_name = f"{table}_crwl"
        # Specify the S3 target for the crawler
        s3_target_path = f"{data_loc}/{table}"

        # Create the crawler
        response = glue_client.create_crawler(
            Name=crawler_name,
            Role=crawler_role,
            Targets={
                'S3Targets': [
                    {
                        'Path': s3_target_path
                    }
                ]
            },
            DatabaseName= database_name,
            SchemaChangePolicy={
                'UpdateBehavior': 'UPDATE_IN_DATABASE',
                'DeleteBehavior': 'DEPRECATE_IN_DATABASE'
            }
        )

        # Start the crawler
        glue_client.start_crawler(Name=crawler_name)


def load_tables(df_dic,data_loc):
    """Function to load dataframes in dictionary to s3
    @param: df_dic dictionary containing mapping of dataframe name and dataframe object
    @param: data_loc Database location
    """
    for table, data_df in df_dic.items():
        data_df.write.format("parquet").mode("overwrite").save(f"{data_loc}/{table}")
```

Loading required lake tables to be used as source for hub ingestion

Loading deliveries

In [11]:

```
deliveries_dyf = glueContext.create_dynamic_frame.from_catalog(database=bronze_database_n
ame, table_name=deliveries_table_name)
deliveries_df = deliveries_dyf.toDF()
deliveries_df.show()
```

```
+--------+------+--------------------+--------------------+----+----+--------------+---
-------------+---------------+-------------+---------+--------+-----------+-----------+
------------+------------+----------+----------+----------------+--------------+-------
--+
|match_id|inning|        batting_team|        bowling_team|over|ball|       batsman|
non_striker|         bowler|is_super_over|wide_runs|bye_runs|legbye_runs|noball_runs|pena
lty_runs|batsman_runs|extra_runs|total_runs|player_dismissed|dismissal_kind|  fielder|
+--------+------+--------------------+--------------------+----+----+--------------+---
-------------+---------------+-------------+---------+--------+-----------+-----------+
------------+------------+----------+----------+----------------+--------------+-------
--+
|     457|     2| Chennai Super Kings|      Mumbai Indians|   4|   1|       M Vijay|
DJ Bravo|     MG Johnson|            0|        0|       0|          0|          0|
0|           0|         0|         0|            null|          null|     null|
|     368|     1|Kolkata Knight Ri...| Chennai Super Kings|   8|   2|     G Gambhir|
BB McCullum|      RA Jadeja|            0|        0|       0|          0|          0|
0|           0|         0|         0|            null|          null|     null|
|       2|     2|Rising Pune Super...|      Mumbai Indians|   6|   6|     AM Rahane|
SPD Smith| MJ McClenaghan|            0|        0|       0|          0|          0|
0|           1|         0|         1|            null|          null|     null|
|     536|     2|     Kings XI Punjab|   Rajasthan Royals|  12|   1|      SE Marsh|
DA Miller|    JP Faulkner|            0|        0|       0|          0|          0|
0|           1|         0|         1|            null|          null|     null|
|   11330|     2| Chennai Super Kings|Royal Challengers...|  19|   4|      MS Dhoni|
DJ Bravo|        N Saini|            0|        0|       0|          0|          0|
0|           2|         0|         2|            null|          null|     null|
|     213|     1|Kolkata Knight Ri...|    Delhi Daredevils|   8|   4|    SC Ganguly|
CH Gayle| PD Collingwood|            0|        0|       0|          0|          0|
0|           1|         0|         1|            null|          null|     null|
|      30|     1|Royal Challengers...|       Gujarat Lions|   9|   1|     KM Jadhav|  AB
de Villiers|      RA Jadeja|            0|        0|       0|          0|          0|
0|           1|         0|         1|            null|          null|     null|
|     470|     2| Chennai Super Kings|      Mumbai Indians|  16|   3|  F du Plessis|
BB McCullum|Harbhajan Singh|            0|        0|       0|          0|          0|
0|           2|         0|         2|            null|          null|     null|
|     596|     2|Kolkata Knight Ri...|Rising Pune Super...|   4|   5|      SA Yadav| Shak
ib Al Hasan|       R Ashwin|            0|        0|       0|          0|          0|
0|           6|         0|         6|            null|          null|     null|
|   11330|     1|Royal Challengers...| Chennai Super Kings|   4|   5|AB de Villiers|
PA Patel|      SN Thakur|            0|        0|       0|          0|          0|
0|           1|         0|         1|            null|          null|     null|
|     574|     1|Royal Challengers...|    Rajasthan Royals|  14|   5| Mandeep Singh|  AB
de Villiers|      SR Watson|            0|        0|       0|          0|          0|
0|           1|         0|         1|            null|          null|     null|
|      90|     1|Royal Challengers...|      Mumbai Indians|   2|   1|    SP Goswami|
MV Boucher|        A Nehra|            0|        0|       0|          0|          0|
0|           2|         0|         2|            null|          null|     null|
|     418|     2|Kolkata Knight Ri...| Chennai Super Kings|   9|   1|      MS Bisla|
JH Kallis|      RA Jadeja|            0|        0|       0|          0|          0|
0|           1|         0|         1|            null|          null|     null|
|     623|     2|    Delhi Daredevils|      Mumbai Indians|  15|   2|     CH Morris|
A Mishra|  R Vinay Kumar|            0|        0|       0|          0|          0|
0|           0|         0|         0|            null|          null|     null|
|    7923|     2|    Delhi Daredevils| Chennai Super Kings|  18|   4|       RR Pant|
V Shankar|        L Ngidi|            0|        0|       0|          0|          0|
0|           0|         0|         0|         RR Pant|        caught|RA Jadeja|
|   11138|     2|Kolkata Knight Ri...| Sunrisers Hyderabad|  20|   3|        S Gill|
AD Russell|Shakib Al Hasan|            0|        0|       0|          0|          0|
0|           6|         0|         6|            null|          null|     null|
|     565|     1| Chennai Super Kings|    Delhi Daredevils|   2|   6|      DR Smith|
BB McCullum|         Z Khan|            0|        0|       0|          0|          0|
0|           0|         0|         0|            null|          null|     null|
|     450|     1|     Kings XI Punjab|      Mumbai Indians|  10|   4|      SE Marsh|   Az
har Mahmood|     KA Pollard|            0|        0|       0|          0|          0|
0|           0|         0|         0|            null|          null|     null|
|     225|     1|     Kings XI Punjab|     Deccan Chargers|  10|   4| KC Sangakkara|DPMD
Jayawardene|      A Symonds|            0|        0|       0|          0|          0|
0|           1|         0|         1|            null|          null|     null|
|     518|     1|      Mumbai Indians|Kolkata Knight Ri...|   3|   3|     RG Sharma|
AP Tare|       UT Yadav|            0|        0|       0|          0|          0|
0|           0|         0|         0|            null|          null|     null|
+--------+------+--------------------+--------------------+----+----+--------------+---
-------------+---------------+-------------+---------+--------+-----------+-----------+
------------+------------+----------+----------+----------------+--------------+-------
--+
only showing top 20 rows
```

Loading player info

In [6]:

```
player_info_dyf = glueContext.create_dynamic_frame.from_catalog(database=bronze_database_
name, table_name=player_info_table_name)
player_info_df = player_info_dyf.toDF()
player_info_df.show(2)
```

```
+-----+------+-------------+---------+--------------------+---------+----------+-----+-
------------+---+--------------------+-------------+--------------------+-----+------+-
----------------+------------------+----------------+------------------+---------------
-+-----------------+----------------+----------------+-----------------+---------------
-+----------------+----------------+----------------+----------------+----------------+
-----------------+---------------+-----------------+---------------+----------------+--
-------------+---------------+----------------+---------------+---------------+--------
-------+---------------+---------------+-----------------+------------------+----------
------+------------------+----------------+-----------------+----------------+---------
-------+-----------------+----------------+----------------+----------------+----------
------+----------------+-----------------------+------------------------+--------------
--------+------------------------+----------------------+-----------------------+------
----------------+----------------------+-----------------------+----------------------+
----------------------+----------------------+----------------------+------------------
----+------------------+-------------------+-----------------+-------------------+-----
------------+------------------+-----------------+-----------------+------------------+
-----------------+-----------------+-----------------+-----------------+---------------
--+----------------+-----------------+---------------+-----------------+---------------
+----------------+---------------+---------------+----------------+---------------+----
-----------+---------------+---------------+---------------+-----------------+---------
---------+-------------------+------------------+------------------+-----------------+-
----------------+-----------------+------------------+----------------+----------------
```

```
+----------------+----------------+----------------+-----------------+-----------------
-+-----------------+-----------------+----------------+----------------+---------------
-+-----------------+---------------+---------------+---------------+---------------+---
--------------+------------------+-------------------+------------------+--------------
----+-----------------+-----------------+-----------------+------------------+---------
-------+----------------+----------------+----------------+-----------------------+----
--------------------+-------------------------+------------------------+---------------
---------+-----------------------+-----------------------+-----------------------+-----
-------------------+----------------------+----------------------+---------------------
-+----------------------+------------------+-------------------+--------------------+--
-----------------+-------------------+------------------+------------------+-----------
-------+-------------------+-----------------+-----------------+-----------------+-----
------------+----------------+-----------------+------------------+-----------------+--
---------------+----------------+----------------+----------------+-----------------+--
-------------+---------------+---------------+---------------+
|index|    id|         name|  country|           full_name|birthdate|birthplace| died|da
te_of_death|age|         major_teams|batting_style|       bowling_style|other|awards|batt
ing_tests_mat|batting_tests_inns|batting_tests_no|batting_tests_runs|batting_tests_hs|bat
ting_tests_ave|batting_tests_bf|batting_tests_sr|batting_tests_100|batting_tests_50|batti
ng_tests_4s|batting_tests_6s|batting_tests_ct|batting_tests_st|batting_odis_mat|batting_o
dis_inns|batting_odis_no|batting_odis_runs|batting_odis_hs|batting_odis_ave|batting_odis_
bf|batting_odis_sr|batting_odis_100|batting_odis_50|batting_odis_4s|batting_odis_6s|batti
ng_odis_ct|batting_odis_st|batting_t20is_mat|batting_t20is_inns|batting_t20is_no|batting_
t20is_runs|batting_t20is_hs|batting_t20is_ave|batting_t20is_bf|batting_t20is_sr|batting_t
20is_100|batting_t20is_50|batting_t20is_4s|batting_t20is_6s|batting_t20is_ct|batting_t20i
s_st|batting_first_class_mat|batting_first_class_inns|batting_first_class_no|batting_firs
t_class_runs|batting_first_class_hs|batting_first_class_ave|batting_first_class_bf|battin
g_first_class_sr|batting_first_class_100|batting_first_class_50|batting_first_class_4s|ba
tting_first_class_6s|batting_first_class_ct|batting_first_class_st|batting_list_a_mat|bat
ting_list_a_inns|batting_list_a_no|batting_list_a_runs|batting_list_a_hs|batting_list_a_a
ve|batting_list_a_bf|batting_list_a_sr|batting_list_a_100|batting_list_a_50|batting_list_
a_4s|batting_list_a_6s|batting_list_a_ct|batting_list_a_st|batting_t20s_mat|batting_t20s_
inns|batting_t20s_no|batting_t20s_runs|batting_t20s_hs|batting_t20s_ave|batting_t20s_bf|b
atting_t20s_sr|batting_t20s_100|batting_t20s_50|batting_t20s_4s|batting_t20s_6s|batting_t
20s_ct|batting_t20s_st|bowling_tests_mat|bowling_tests_inns|bowling_tests_balls|bowling_t
ests_runs|bowling_tests_wkts|bowling_tests_bbi|bowling_tests_bbm|bowling_tests_ave|bowlin
g_tests_econ|bowling_tests_sr|bowling_tests_4w|bowling_tests_5w|bowling_tests_10|bowling_
odis_mat|bowling_odis_inns|bowling_odis_balls|bowling_odis_runs|bowling_odis_wkts|bowling
_odis_bbi|bowling_odis_bbm|bowling_odis_ave|bowling_odis_econ|bowling_odis_sr|bowling_odi
s_4w|bowling_odis_5w|bowling_odis_10|bowling_t20is_mat|bowling_t20is_inns|bowling_t20is_b
alls|bowling_t20is_runs|bowling_t20is_wkts|bowling_t20is_bbi|bowling_t20is_bbm|bowling_t2
0is_ave|bowling_t20is_econ|bowling_t20is_sr|bowling_t20is_4w|bowling_t20is_5w|bowling_t20
is_10|bowling_first_class_mat|bowling_first_class_inns|bowling_first_class_balls|bowling_
first_class_runs|bowling_first_class_wkts|bowling_first_class_bbi|bowling_first_class_bbm
|bowling_first_class_ave|bowling_first_class_econ|bowling_first_class_sr|bowling_first_cl
ass_4w|bowling_first_class_5w|bowling_first_class_10|bowling_list_a_mat|bowling_list_a_in
ns|bowling_list_a_balls|bowling_list_a_runs|bowling_list_a_wkts|bowling_list_a_bbi|bowlin
g_list_a_bbm|bowling_list_a_ave|bowling_list_a_econ|bowling_list_a_sr|bowling_list_a_4w|b
owling_list_a_5w|bowling_list_a_10|bowling_t20s_mat|bowling_t20s_inns|bowling_t20s_balls|
bowling_t20s_runs|bowling_t20s_wkts|bowling_t20s_bbi|bowling_t20s_bbm|bowling_t20s_ave|bo
wling_t20s_econ|bowling_t20s_sr|bowling_t20s_4w|bowling_t20s_5w|bowling_t20s_10|
+-----+------+-------------+---------+--------------------+---------+----------+-----+-
------------+---+--------------------+-------------+--------------------+-----+------+-
----------------+------------------+----------------+------------------+---------------
-+-----------------+----------------+----------------+-----------------+---------------
-+----------------+----------------+----------------+----------------+----------------+
-----------------+---------------+-----------------+---------------+----------------+--
-------------+---------------+----------------+---------------+---------------+--------
-------+---------------+---------------+-----------------+------------------+----------
------+------------------+----------------+-----------------+----------------+---------
-------+-----------------+----------------+----------------+----------------+----------
------+----------------+-----------------------+------------------------+--------------
--------+------------------------+----------------------+-----------------------+------
----------------+----------------------+-----------------------+----------------------+
----------------------+----------------------+----------------------+------------------
----+------------------+-------------------+-----------------+-------------------+-----
------------+------------------+-----------------+-----------------+------------------+
-----------------+-----------------+-----------------+-----------------+---------------
--+----------------+-----------------+---------------+-----------------+---------------
+----------------+---------------+---------------+----------------+---------------+----
-----------+---------------+---------------+---------------+-----------------+---------
---------+-------------------+------------------+------------------+-----------------+-
```

