Goal of this session is to create a cumulative table of bootcamp.nba\_player\_seasons

Thought process and steps followed

Understanding the initial table, schema, datatypes, size of the table

SELECT \* FROM bootcamp.nba\_player\_seasons

LIMIT 5

Output:

| **player\_name** | **age** | **height** | **weight** | **college** | **country** | **draft\_year** | **draft\_round** | **draft\_number** | **gp** | **pts** | **Reb** | **ast** | **netrtg** | **oreb\_pct** | **dreb\_pct** | **usg\_pct** | **ts\_pct** | **ast\_pct** | **season** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A.C. Green | 33 | 6-9 | 225 | Oregon State | USA | 1985 | 1 | 23 | 83 | 7.2 | 7.9 | 0.8 | -7.4 | 8.9 | 8.9 | 18.4 | 11.8 | 4.5 | 1996 |
| Aaron McKie | 24 | 6-5 | 209 | Temple | USA | 1994 | 1 | 17 | 83 | 5.2 | 2.7 | 1.9 | 3.7 | 2.6 | 2.6 | 11.3 | 14.2 | 16.3 | 1996 |
| Aaron Williams | 25 | 6-9 | 225 | Xavier | USA | Undrafted | Undrafted | Undrafted | 33 | 6.2 | 4.3 | 0.5 | -9.3 | 11.3 | 11.3 | 14.4 | 16.1 | 5.1 | 1996 |
| Acie Earl | 27 | 6-11 | 240 | Iowa | USA | 1993 | 1 | 19 | 47 | 4 | 2 | 0.4 | -6.4 | 6.7 | 6.7 | 12.2 | 22 | 7.7 | 1996 |
| Adam Keefe | 27 | 6-9 | 241 | Stanford | USA | 1992 | 1 | 10 | 62 | 3.8 | 3.5 | 0.5 | 7.2 | 9.6 | 9.6 | 15.8 | 12.4 | 5.1 | 1996 |

Observations:

Draft\_year, draft\_round, draft\_number are strings as there are values like ‘undefined’

There are 12869 records in nba\_player\_seasons table

Now, lets create the cumulative table, for my case, that is Prithvi.nba\_players. Idea is to have all the constant/unchanging columns coalesced and the changing/temporal columns as a part of an array.

As part of this lecture, only the player\_name, height, college,country,draft\_year,draft\_round,draft\_number were considered in the unchanging dimensions. For temporal dimensions, we tracked season,age,weight,gp,pts,reb,ast in the form of an seasons array.

We also have a column called ‘is\_active’ to track if the player is playing in the current season or not.

Another column called ‘years\_since\_last\_active’ and ‘current season’ to keep track of the current season.

Seasons array will be a array of ROW objects, with a definition:

ROW(season integer, age integer, weight integer, gp integer, pts double, reb double, ast double)

create schema prithvi

create table prithvi.nba\_players(

player\_name varchar,

height varchar,

college varchar,

country varchar,

draft\_year varchar,

draft\_round varchar,

draft\_number varchar,

seasons array(

ROW(

season integer,

age integer,

weight integer,

gp integer,

pts double,

reb double,

ast double

)

),

is\_active integer,

year\_since\_last\_active integer,

current\_season integer

)

with

(

format = 'PARQUET',

partitioning = ARRAY['current\_season'])

SHOW STATS FOR (

SELECT

\*

FROM

prithvi.nba\_players

)

Now the cumulative table schema is created, time to populate it with data. Let us check what is the first record time wise in the bootcamp.nba\_player\_seasons, aka first season.

select \*

from bootcamp.nba\_player\_seasons

order by

season asc

| **player\_name** | **age** | **height** | **weight** | **college** | **country** | **draft\_year** | **draft\_round** | **draft\_number** | **gp** | **pts** | **reb** | **ast** | **netrtg** | **oreb\_pct** | **dreb\_pct** | **usg\_pct** | **ts\_pct** | **ast\_pct** | **season** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Zan Tabak | 27 | 7-0 | 245 | None | Croatia | 1991 | 2 | 51 | 13 | 6.5 | 3.8 | 1.1 | -9.4 | 9.4 | 9.4 | 12.9 | 21.5 | 11.1 | 1996 |
| Anfernee Hardaway | 25 | 6-7 | 215 | Memphis | USA | 1993 | 1 | 3 | 59 | 20.5 | 4.5 | 5.6 | 4.5 | 3.7 | 3.7 | 8.4 | 25 | 27 | 1996 |
| Brett Szabo | 29 | 6-11 | 230 | Augustana (SD) | USA | Undrafted | Undrafted | Undrafted | 70 | 2.2 | 2.4 | 0.2 | -11.1 | 7.9 | 7.9 | 16.8 | 11.7 | 3.9 | 1996 |

As seen above, the first season is 1996.