```
----------------+-----------------+------------------+----------------+-----------------
+----------------+----------------+----------------+-----------------+------------------
-+-----------------+-----------------+-----------------+---------------+----------------
-+-----------------+----------------+----------------+-------------+---------------+----
--------------+------------------+-------------------+------------------+---------------
----+-----------------+-----------------+-----------------+------------------+----------
-------+----------------+----------------+----------------+-----------------------+-----
--------------------+-------------------------+------------------------+----------------
---------+-----------------------+-----------------------+-----------------------+------
-------------------+----------------------+----------------------+----------------------
-+----------------------+------------------+-------------------+--------------------+---
-----------------+-------------------+------------------+------------------+------------
-------+-------------------+-----------------+-----------------+-----------------+------
------------+----------------+-----------------+------------------+-----------------+---
---------------+----------------+----------------+----------------+-----------------+---
-------------+---------------+---------------+---------------+
|37362|527291| Sumit Ruikar|    India|Sumit Sureshrao R...| 06/06/90|    Nagpur|Alive|
null| 29|['Vidarbha,', 'Vi...|Left-hand bat|Slow left-arm ort...| null|    []|
null|              null|            null|              null|            null|
null|            null|            null|             null|            null|            nu
ll|            null|            null|            null|            null|             null
|           null|             null|           null|            null|           null|
null|            null|           null|           null|           null|           null|
null|             null|              null|            null|              null|
null|             null|            null|            null|             null|
null|            null|            null|            null|            null|
27|                    40.0|                     6|                   837.0|
67*|                  24.61|                  1754|                 47.71|
0|                   6.0|                    86|                   3.0|
5|                   0.0|                27|                 21|                6|
216.0|               18|              14.4|              292|            73.97|
0|              0.0|               16|              1.0|               18|
0.0|              13|              8.0|              2|             50.0|             20
|            8.33|             61|          81.96|               0|            0.0|
2|            0.0|              6|            0.0|             null|              null|
null|              null|              null|             null|             null|
null|              null|            null|            null|            null|
null|            null|             null|              null|             null|
null|            null|            null|            null|             null|           nul
l|           null|           null|           null|             null|              null|
null|              null|              null|             null|             null|
null|              null|            null|            null|            null|
null|                     27|                      43|                     5548|
2646|                      84|                  7/112|                 10/101|
31.5|                    2.86|                    66|                     1|
6|                     1|                27|                 27|                1475|
1072|                 36|            Jun-26|            Jun-26|             29.77|
4.36|             40.9|                0|                1|                0|
13|               13|               246|              288|                6|          Fe
b-14|          Feb-14|              48|             7.02|             41|              0
|              0|              0|
|48407|623629|Sanesh de Mel|Sri Lanka|       Sanesh de Mel| 11/10/94|   Colombo|Alive|
null| 25|   S.Thomas' College|Left-hand bat|                null| null|    []|
null|              null|            null|              null|            null|
null|            null|            null|             null|            null|            nu
ll|            null|            null|            null|            null|             null
|           null|             null|           null|            null|           null|
null|            null|           null|           null|           null|           null|
null|             null|              null|            null|              null|
null|             null|            null|            null|             null|
null|            null|            null|            null|            null|
null|                    null|                  null|                    null|
null|                   null|                  null|                  null|
null|                  null|                  null|                  null|
null|                  null|              null|               null|             null|
null|             null|              null|             null|             null|
null|             null|             null|             null|             null|
null|            null|             null|           null|             null|           nul
l|            null|           null|           null|            null|           null|
null|           null|           null|           null|             null|              nul
l|               null|              null|              null|             null|
null|             null|              null|            null|            null|
null|             null|              null|            null|            null|
null|            null|            null|             null|              null|
null|             null|            null|            null|            null|
null|           null|           null|           null|           null|             null|
null|               null|              null|              null|             null|
null|             null|              null|            null|            null|
null|            null|                   null|                    null|
null|                    null|                    null|                   null|
null|                   null|                    null|                  null|
null|                  null|                  null|              null|               nu
ll|                null|               null|               null|              null|
null|              null|               null|             null|             null|
null|             null|            null|             null|              null|
null|             null|            null|            null|            null|
null|           null|           null|           null|
+-----+------+-------------+---------+--------------------+---------+----------+-----+-
------------+---+--------------------+-------------+--------------------+-----+------+-
----------------+------------------+----------------+------------------+---------------
-+-----------------+----------------+----------------+-----------------+---------------
-+----------------+----------------+----------------+----------------+----------------+
-----------------+---------------+-----------------+---------------+----------------+--
-------------+---------------+----------------+---------------+---------------+--------
-------+---------------+---------------+-----------------+------------------+----------
------+------------------+----------------+-----------------+----------------+---------
-------+-----------------+----------------+----------------+----------------+----------
------+----------------+-----------------------+------------------------+--------------
--------+------------------------+----------------------+-----------------------+------
----------------+----------------------+-----------------------+----------------------+
----------------------+----------------------+----------------------+------------------
----+------------------+-------------------+-----------------+-------------------+-----
------------+------------------+-----------------+-----------------+------------------+
-----------------+-----------------+-----------------+-----------------+---------------
--+----------------+-----------------+---------------+-----------------+---------------
+----------------+---------------+---------------+----------------+---------------+----
-----------+---------------+---------------+---------------+-----------------+---------
---------+-------------------+------------------+------------------+-----------------+-
----------------+-----------------+------------------+----------------+----------------
+----------------+----------------+----------------+-----------------+-----------------
-+-----------------+-----------------+----------------+----------------+---------------
-+-----------------+---------------+---------------+---------------+---------------+---
--------------+------------------+-------------------+------------------+--------------
----+-----------------+-----------------+-----------------+------------------+---------
-------+----------------+----------------+----------------+-----------------------+----
--------------------+-------------------------+------------------------+---------------
---------+-----------------------+-----------------------+-----------------------+-----
-------------------+----------------------+----------------------+---------------------
-+----------------------+------------------+-------------------+--------------------+--
-----------------+-------------------+------------------+------------------+-----------
-------+-------------------+-----------------+-----------------+-----------------+-----
------------+----------------+-----------------+------------------+-----------------+--
---------------+----------------+----------------+----------------+-----------------+--
-------------+---------------+---------------+---------------+
only showing top 2 rows
```

Loading matches

In [7]:

```
matches_dyf = glueContext.create_dynamic_frame.from_catalog(database=bronze_database_name
, table_name=matches_table_name)
matches_df = matches_dyf.toDF()
matches_df.show(2)
```

```
+-------------+---------------+---------+---------------+-------+--------+------+------
---------+--------------+-----------+----------+-----------+---+----------+------------
---+-------------------+--------------------+-------------------+
|toss_decision|         winner|     city|player_of_match|umpire3|  Season|result|
umpire2|win_by_wickets|win_by_runs|dl_applied|    umpire1| id|      date|          team1|
team2|               venue|        toss_winner|
+-------------+---------------+---------+---------------+-------+--------+------+------
---------+--------------+-----------+----------+-----------+---+----------+------------
---+-------------------+--------------------+-------------------+
---+--------------------+---------------------+-------------------+
|        field|Kings XI Punjab|  Chennai|       J Theron|        |IPL-2010|   tie|      D
J Harper|             0|          0|         0|K Hariharan|190|21-03-2010|Kings XI Punjab
|Chennai Super Kings|MA Chidambaram St...|Chennai Super Kings|
|          bat|Deccan Chargers|Centurion|      RG Sharma|        |IPL-2009|normal|HDPK Dha
rmasena|             0|         19|         0|  MR Benson|147|06-05-2009|Deccan Chargers|
Mumbai Indians|     SuperSport Park|    Deccan Chargers|
+-------------+---------------+---------+---------------+--------+--------+------+------
---------+-------------+-----------+----------+-----------+---+----------+------------
---+--------------------+--------------------+--------------------+
only showing top 2 rows
```

Building silver tables

Calculating extra fields

In [12]:

```
# is the ball countable
deliveries_df = deliveries_df.withColumn("is_real_ball", when((col("wide_runs") == 0) &
(col("noball_runs") == 0), 1).otherwise(0))
```

Crating temp views

In [14]:

```
deliveries_df.createOrReplaceTempView('deliveries')
```

In [30]:

```
matches_df.createOrReplaceTempView('matches')
```

Building match flow

Match flow inning stat

In [17]:

```
# Powerplay calculation
inning_win = Window.partitionBy("match_id", "inning")

match_flow_df_inning = deliveries_df.withColumn("powerplay_run", when(col("over") <=6, c
ol("total_runs")).otherwise(0))\
    .withColumn("powerplay_wicket", when((col("over") <=6) & (col("player_dismissed").is
NotNull()), 1).otherwise(0))\
    .withColumn("total_powerplay_run", sum("powerplay_run").over(inning_win))\
    .withColumn("total_powerplay_wicket", sum("powerplay_wicket").over(inning_win))



def calculate_balls_by_x_runs(match_flow_df_inning, run):
    # Runs after number of balls calculation
    inning_over_ball_win = Window.partitionBy("match_id", "inning").orderBy("over","ball
")
    over_run_win = Window.partitionBy("match_id", "inning", f"over_{run}_flag")

    match_flow_df_inning = match_flow_df_inning.withColumn("cumulative_runs", sum(col("to
tal_runs")).over(inning_over_ball_win))\
        .withColumn("cumulative_balls", sum(col("is_real_ball")).over(inning_over_ball_wi
n))\
        .withColumn(f"over_{run}_flag", when(col("cumulative_runs") >=run, 1).otherwise(
lit(None)))\
        .withColumn(f"cumulative_balls_{run}", when(col("cumulative_runs") >=run, col("c
umulative_balls")).otherwise(lit(None)))\
        .withColumn(f"{run}_in_balls", min(col(f"cumulative_balls_{run}")).over(over_run
_win))\
        .withColumn(f"{run}_in_balls", min(col(f"{run}_in_balls")).over(over_run_win))\
        .withColumn(f"{run}_in_balls", last(col(f"{run}_in_balls"), ignorenulls=True).ov
er(inning_win))

    return match_flow_df_inning

match_flow_df_inning = calculate_balls_by_x_runs(match_flow_df_inning, 50)
match_flow_df_inning = calculate_balls_by_x_runs(match_flow_df_inning, 100)
match_flow_df_inning = calculate_balls_by_x_runs(match_flow_df_inning, 150)
match_flow_df_inning = calculate_balls_by_x_runs(match_flow_df_inning, 200)
match_flow_df_inning = calculate_balls_by_x_runs(match_flow_df_inning, 250)

match_flow_df_inning = match_flow_df_inning.select("match_id","inning","batting_team","bo
wling_team","total_powerplay_run","total_powerplay_wicket","50_in_balls","100_in_balls","
150_in_balls","200_in_balls","250_in_balls")\
                        .dropDuplicates()

match_flow_df_inning.filter("match_id=1").show()
```

```
+--------+------+--------------------+--------------------+-------------------+--------
--------------+-----------+------------+------------+------------+------------+
|match_id|inning|        batting_team|        bowling_team|total_powerplay_run|total_powe
rplay_wicket|50_in_balls|100_in_balls|150_in_balls|200_in_balls|250_in_balls|
+--------+------+--------------------+--------------------+-------------------+--------
--------------+-----------+------------+------------+------------+------------+
|       1|     2|Royal Challengers...| Sunrisers Hyderabad|                 54|
1|         32|          61|          95|        null|        null|
|       1|     1| Sunrisers Hyderabad|Royal Challengers...|                 59|
1|         32|          69|          89|         118|        null|
+--------+------+--------------------+--------------------+-------------------+--------
--------------+-----------+------------+------------+------------+------------+
```

Match flow batsman stat

In [20]:

```
def calculate_balls_by_x_batsman_runs(match_flow_df_batsman_stats, run):
    # Runs after number of balls calculation
    inning_batsman_over_ball_win = Window.partitionBy("match_id", "inning", "batsman").o
rderBy("over","ball")
    over_run_win = Window.partitionBy("match_id", "inning", "batsman", f"over_{run}_flag
")

    match_flow_df_batsman_stats = match_flow_df_batsman_stats.withColumn("cumulative_bat
sman_runs", sum(col("batsman_runs")).over(inning_batsman_over_ball_win))\
        .withColumn("cumulative_batsman_balls", sum(col("is_real_ball")).over(inning_bat
sman_over_ball_win))\
        .withColumn(f"over_{run}_flag", when(col("cumulative_batsman_runs") >=run, 1).ot
herwise(lit(None)))\
        .withColumn(f"cumulative_batsman_balls_{run}", when(col("cumulative_batsman_runs
") >=run, col("cumulative_batsman_balls")).otherwise(lit(None)))\
        .withColumn(f"batsman_{run}_in_balls", min(col(f"cumulative_batsman_balls_{run}"
)).over(over_run_win))

    return match_flow_df_batsman_stats

match_flow_df_batsman_stats = calculate_balls_by_x_batsman_runs(deliveries_df, 50)
match_flow_df_batsman_stats = calculate_balls_by_x_batsman_runs(match_flow_df_batsman_sta
ts, 100)
match_flow_df_batsman_stats = calculate_balls_by_x_batsman_runs(match_flow_df_batsman_sta
ts, 150)
match_flow_df_batsman_stats = calculate_balls_by_x_batsman_runs(match_flow_df_batsman_sta
ts, 200)
match_flow_df_batsman_stats = calculate_balls_by_x_batsman_runs(match_flow_df_batsman_sta
ts, 250)
```

```
match_flow_df_batsman_stats = match_flow_df_batsman_stats.filter("coalesce(batsman_50_in_
balls,batsman_100_in_balls,batsman_150_in_balls,batsman_200_in_balls,batsman_250_in_balls
) is not null")

match_flow_df_batsman_stats = match_flow_df_batsman_stats.select("match_id", "inning", "
batsman","batsman_50_in_balls","batsman_100_in_balls","batsman_150_in_balls","batsman_200
_in_balls","batsman_250_in_balls")\
                                .dropDuplicates()

match_flow_df_batsman_stats.filter("match_id=1").show()
```

```
+--------+------+------------+-------------------+--------------------+-----------------
----+--------------------+--------------------+
|match_id|inning|     batsman|batsman_50_in_balls|batsman_100_in_balls|batsman_150_in_bal
ls|batsman_200_in_balls|batsman_250_in_balls|
+--------+------+------------+-------------------+--------------------+-----------------
----+--------------------+--------------------+
|       1|     1|MC Henriques|                 34|                null|                n
ull|                null|                null|
|       1|     1|Yuvraj Singh|                 23|                null|                n
ull|                null|                null|
+--------+------+------------+-------------------+--------------------+-----------------
----+--------------------+--------------------+
```

Match flow wicket partnership stat

In [21]:

```
def calculate_wicket_partner_by_x_runs(match_flow_df_inning_wicket, run):
    # Runs after number of balls calculation
    inning_over_ball_win = Window.partitionBy("match_id", "inning").orderBy("over","ball
")
    inning_over_wicket_win = Window.partitionBy("match_id", "inning", "cumulative_player
_dismissed").orderBy("over","ball")
    over_run_win = Window.partitionBy("match_id", "inning", "cumulative_player_dismissed
", f"over_{run}_flag")

    match_flow_df_inning_wicket = match_flow_df_inning_wicket.withColumn("cumulative_pla
yer_dismissed", count(col("player_dismissed")).over(inning_over_ball_win))\
        .withColumn("cumulative_player_dismissed", when(col("cumulative_player_dismissed"
).isNull(),0).otherwise(col("cumulative_player_dismissed")))\
        .withColumn("cumulative_wicket_balls", sum(col("is_real_ball")).over(inning_over
_wicket_win))\
        .withColumn("cumulative_wicket_runs", sum(col("total_runs")).over(inning_over_wi
cket_win))\
        .withColumn(f"over_{run}_flag", when(col("cumulative_wicket_runs") >=run, 1).oth
erwise(lit(None)))\
        .withColumn(f"cumulative_wicket_balls_{run}", when(col("cumulative_wicket_runs")
>=run, col("cumulative_wicket_balls")).otherwise(lit(None)))\
        .withColumn(f"wicket_{run}_in_balls", min(col(f"cumulative_wicket_balls_{run}"))
.over(over_run_win))


    return match_flow_df_inning_wicket

match_flow_df_inning_wicket = calculate_wicket_partner_by_x_runs(deliveries_df, 50)

match_flow_df_inning_wicket = match_flow_df_inning_wicket.withColumn("batting_pair", sort
_array(array("batsman", "non_striker")))
match_flow_df_inning_wicket = match_flow_df_inning_wicket.select("match_id","inning",col(
"batting_pair")[0].alias("batsman_1"), col("batting_pair")[1].alias("batsman_2"),"cumula
tive_player_dismissed","wicket_50_in_balls")\
                                .filter("wicket_50_in_balls is not null")\
                                .dropDuplicates()

match_flow_df_inning_wicket.filter("match_id=1").show()
```

```
+--------+------+------------+-------------+---------------------------+----------------
---+
|match_id|inning|   batsman_1|    batsman_2|cumulative_player_dismissed|wicket_50_in_ball
s|
+--------+------+------------+-------------+----------------------------+---------------
---+
|       1|     2|    CH Gayle|Mandeep Singh|                          0|
32|
|       1|     2|   KM Jadhav|      TM Head|                          2|
27|
|       1|     1|MC Henriques|     S Dhawan|                          1|
37|
|       1|     1|MC Henriques| Yuvraj Singh|                          2|
25|
+--------+------+------------+-------------+----------------------------+---------------
---+
```

Score cards

Batting score card

In [22]:

```
battin_scorecard_df = spark.sql("""with bs as (
select distinct match_id, batting_team, batsman,
sum(total_runs) over(partition by match_id, batting_team) runs,
count(1) over(partition by match_id, batting_team)  total_balls,
count(player_dismissed) over(partition by match_id, batting_team) wickets,
sum(total_runs) over(partition by match_id, batting_team, batsman) player_runs,
count(1) over(partition by match_id, batting_team, batsman) player_balls,
sum(case when total_runs=4 then 1 else 0 end) over(partition by match_id, batting_team, b
atsman) fours,
sum(case when total_runs=6 then 1 else 0 end) over(partition by match_id, batting_team, b
atsman) sixes
from deliveries where match_id =1
)
select distinct *, round(player_runs/player_balls*100,2) sr, round(runs/total_balls*100,2
) rr from bs
""")

battin_scorecard_df.filter("match_id=1").show()
```

```
+--------+--------------------+-------------+----+-----------+-------+-----------+-----
-------+-----+-----+------+-------+
|match_id|        batting_team|      batsman|runs|total_balls|wickets|player_runs|player_
balls|fours|sixes|    sr|    rr|
+--------+--------------------+-------------+----+-----------+-------+-----------+-----
-------+-----+-----+------+-------+
|       1| Sunrisers Hyderabad|  BCJ Cutting| 207|        125|      4|         16|
6|    0|    2|266.67| 165.6|
|       1| Sunrisers Hyderabad|    DA Warner| 207|        125|      4|         17|
9|    2|    1|188.89| 165.6|
|       1| Sunrisers Hyderabad|     DJ Hooda| 207|        125|      4|         16|
12|    0|    1|133.33| 165.6|
|       1| Sunrisers Hyderabad| MC Henriques| 207|        125|      4|         52|
37|    3|    2|140.54| 165.6|
|       1| Sunrisers Hyderabad|     S Dhawan| 207|        125|      4|         41|
31|    5|    0|132.26| 165.6|
|       1| Sunrisers Hyderabad| Yuvraj Singh| 207|        125|      4|         65|
30|    7|    3|216.67| 165.6|
|       1|Royal Challengers...|  A Choudhary| 172|        123|     10|          6|
2|    0|    1| 300.0|139.84|
|       1|Royal Challengers...|     CH Gayle| 172|        123|     10|         34|
23|    2|    3|147.83|139.84|
|       1|Royal Challengers...|    KM Jadhav| 172|        123|     10|         34|
18|    4|    1|188.89|139.84|
|       1|Royal Challengers...|Mandeep Singh| 172|        123|     10|         24|
16|    5|    0| 150.0|139.84|
|       1|Royal Challengers...|    S Aravind| 172|        123|     10|          0|
2|    0|    0|   0.0|139.84|
|       1|Royal Challengers...|    SR Watson| 172|        123|     10|         22|
17|    1|    1|129.41|139.84|
|       1|Royal Challengers...|    STR Binny| 172|        123|     10|         11|
10|    0|    1| 110.0|139.84|
10|    0|    1| 110.0|139.84|
|       1|Royal Challengers...|  Sachin Baby| 172|        123|     10|           1|
3|    0|    0| 33.33|139.84|
|       1|Royal Challengers...|      TM Head| 172|        123|     10|          30|
22|    3|    0|136.36|139.84|
|       1|Royal Challengers...|     TS Mills| 172|        123|     10|           7|
3|    0|    1|233.33|139.84|
|       1|Royal Challengers...|    YS Chahal| 172|        123|     10|           3|
7|    0|    0| 42.86|139.84|
+--------+--------------------+-------------+----+-----------+-------+-----------+-----
-------+-----+-----+------+-------+
```

In [23]:

```
bowling_scorecard_df = spark.sql("""with bs as (
select distinct match_id, bowling_team, bowler,
concat(cast(floor(count(1) over(partition by match_id, bowling_team, bowler)/6) as string
),'.',cast(mod(count(1) over(partition by match_id, bowling_team, bowler),6) as string))
overs,
sum(total_runs) over(partition by match_id, bowling_team, bowler) runs,
case when sum(total_runs) over(partition by match_id, bowling_team, bowler, `over`) >= 1
then 0 else 1 end maiden,
count(player_dismissed) over(partition by match_id, bowling_team, bowler) wickets,
sum(case when total_runs=4 then 1 else 0 end) over(partition by match_id, bowling_team, b
owler) fours,
sum(case when total_runs=6 then 1 else 0 end) over(partition by match_id, bowling_team, b
owler) sixes,
sum(case when total_runs=0 then 1 else 0 end) over(partition by match_id, bowling_team, b
owler) zeros,
sum(case when wide_runs>0 then 1 else 0 end) over(partition by match_id, bowling_team, bo
wler) wides,
sum(case when noball_runs>0 then 1 else 0 end) over(partition by match_id, bowling_team,
bowler) noballs
from deliveries where match_id=1)
select distinct *, round(runs/overs,2) economy from bs
""")

bowling_scorecard_df.filter("match_id=1").show()
```

```
+--------+--------------------+------------+-----+----+------+-------+-----+-----+-----
+-----+-------+-------+
|match_id|        bowling_team|      bowler|overs|runs|maiden|wickets|fours|sixes|zeros|w
ides|noballs|economy|
+--------+--------------------+------------+-----+----+------+-------+-----+-----+-----
+-----+-------+-------+
|       1|Royal Challengers...| A Choudhary|  4.4|  55|     0|      1|    5|    3|    5|
3|      1|   12.5|
|       1|Royal Challengers...|    TS Mills|  4.1|  32|     0|      1|    3|    1|   10|
1|      0|    7.8|
|       1|Royal Challengers...|   YS Chahal|  4.0|  22|     0|      1|    0|    0|    5|
0|      0|    5.5|
|       1|Royal Challengers...|   STR Binny|  1.0|  10|     0|      1|    1|    0|    1|
0|      0|   10.0|
|       1| Sunrisers Hyderabad|     A Nehra|  4.1|  42|     0|      2|    6|    1|    8|
1|      0|  10.24|
|       1| Sunrisers Hyderabad|MC Henriques|  2.1|  20|     0|      1|    2|    0|    1|
0|      1|   9.52|
|       1| Sunrisers Hyderabad| Rashid Khan|  4.0|  36|     0|      2|    4|    1|    8|
0|      0|    9.0|
|       1|Royal Challengers...|   SR Watson|  3.0|  41|     0|      0|    5|    2|    3|
0|      0|  13.67|
|       1|Royal Challengers...|     TM Head|  1.0|  11|     0|      0|    0|    1|    0|
0|      0|   11.0|
|       1| Sunrisers Hyderabad|     B Kumar|  4.1|  28|     0|      2|    1|    2|   11|
1|      0|   6.83|
|       1| Sunrisers Hyderabad| BCJ Cutting|  4.0|  35|     0|      1|    2|    3|   10|
2|      0|   8.75|
|       1| Sunrisers Hyderabad|Bipul Sharma|  1.0|   4|     0|      1|    0|    0|    2|
0|      0|    4.0|
|       1|Royal Challengers...|   S Aravind|  3.0|  36|     0|      0|    3|    2|    2|
0|      0|   12.0|
|       1| Sunrisers Hyderabad|    DJ Hooda|  1.0|   7|     0|      1|    0|    1|    4|
0|      0|    7.0|
0|       0|     7.0|
+---------+---------------------+------------+-----+----+------+-------+-----+-----+-----
+-----+-------+--------+
```

In [50]:

```
points_table_df = spark.sql("select distinct season, winner as team, count(1) as points f
rom matches group by season, winner")
points_table_df.show()
```

```
+--------+--------------------+------+
|  season|                team|points|
+--------+--------------------+------+
|IPL-2011| Chennai Super Kings|    11|
|IPL-2011|    Rajasthan Royals|     6|
|IPL-2016| Sunrisers Hyderabad|    11|
|IPL-2008|     Kings XI Punjab|    10|
|IPL-2011|Kochi Tuskers Kerala|     6|
|IPL-2015|Royal Challengers...|     8|
|IPL-2013|     Kings XI Punjab|     8|
|IPL-2014|     Kings XI Punjab|    12|
|IPL-2008|      Mumbai Indians|     7|
|IPL-2010|Royal Challengers...|     8|
|IPL-2008|    Rajasthan Royals|    13|
|IPL-2011|Kolkata Knight Ri...|     8|
|IPL-2013| Sunrisers Hyderabad|    10|
|IPL-2010|      Mumbai Indians|    11|
|IPL-2017|Rising Pune Super...|    10|
|IPL-2010|    Delhi Daredevils|     7|
|IPL-2013|      Mumbai Indians|    13|
|IPL-2019| Chennai Super Kings|    10|
|IPL-2011|Royal Challengers...|    10|
|IPL-2018|    Delhi Daredevils|     5|
+--------+--------------------+------+
only showing top 20 rows
```

Match details

In [41]:

```
match_details_df = spark.sql("""select venue, team1, team2, toss_decision, season, player
_of_match, date,
                        concat_ws(', ', umpire1, umpire2, umpire3) as umpires,
                        winner
                        from matches
                   """)
match_details_df.filter("season='IPL-2017' and date='15-04-2017'").show()
```

```
+-----------------+--------------------+--------------------+-------------+--------+-----
----------+----------+--------------------+--------------------+
|            venue|               team1|               team2|toss_decision|  season|player
_of_match|      date|             umpires|              winner|
+-----------------+--------------------+--------------------+-------------+--------+-----
----------+----------+--------------------+--------------------+
|     Eden Gardens|Kolkata Knight Ri...|Sunrisers Hyderabad|        field|IPL-2017|     RV
Uthappa|15-04-2017|AY Dandekar, NJ L...|Kolkata Knight Ri...|
|Feroz Shah Kotla|    Delhi Daredevils|    Kings XI Punjab|          bat|IPL-2017|     CJ
Anderson|15-04-2017|YC Barde, Nitin M...|    Delhi Daredevils|
+-----------------+--------------------+--------------------+-------------+--------+-----
----------+----------+--------------------+--------------------+
```

Writing result dataframes to lake

Loading all the tables

In [49]:

```
# Building map for crawler and loader functions
```

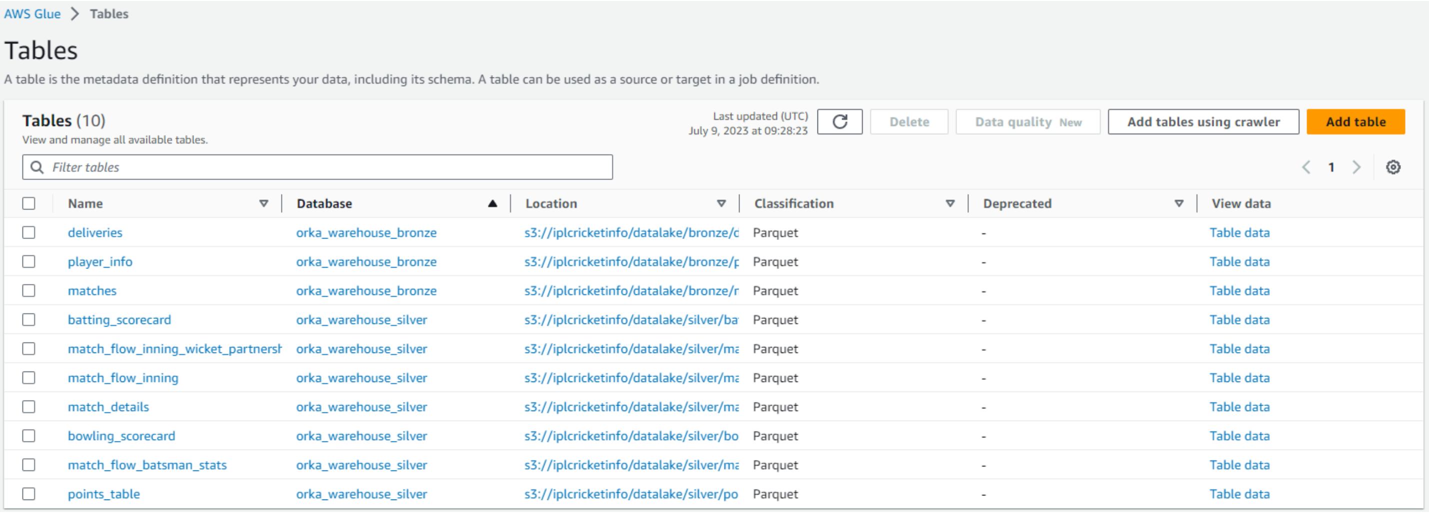
```
match_flow_silver_tables_to_load = {"match_flow_inning":match_flow_df_inning,
                                    "match_flow_batsman_stats":match_flow_df_batsman_sta
ts,
                                    "match_flow_inning_wicket_partnership":match_flow_df
_inning_wicket,
                                    "batting_scorecard":battin_scorecard_df,
                                    "bowling_scorecard":bowling_scorecard_df,
                                    "points_table":points_table_df,
                                    "match_details":match_details_df
                                   }