Bootcamp.nba\_player\_seasons is our fact table for all the players. Let’s load the data incrementally. For the first iteration, the query could look something like this:

with last\_season as(

select \*

from prithvi.nba\_players

where current\_season = 1995

),

live\_season as(

select \*

from bootcamp.nba\_player\_seasons

where season =1996

)

select \*

from last\_season las

full outer join live\_season liv

on las.player\_name = liv.player\_name

| **player\_name** | **height** | **college** | **country** | **draft\_year** | **draft\_round** | **draft\_number** | **seasons** | **is\_active** | **year\_since\_last\_active** | **current\_season** | **player\_name** | **age** | **height** | **weight** | **college** | **country** | **draft\_year** | **draft\_round** | **draft\_number** | **gp** | **pts** | **reb** | **ast** | **netrtg** | **oreb\_pct** | **dreb\_pct** | **usg\_pct** | **ts\_pct** | **ast\_pct** | **season** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| null | null | null | null | null | null | null | null | null | null | null | A.C. Green | 33 | 6-9 | 225 | Oregon State | USA | 1985 | 1 | 23 | 83 | 7.2 | 7.9 | 0.8 | -7.4 | 8.9 | 8.9 | 18.4 | 11.8 | 4.5 | 1996 |
| null | null | null | null | null | null | null | null | null | null | null | Aaron McKie | 24 | 6-5 | 209 | Temple | USA | 1994 | 1 | 17 | 83 | 5.2 | 2.7 | 1.9 | 3.7 | 2.6 | 2.6 | 11.3 | 14.2 | 16.3 | 1996 |

We encounter nulls from the empty Prithvi\_nba\_players table (as it’s the first iteration its empty, from the next iteration, this table represents the past) and the records from the first season from our fact table(bootcamp.nba\_player\_seasons)

Also, all the unchanging dimensions are to be coalesced. The coalesce command handles null values and returns the first non\_null value from the list of items supplied to it.

We update the query as follows:

with last\_season as(

select \*

from prithvi.nba\_players

where current\_season = 1995

),

live\_season as(

select \*

from bootcamp.nba\_player\_seasons

where season =1996

)

select

coalesce(las.player\_name,liv.player\_name) as player\_name,

coalesce(las.height,liv.height) as height,

coalesce(las.college,liv.college) as college,

coalesce(las.country,liv.country) as country,

coalesce(las.draft\_year,liv.draft\_year) as draft\_year,

coalesce(las.draft\_round, liv.draft\_round) as draft\_round,

coalesce(las.draft\_number,liv.draft\_number) as draft\_number,

null as seasons,

case when liv.season is not null then 1 else 0 end as is\_active,

case when liv.season is not null then 0 else las.year\_since\_last\_active+1 end as year\_since\_last\_active,

coalesce(liv.season, las.current\_season+1) as current\_season

from last\_season las

full outer join live\_season liv

on las.player\_name = liv.player\_name

In the above query, we coalesced all the unchanging dimensions, kept seasons array as null for now(will work on the next iteration). Lets delve deeper into the last three columns.

* Is\_active: case when liv.season is not null then 1 else 0 end

If the season column from the live season is null, it means player is not playing in the current season and is\_active column is assigned 0. If the season column has a value, it means he is playing, and its set to 1

* Years\_since\_last\_active: case when liv.season is not null then 0 else las.year\_since\_last\_active+1end

If the season in live season is not null, it means the player is playing right now and it will be set to 0. If its null, it means he is not playing, we take the entry from his last season i.e las.year\_since\_last\_active and add a 1 to it.

* Current\_season: coalesce(liv.season, las.current\_season+1)

We know that current season never be null.

We also know coalesce returns the first non-null value from the list. Our list here has the current season: liv.season and las.current\_season+1.