load_tables(match_flow_silver_tables_to_load,hub_loc)
crawl_tables(glue_client, match_flow_silver_tables_to_load, hub_loc, hub_database_name)
```

*

Now that hub is loaded, lets look at results in ATHENA

*

Looking at loaded databases and tables in glue catalog



AWS Glue > Tables

Tables

A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

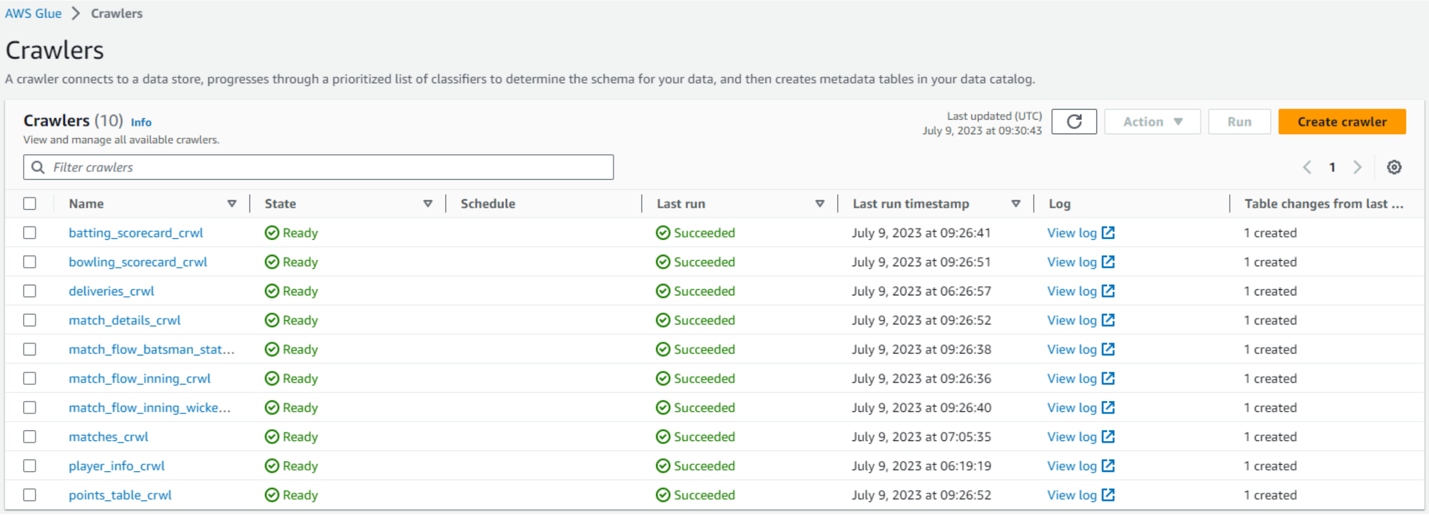
Tables (10)

View and manage all available tables.

Last updated (UTC)
July 9, 2023 at 09:28:23

□	Name	Database	Location	Classification	Deprecated	View data
□	deliveries	orka_warehouse_bronze	s3://iplcricketinfo/datalake/bronze/c	Parquet	-	Table data
□	player_info	orka_warehouse_bronze	s3://iplcricketinfo/datalake/bronze/p	Parquet	-	Table data
□	matches	orka_warehouse_bronze	s3://iplcricketinfo/datalake/bronze/r	Parquet	-	Table data
□	batting_scorecard	orka_warehouse_silver	s3://iplcricketinfo/datalake/silver/ba	Parquet	-	Table data
□	match_flow_inning_wicket_partnerst	orka_warehouse_silver	s3://iplcricketinfo/datalake/silver/ma	Parquet	-	Table data
□	match_flow_inning	orka_warehouse_silver	s3://iplcricketinfo/datalake/silver/ma	Parquet	-	Table data
□	match_details	orka_warehouse_silver	s3://iplcricketinfo/datalake/silver/ma	Parquet	-	Table data
□	bowling_scorecard	orka_warehouse_silver	s3://iplcricketinfo/datalake/silver/bo	Parquet	-	Table data
□	match_flow_batsman_stats	orka_warehouse_silver	s3://iplcricketinfo/datalake/silver/ma	Parquet	-	Table data
□	points_table	orka_warehouse_silver	s3://iplcricketinfo/datalake/silver/po	Parquet	-	Table data

Looking at crawlers created



AWS Glue > Crawlers

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

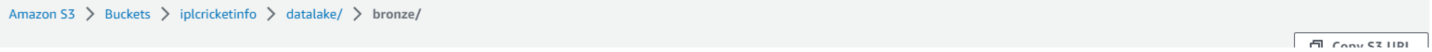
Crawlers (10) Info

View and manage all available crawlers.

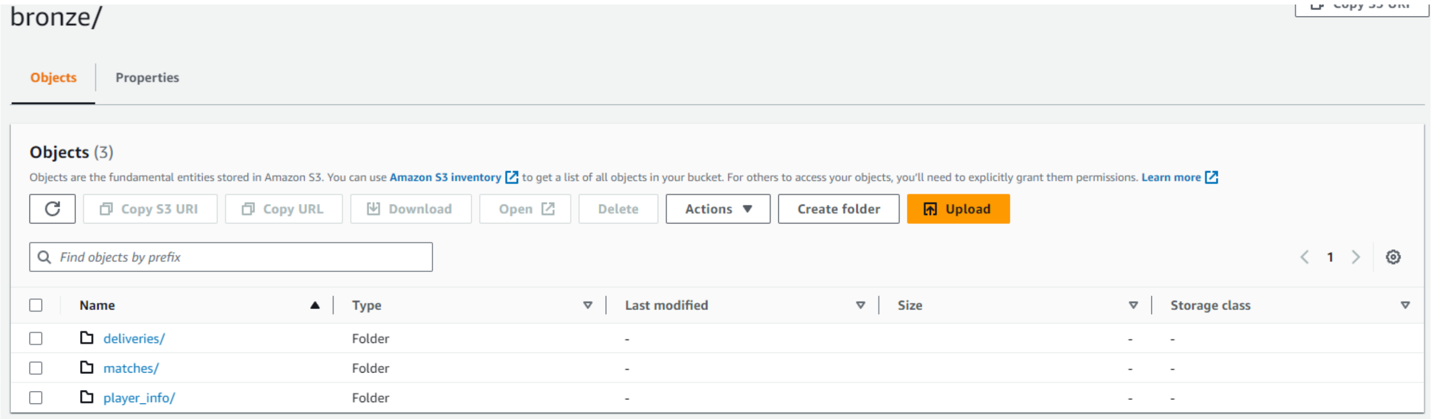
Last updated (UTC)
July 9, 2023 at 09:30:43

□	Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from last ...
□	batting_scorecard_crwl	✓ Ready		✓ Succeeded	July 9, 2023 at 09:26:41	View log	1 created
□	bowling_scorecard_crwl	✓ Ready		✓ Succeeded	July 9, 2023 at 09:26:51	View log	1 created
□	deliveries_crwl	✓ Ready		✓ Succeeded	July 9, 2023 at 06:26:57	View log	1 created
□	match_details_crwl	✓ Ready		✓ Succeeded	July 9, 2023 at 09:26:52	View log	1 created
□	match_flow_batsman_stat...	✓ Ready		✓ Succeeded	July 9, 2023 at 09:26:38	View log	1 created
□	match_flow_inning_crwl	✓ Ready		✓ Succeeded	July 9, 2023 at 09:26:36	View log	1 created
□	match_flow_inning_wicke...	✓ Ready		✓ Succeeded	July 9, 2023 at 09:26:40	View log	1 created
□	matches_crwl	✓ Ready		✓ Succeeded	July 9, 2023 at 07:05:35	View log	1 created
□	player_info_crwl	✓ Ready		✓ Succeeded	July 9, 2023 at 06:19:19	View log	1 created
□	points_table_crwl	✓ Ready		✓ Succeeded	July 9, 2023 at 09:26:52	View log	1 created

Looking at external s3 location of lake



Amazon S3 > Buckets > iplcricketinfo > datalake/ > bronze/



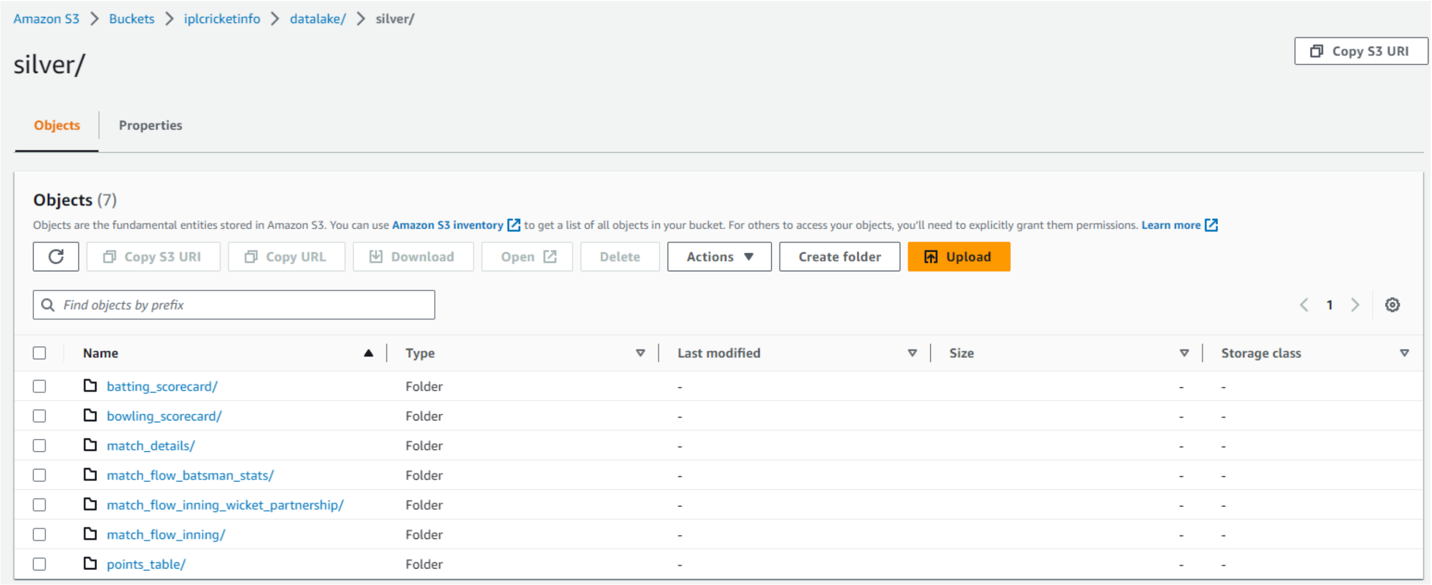
bronze/

Objects (3)

Objects are the fundamental entities stored in Amazon S3. You can use **Amazon S3 inventory** to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. **Learn more**

□	Name ▲	Type	Last modified	Size	Storage class
□	deliveries/	Folder	-	-	-
□	matches/	Folder	-	-	-
□	player_info/	Folder	-	-	-

Looking at external s3 location of lake



Amazon S3 > Buckets > iplcricketinfo > datalake/ > silver/

silver/

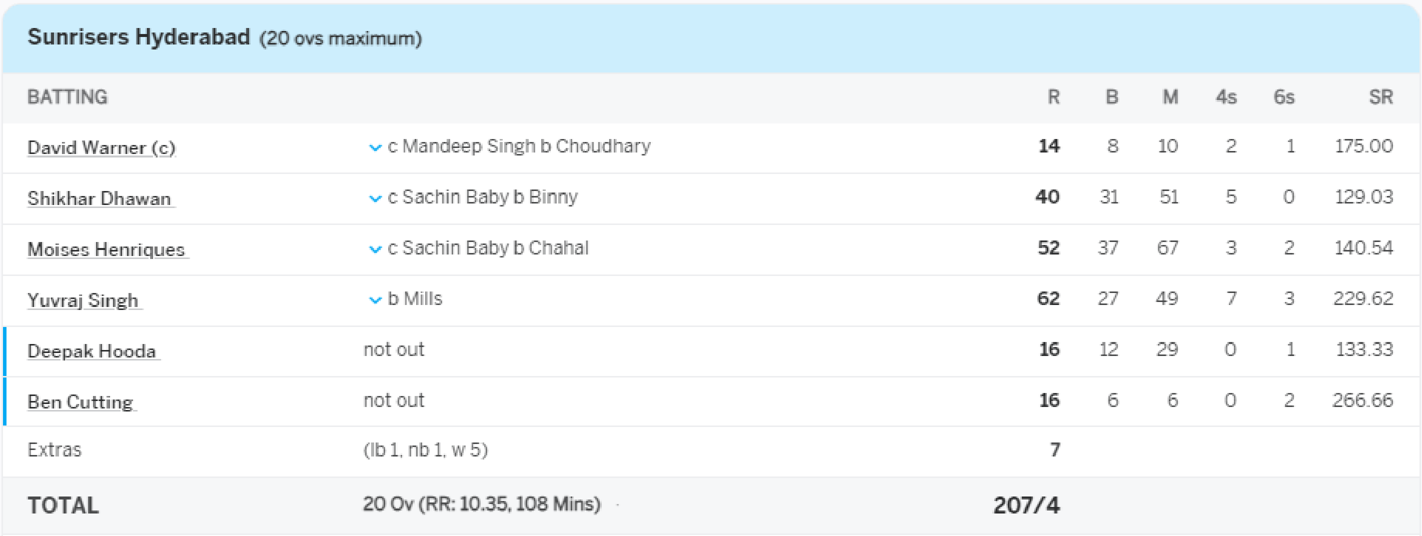
Objects (7)

Objects are the fundamental entities stored in Amazon S3. You can use **Amazon S3 inventory** to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. **Learn more**

□	Name ▲	Type	Last modified	Size	Storage class
□	batting_scorecard/	Folder	-	-	-
□	bowling_scorecard/	Folder	-	-	-
□	match_details/	Folder	-	-	-
□	match_flow_batsman_stats/	Folder	-	-	-
□	match_flow_inning_wicket_partnership/	Folder	-	-	-
□	match_flow_inning/	Folder	-	-	-
□	points_table/	Folder	-	-	-

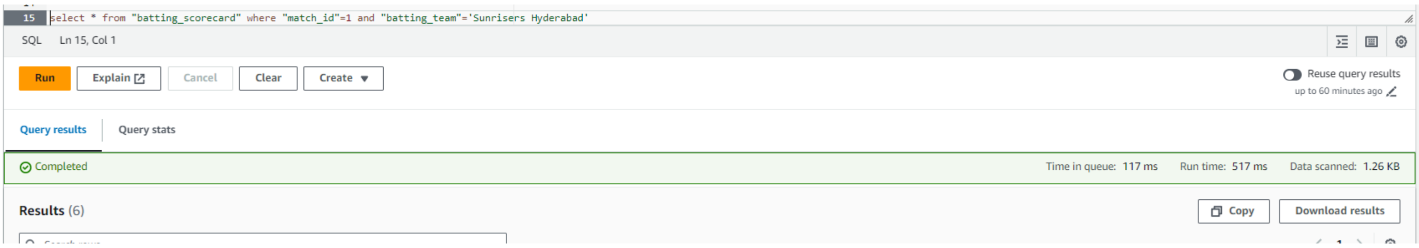
Looking at athena result for a match and comparing it with results from espn page

Batting score card



Sunrisers Hyderabad (20 ovs maximum)

BATTING		R	B	M	4s	6s	SR
David Warner (c)	c Mandeep Singh b Choudhary	**14**	8	10	2	1	175.00
Shikhar Dhawan	c Sachin Baby b Binny	**40**	31	51	5	0	129.03
Moises Henriques	c Sachin Baby b Chahal	**52**	37	67	3	2	140.54
Yuvraj Singh	b Mills	**62**	27	49	7	3	229.62
Deepak Hooda	not out	**16**	12	29	0	1	133.33
Ben Cutting	not out	**16**	6	6	0	2	266.66
Extras	(lb 1, nb 1, w 5)	**7**					
TOTAL	20 Ov (RR: 10.35, 108 Mins)	**207/4**					

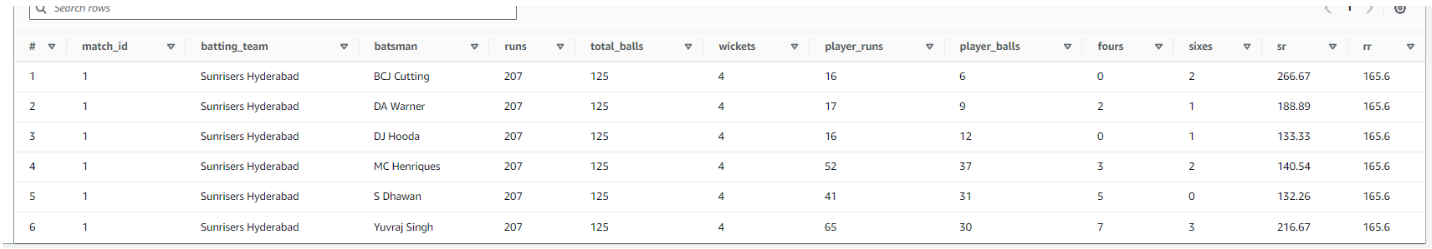