Now 2 cases arise here: when liv.season is not null, that means a player is playing right now, it will be set to liv.season.

When liv.season is null, which means the player is not playing right now. This happens when the player plays last season and does not in the present season. When we join the tables as in the query, the las.current\_season column will be non-null and the liv.season will be null(right table where the player entry will be missing and all nulls). In this case current season will be rightly set to las.current\_season+1.

| **player\_name** | **height** | **college** | **country** | **draft\_year** | **draft\_round** | **draft\_number** | **seasons** | **is\_active** | **year\_since\_last\_active** | **current\_season** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A.C. Green | 6-9 | Oregon State | USA | 1985 | 1 | 23 | null | 1 | 0 | 1996 |
| Aaron McKie | 6-5 | Temple | USA | 1994 | 1 | 17 | null | 1 | 0 | 1996 |
| Aaron Williams | 6-9 | Xavier | USA | Undrafted | Undrafted | Undrafted | null | 1 | 0 | 1996 |

Final query for cumulative query would be:

with last\_season as(

select \*

from prithvi.nba\_players

where current\_season = 1995

),

live\_season as(

select \*

from bootcamp.nba\_player\_seasons

where season =1996

)

select

coalesce(las.player\_name,liv.player\_name) as player\_name,

coalesce(las.height,liv.height) as height,

coalesce(las.college,liv.college) as college,

coalesce(las.country,liv.country) as country,

coalesce(las.draft\_year,liv.draft\_year) as draft\_year,

coalesce(las.draft\_round, liv.draft\_round) as draft\_round,

coalesce(las.draft\_number,liv.draft\_number) as draft\_number,

case

when liv.season is null then las.seasons

when liv.season is not null and las.seasons is null then ARRAY[Row(liv.season,liv.age,liv.weight,liv.gp,liv.pts,liv.reb,liv.ast)]

when liv.season is not null and las.seasons is not null then ARRAY[Row(liv.season,liv.age,liv.weight,liv.gp,liv.pts,liv.reb,liv.ast)]||las.seasons

end as seasons,

case when liv.season is not null then 1 else 0 end as is\_active,

case when liv.season is not null then 0 else las.year\_since\_last\_active+1 end as year\_since\_last\_active,

coalesce(liv.season, las.current\_season+1) as current\_season

from last\_season las

full outer join live\_season liv

on las.player\_name = liv.player\_name

Looking at the seasons column:

case

when liv.season is null then las.seasons

when liv.season is not null and las.seasons is null then ARRAY[Row(liv.season,liv.age,liv.weight,liv.gp,liv.pts,liv.reb,liv.ast)]

when liv.season is not null and las.seasons is not null then ARRAY[Row(liv.season,liv.age,liv.weight,liv.gp,liv.pts,liv.reb,liv.ast)]||las.seasons end as seasons

There are 3 cases here:

* When the current season is null, then we just take the seasons array from last season
* When the liv.season is not null, which means we have an entry in the current season and the previous season’s array (las.seasons)is null. This is the first time we come across this player, we create an array element using the Array[Row()] syntax
* When the liv.season is not null and las.seasons is not null. This is a player who played both the last season and the current season. Then we use || operator to concatenate the present season(Array[Row()]) nd las.seasons

The above query works with no issues. Now we insert this data into Prithvi.nba\_players cumulative table. We start with year 1995 ,1996 and end with 2001,2002 respectively for last season and live season respectively.

We need run the query sequentially, as it’s a cumulative table design

Insert into prithvi.nba\_players

with last\_season as(

select \*

from prithvi.nba\_players

where current\_season = 1995

),

live\_season as(

select \*

from bootcamp.nba\_player\_seasons

where season =1996

)

select

coalesce(las.player\_name,liv.player\_name) as player\_name,

coalesce(las.height,liv.height) as height,

coalesce(las.college,liv.college) as college,

coalesce(las.country,liv.country) as country,

coalesce(las.draft\_year,liv.draft\_year) as draft\_year,

coalesce(las.draft\_round, liv.draft\_round) as draft\_round,

coalesce(las.draft\_number,liv.draft\_number) as draft\_number,

case

when liv.season is null then las.seasons

when liv.season is not null and las.seasons is null then ARRAY[Row(liv.season,liv.age,liv.weight,liv.gp,liv.pts,liv.reb,liv.ast)]