```
select * from "batting_scorecard" where "match_id"=1 and "batting_team"='Sunrisers Hyderabad'
```

SQL Ln 15, Col 1

Completed Time in queue: 117 ms Run time: 517 ms Data scanned: 1.26 KB

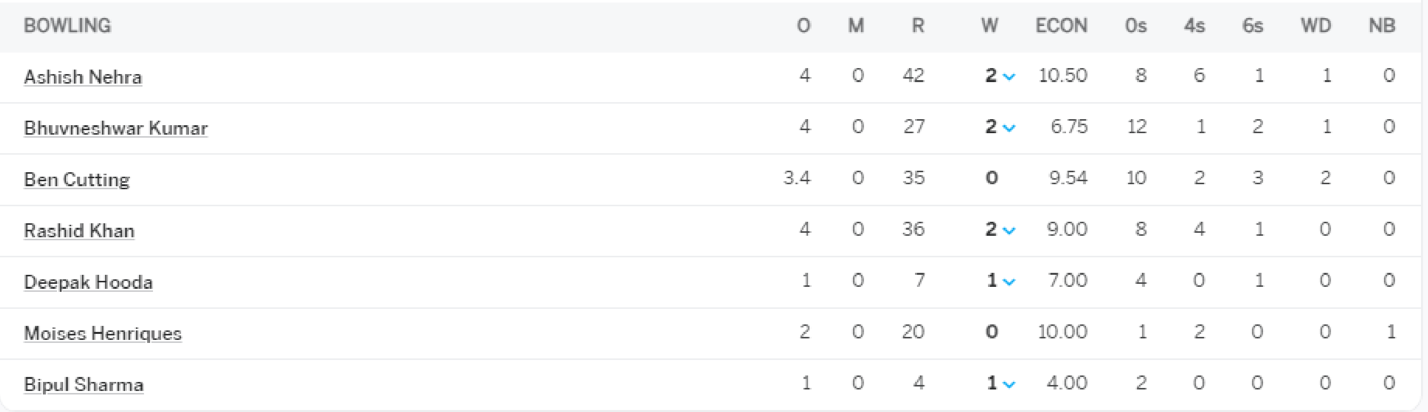
Results (6)



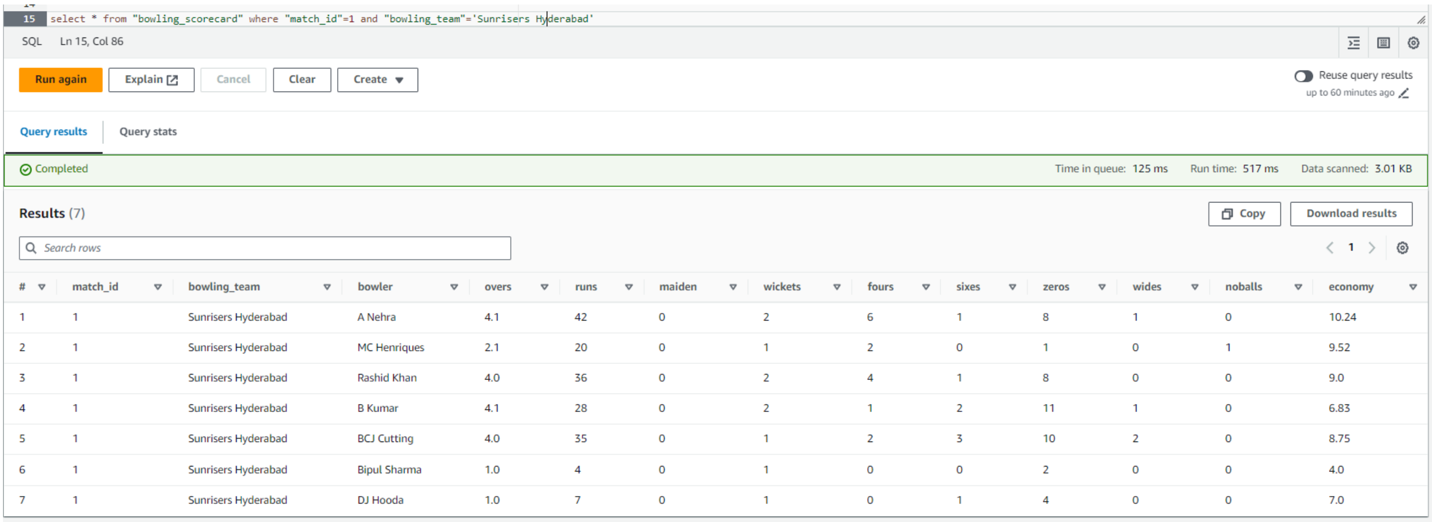
#	match_id	batting_team	batsman	runs	total_balls	wickets	player_runs	player_balls	fours	sixes	sr	rr
1	1	Sunrisers Hyderabad	BCJ Cutting	207	125	4	16	6	0	2	266.67	165.6
2	1	Sunrisers Hyderabad	DA Warner	207	125	4	17	9	2	1	188.89	165.6
3	1	Sunrisers Hyderabad	DJ Hooda	207	125	4	16	12	0	1	133.33	165.6
4	1	Sunrisers Hyderabad	MC Henriques	207	125	4	52	37	3	2	140.54	165.6
5	1	Sunrisers Hyderabad	S Dhawan	207	125	4	41	31	5	0	132.26	165.6
6	1	Sunrisers Hyderabad	Yuvraj Singh	207	125	4	65	30	7	3	216.67	165.6

There's minor discrepencies in total_balls, player_runs due to extras(No's, Wide's, Bye's) being counted which I didn't correct.

Bowling score card



BOWLING	O	M	R	W	ECON	0s	4s	6s	WD	NB
Ashish Nehra	4	0	42	2	10.50	8	6	1	1	0
Bhuvneshwar Kumar	4	0	27	2	6.75	12	1	2	1	0
Ben Cutting	3.4	0	35	0	9.54	10	2	3	2	0
Rashid Khan	4	0	36	2	9.00	8	4	1	0	0
Deepak Hooda	1	0	7	1	7.00	4	0	1	0	0
Moises Henriques	2	0	20	0	10.00	1	2	0	0	1
Bipul Sharma	1	0	4	1	4.00	2	0	0	0	0



```
select * from "bowling_scorecard" where "match_id"=1 and "bowling_team"='Sunrisers Hyderabad'
```

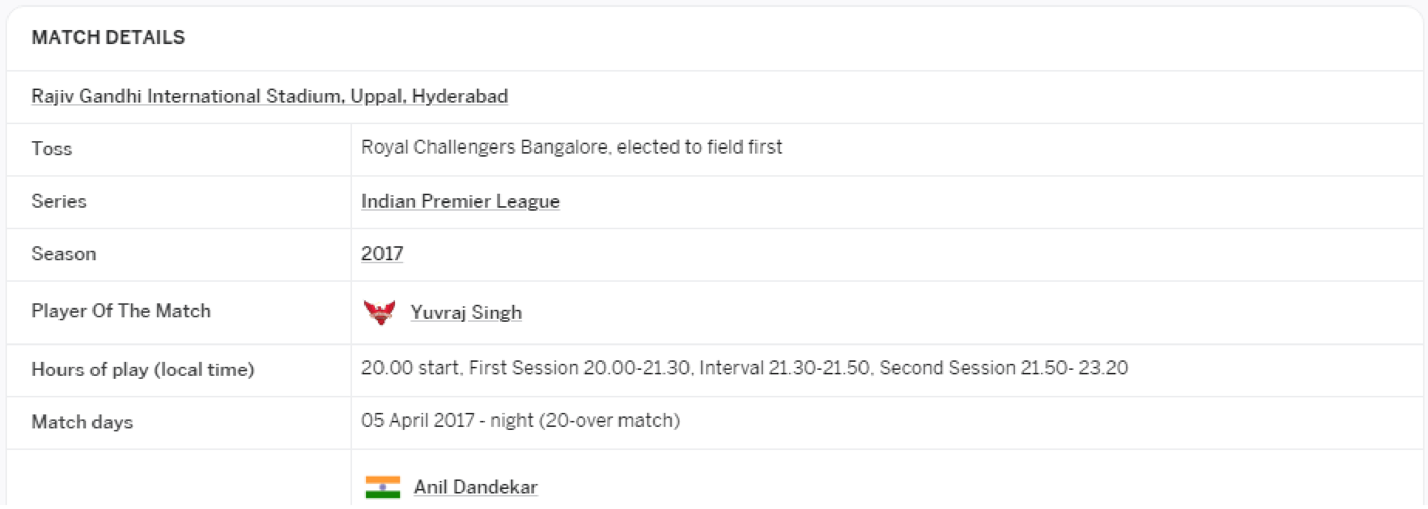
Completed   Time in queue: 125 ms   Run time: 517 ms   Data scanned: 3.01 KB

Results (7)

#	match_id	bowling_team	bowler	overs	runs	maiden	wickets	fours	sixes	zeros	wides	noballs	economy
1	1	Sunrisers Hyderabad	A Nehra	4.1	42	0	2	6	1	8	1	0	10.24
2	1	Sunrisers Hyderabad	MC Henriques	2.1	20	0	1	2	0	1	0	1	9.52
3	1	Sunrisers Hyderabad	Rashid Khan	4.0	36	0	2	4	1	8	0	0	9.0
4	1	Sunrisers Hyderabad	B Kumar	4.1	28	0	2	1	2	11	1	0	6.83
5	1	Sunrisers Hyderabad	BCJ Cutting	4.0	35	0	1	2	3	10	2	0	8.75
6	1	Sunrisers Hyderabad	Bipul Sharma	1.0	4	0	1	0	0	2	0	0	4.0
7	1	Sunrisers Hyderabad	DJ Hooda	1.0	7	0	1	0	1	4	0	0	7.0

Here too there is minor discrepency in overs, due to extras(No's, Wide's, Bye's) being counted which I didn't correct.

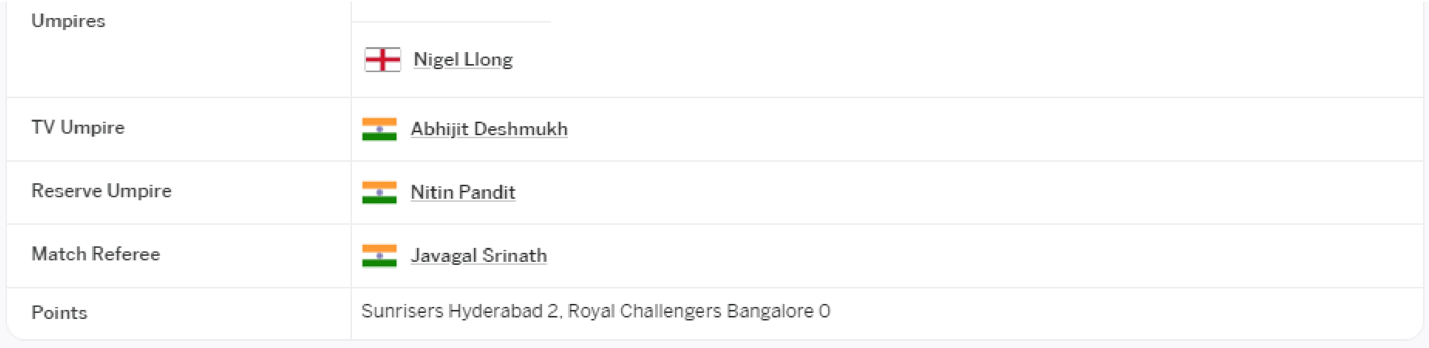
Match details



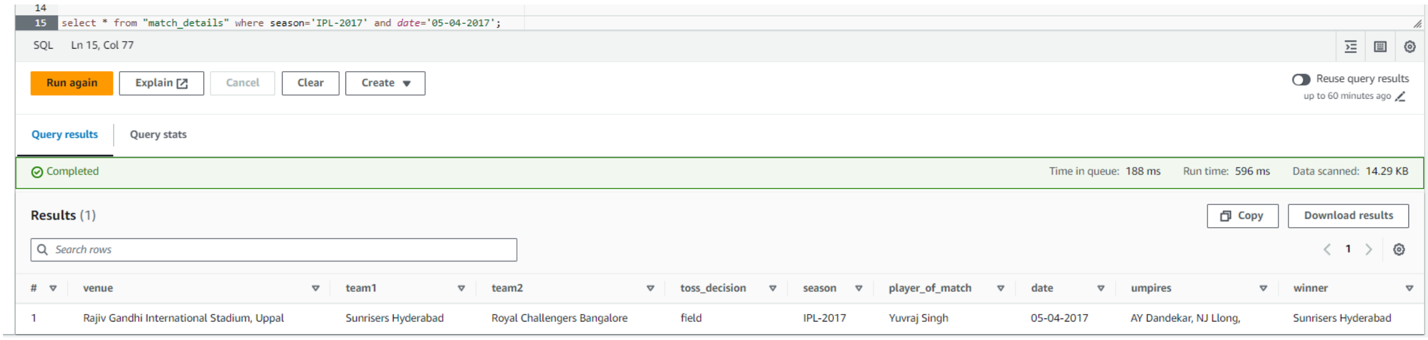
MATCH DETAILS

Rajiv Gandhi International Stadium, Uppal, Hyderabad

Toss	Royal Challengers Bangalore, elected to field first
Series	Indian Premier League
Season	2017
Player Of The Match	Yuvraj Singh
Hours of play (local time)	20.00 start, First Session 20.00-21.30, Interval 21.30-21.50, Second Session 21.50- 23.20
Match days	05 April 2017 - night (20-over match)
	Anil Dandekar



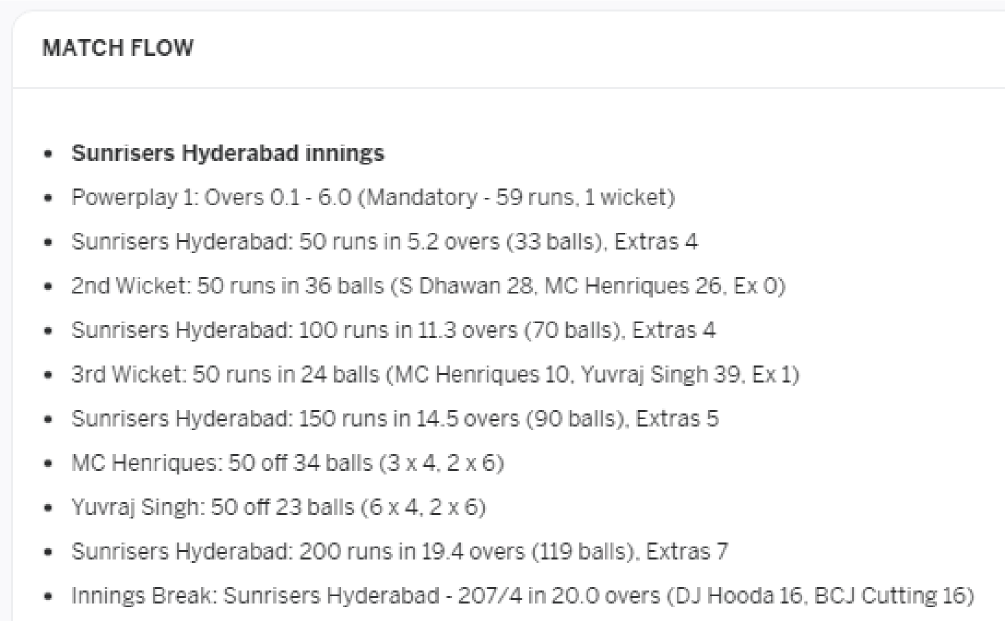
Umpires	Nigel Llong
TV Umpire	Abhijit Deshmukh
Reserve Umpire	Nitin Pandit
Match Referee	Javagal Srinath
Points	Sunrisers Hyderabad 2, Royal Challengers Bangalore 0



```sql
select * from "match_details" where season='IPL-2017' and date='05-04-2017';
```

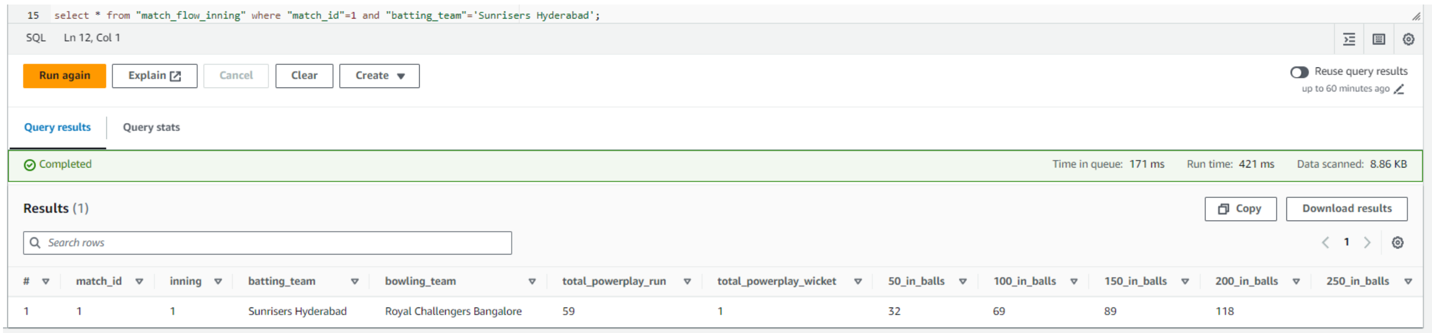
#	venue	team1	team2	toss_decision	season	player_of_match	date	umpires	winner
1	Rajiv Gandhi International Stadium, Uppal	Sunrisers Hyderabad	Royal Challengers Bangalore	field	IPL-2017	Yuvraj Singh	05-04-2017	AY Dandekar, NJ Llong,	Sunrisers Hyderabad

Match flow stats



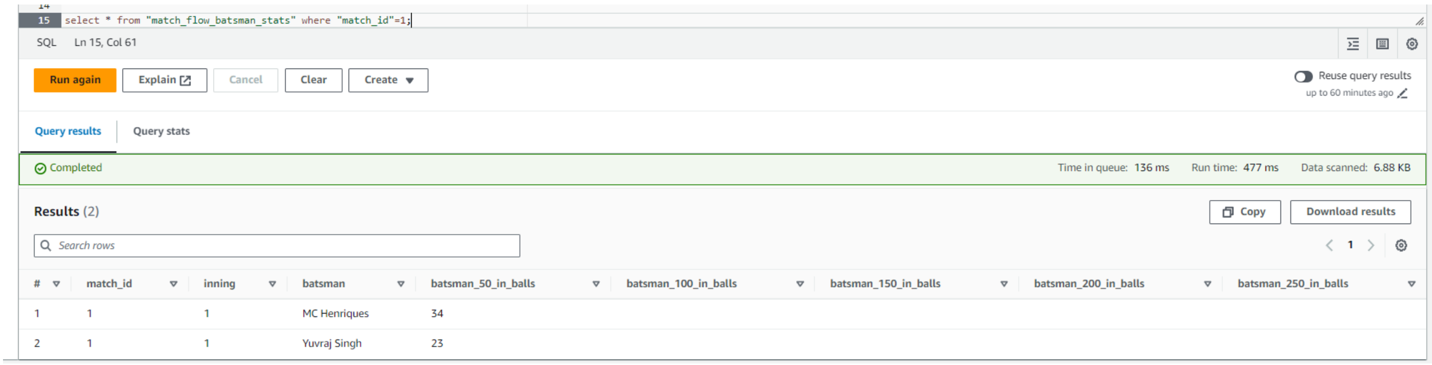
MATCH FLOW

- **Sunrisers Hyderabad innings**
- Powerplay 1: Overs 0.1 - 6.0 (Mandatory - 59 runs, 1 wicket)
- Sunrisers Hyderabad: 50 runs in 5.2 overs (33 balls), Extras 4
- 2nd Wicket: 50 runs in 36 balls (S Dhawan 28, MC Henriques 26, Ex 0)
- Sunrisers Hyderabad: 100 runs in 11.3 overs (70 balls), Extras 4
- 3rd Wicket: 50 runs in 24 balls (MC Henriques 10, Yuvraj Singh 39, Ex 1)
- Sunrisers Hyderabad: 150 runs in 14.5 overs (90 balls), Extras 5
- MC Henriques: 50 off 34 balls (3 x 4, 2 x 6)
- Yuvraj Singh: 50 off 23 balls (6 x 4, 2 x 6)
- Sunrisers Hyderabad: 200 runs in 19.4 overs (119 balls), Extras 7
- Innings Break: Sunrisers Hyderabad - 207/4 in 20.0 overs (DJ Hooda 16, BCJ Cutting 16)



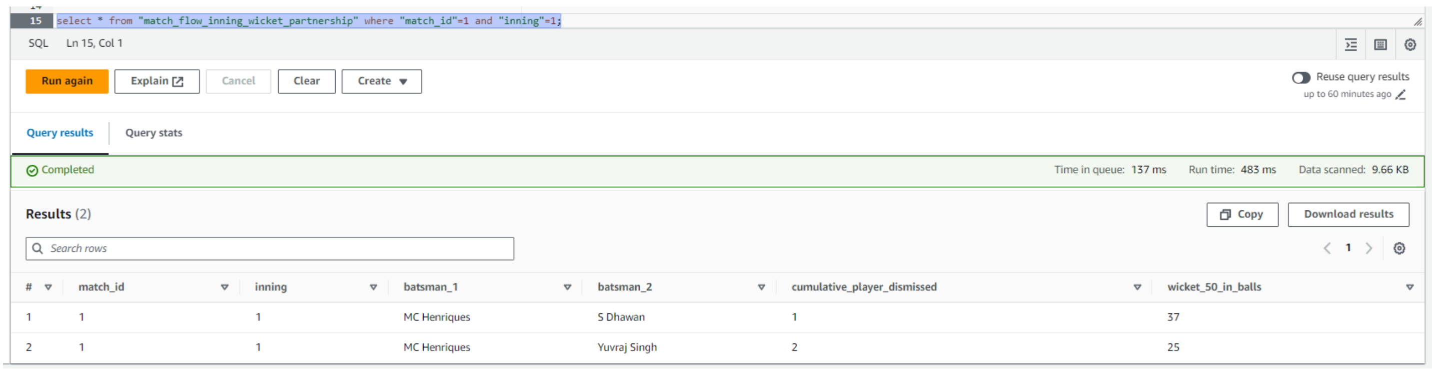
```sql
select * from "match_flow_inning" where "match_id"=1 and "batting_team"='Sunrisers Hyderabad';
```

#	match_id	inning	batting_team	bowling_team	total_powerplay_run	total_powerplay_wicket	50_in_balls	100_in_balls	150_in_balls	200_in_balls	250_in_balls
1	1	1	Sunrisers Hyderabad	Royal Challengers Bangalore	59	1	32	69	89	118	