when liv.season is not null and las.seasons is not null then ARRAY[Row(liv.season,liv.age,liv.weight,liv.gp,liv.pts,liv.reb,liv.ast)]||las.seasons

end as seasons,

case when liv.season is not null then 1 else 0 end as is\_active,

case when liv.season is not null then 0 else las.year\_since\_last\_active+1 end as year\_since\_last\_active,

coalesce(liv.season, las.current\_season+1) as current\_season

from last\_season las

full outer join live\_season liv

on las.player\_name = liv.player\_name

Inserts 441 rows

Now increment the years by 1

Insert into prithvi.nba\_players

with last\_season as(

select \*

from prithvi.nba\_players

where current\_season = 1996

),

live\_season as(

select \*

from bootcamp.nba\_player\_seasons

where season =1997

)

select

coalesce(las.player\_name,liv.player\_name) as player\_name,

coalesce(las.height,liv.height) as height,

coalesce(las.college,liv.college) as college,

coalesce(las.country,liv.country) as country,

coalesce(las.draft\_year,liv.draft\_year) as draft\_year,

coalesce(las.draft\_round, liv.draft\_round) as draft\_round,

coalesce(las.draft\_number,liv.draft\_number) as draft\_number,

case

when liv.season is null then las.seasons

when liv.season is not null and las.seasons is null then ARRAY[Row(liv.season,liv.age,liv.weight,liv.gp,liv.pts,liv.reb,liv.ast)]

when liv.season is not null and las.seasons is not null then ARRAY[Row(liv.season,liv.age,liv.weight,liv.gp,liv.pts,liv.reb,liv.ast)]||las.seasons

end as seasons,

case when liv.season is not null then 1 else 0 end as is\_active,

case when liv.season is not null then 0 else las.year\_since\_last\_active+1 end as year\_since\_last\_active,

coalesce(liv.season, las.current\_season+1) as current\_season

from last\_season las

full outer join live\_season liv

on las.player\_name = liv.player\_name

Inserts 527 rows

The process continues till the end, this dataset has till the year 2022. As part of this lecture, we do this only till year 2002, to get Michael Jordan’s complete record in, as he retired that year.

After running the above query repetitively till year 2002, we get 4642 rows.

Check if your columns are working as expected:

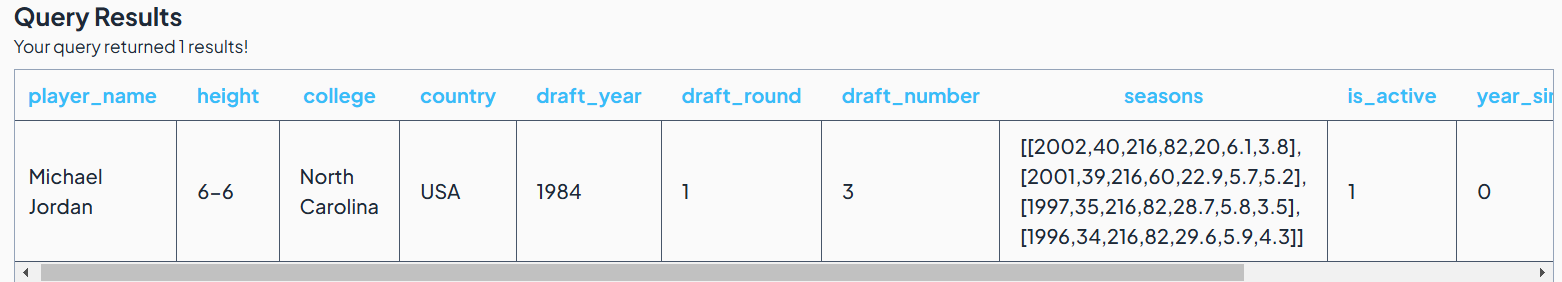
select \* from prithvi.nba\_players

where current\_season =2002

and player\_name = 'Michael Jordan'

**player\_nameheightcollegecountrydraft\_yeardraft\_rounddraft\_numberseasonsis\_activeyear\_since\_last\_activecurrent\_season**