```sql
select * from "match_flow_batsman_stats" where "match_id"=1;
```

#	match_id	inning	batsman	batsman_50_in_balls	batsman_100_in_balls	batsman_150_in_balls	batsman_200_in_balls	batsman_250_in_balls
1	1	1	MC Henriques	34				
2	1	1	Yuvraj Singh	23				



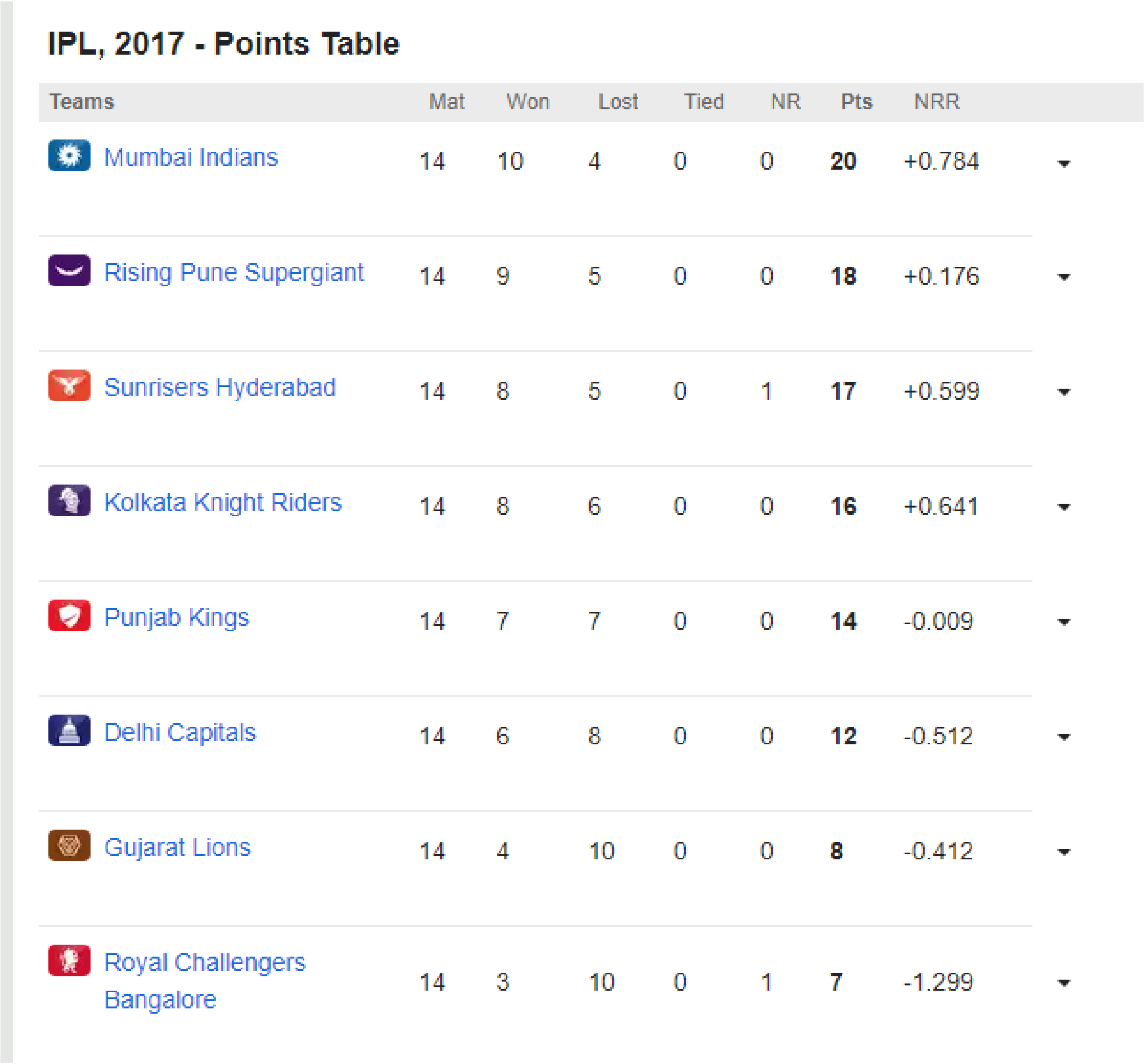
```
select * from "match_flow_inning_wicket_partnership" where "match_id"=1 and "inning"=1;
```

Completed Time in queue: 137 ms Run time: 483 ms Data scanned: 9.66 KB

Results (2)

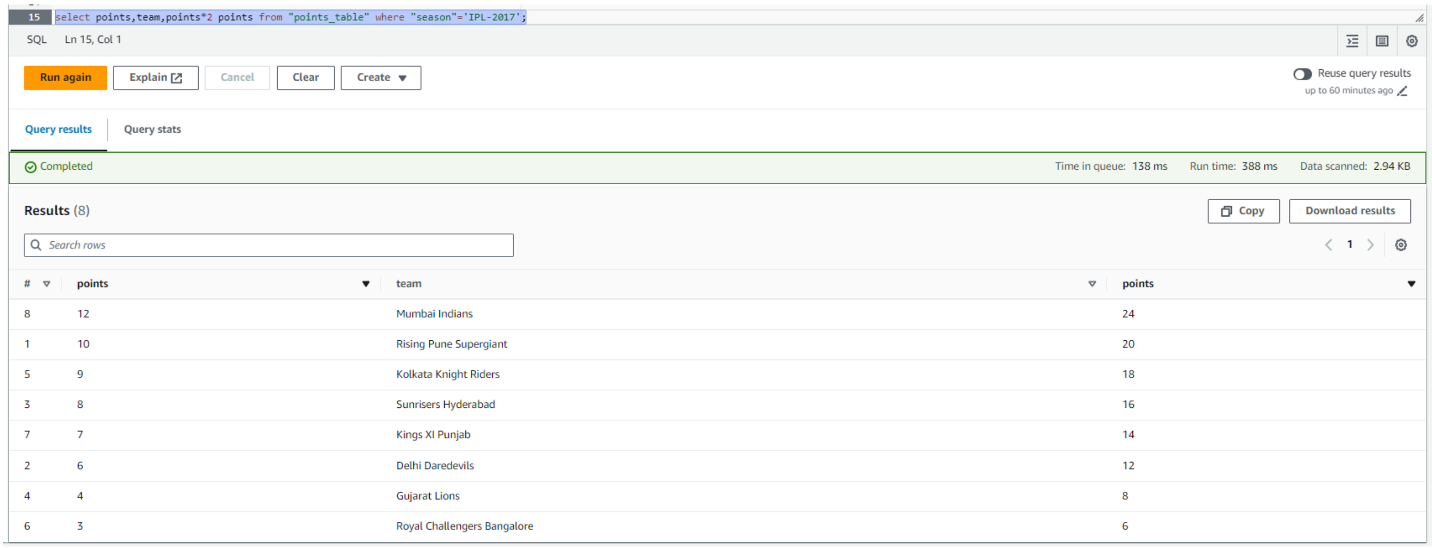
#	match_id	inning	batsman_1	batsman_2	cumulative_player_dismissed	wicket_50_in_balls
1	1	1	MC Henriques	S Dhawan	1	37
2	1	1	MC Henriques	Yuvraj Singh	2	25

Points table



IPL, 2017 - Points Table

Teams	Mat	Won	Lost	Tied	NR	Pts	NRR
Mumbai Indians	14	10	4	0	0	**20**	+0.784
Rising Pune Supergiant	14	9	5	0	0	**18**	+0.176
Sunrisers Hyderabad	14	8	5	0	1	**17**	+0.599
Kolkata Knight Riders	14	8	6	0	0	**16**	+0.641
Punjab Kings	14	7	7	0	0	**14**	-0.009
Delhi Capitals	14	6	8	0	0	**12**	-0.512
Gujarat Lions	14	4	10	0	0	**8**	-0.412
Royal Challengers Bangalore	14	3	10	0	1	**7**	-1.299



```
select points,team,points*2 points from "points_table" where "season"='IPL-2017';
```

Completed Time in queue: 138 ms Run time: 388 ms Data scanned: 2.94 KB

Results (8)

#	points	team	points
8	12	Mumbai Indians	24
1	10	Rising Pune Supergiant	20
5	9	Kolkata Knight Riders	18
3	8	Sunrisers Hyderabad	16
7	7	Kings XI Punjab	14
2	6	Delhi Daredevils	12
4	4	Gujarat Lions	8
6	3	Royal Challengers Bangalore	6

Here, there is dicrepancy because qualifier matches are getting counted toward points which I didn't correct.

Finishing up

In [52]:

```
job.commit()
```