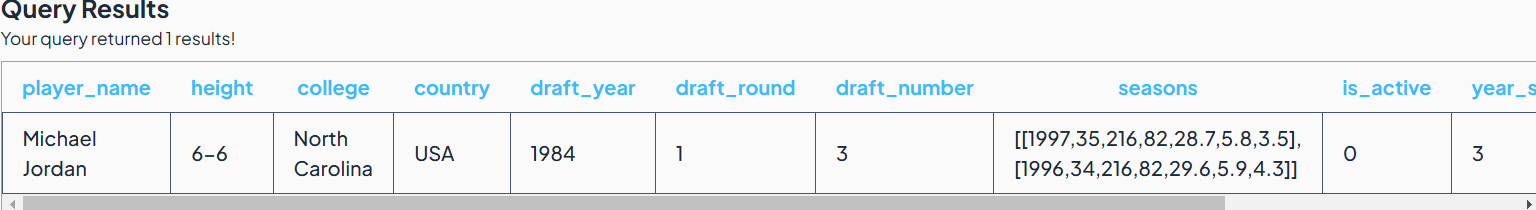
Michael Jordan6-6North CarolinaUSA198413[[2002,40,216,82,20,6.1,3.8],[2001,39,216,60,22.9,5.7,5.2],[1997,35,216,82,28.7,5.8,3.5],[1996,34,216,82,29.6,5.9,4.3]]102002



select \* from prithvi.nba\_players

where current\_season =2000

and player\_name = 'Michael Jordan'



**player\_nameheightcollegecountrydraft\_yeardraft\_rounddraft\_numberseasonsis\_activeyear\_since\_last\_activecurrent\_season**

Michael Jordan6-6North CarolinaUSA198413[[1997,35,216,82,28.7,5.8,3.5],[1996,34,216,82,29.6,5.9,4.3]]032000

Working as expected! Michael Jordan retired in the year 1997 and only returned to play in 2001

Thanks to cumulative table design, we have all the information pertaining to a single player in a single row. This was accomplished using complex datatype like ARRAY.

This array of rows is very cumbersome to read. We can use a cross join along with unnest to explode out the contents relating to a player using the following query:

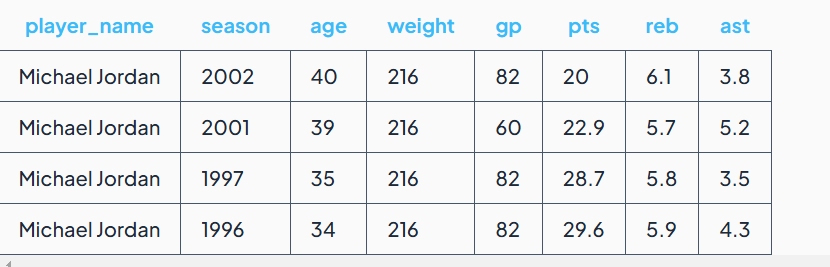
select player\_name, t.\*

from prithvi.nba\_players

cross join unnest(seasons) as t

where current\_season =2002

and player\_name = 'Michael Jordan'



Now, we have the data, what if we want to get the sum of the total points during his career?

This can be done as follows:

select player\_name, sum(t.gp\*t.pts) as total\_points

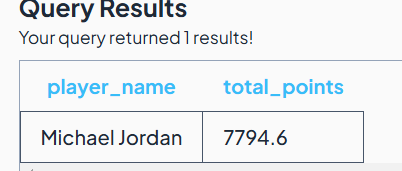
from prithvi.nba\_players

cross join unnest(seasons) as t

where current\_season =2002

and player\_name = 'Michael Jordan'

group by player\_name



Time taken:0.7 sec

But there is a problem here. We took all the pain to build a cumulative table to avoid shuffle(shuffling of data) which generally takes lot of time to execute and is compute intensive. Here, in the above query, we use group by, which triggers shuffle=ing of the data, hence not recommended.

There is another way to do this, using reduce method with works with array data structure.

select player\_name,

reduce(seasons,0,(s,r) -> s +r.pts\*r.gp, s->s) as total\_points

from prithvi.nba\_players

where current\_season =2002

and player\_name = 'Michael Jordan'

Time taken: 0.07 s

Sample data from the cumulative table for Michael Jordan:

select \* from prithvi.nba\_players where player\_name = 'Michael Jordan'

