**Analyze the data with Azure HDInsight**

**By**

**Israel Nolazco**

Before we get started, we need to start by setting up an environment for hive. The image below shows the hive environment created in Microsoft Azure.

A screenshot of a cell phone

Description automatically generated

1. What is the total number of baseball players?

Theory:

SELECT count(\*) FROM People

Actual:

A screenshot of a cell phone

Description automatically generated

1. How many players were born in the year 1960 and earlier?

Theory:

SELECT count(\*) FROM People

WHERE BirthYear < = 1960

Actual:

A screenshot of a cell phone

Description automatically generated

1. How many players were born in the USA?

Theory:

SELECT count(\*) FROM People

Where BithCountry = USA

Actual:

A screenshot of a cell phone

Description automatically generated

1. How many players were born outside the USA?

Theory:

SELECT count(\*) FROM People

Where BithCountry < > USA

Actual:

As seen in the image below the actual piece of code is different. The command WHERE NOT is the correct one for the results we are trying to obtain.

A screenshot of a cell phone

Description automatically generated

1. Display the number of players born in each year starting from 1960 thru 2000. For example, the output should show: 1980 4 ( where 4 is the number of players born in 1980)

Theory:

SELECT BirthYear, count (\*) from People

WHERE BirthYear >= 1960 and Birth Year < =2000

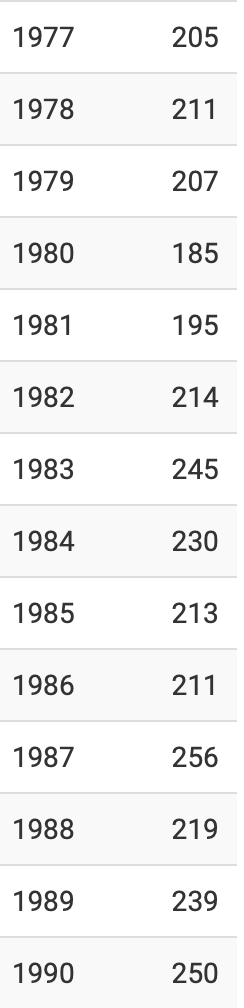
GROUP by BirthYear

Order by BirthYear

Actual:

A screenshot of a social media post

Description automatically generated



1. How many players and managers were inducted into the Hall of Fame?

Theory: For this question the halloffame.csv file was uploaded.

SELECT inducted, category count(\*) FROM HallOfFam

Where inducted = Y

Where category = Player AND Managers

Actual:

select count(inducted), category from halloffame

where inducted = 'Y'

group by category;

It looks like creating a where statement with both categories of manager and player creates an error. I would need to further research the appropriate command.

A screenshot of a social media post

Description automatically generated

1. Provide a list of all players for any team and from any year. For example, print the list of players who played for Chicago Cubs in 2000.

Theory.

SELECT Appearances.yearID, Appearances.teamID, Appearances.playerID, Teams.name, People.nameGiven

INNER JOIN Teams ON Appearances.TeamID = Teams.teamID

INNER JOIN People ON Appearances.playerID = People.playerID

The overall problem with this code is that is the results recreate the event of the player appear in that year. That is to say if the player played 5 times in a year. Then the name will show 5 times in the same year. This just expands the number of entries by each player appear each year. Therefore, an AND statement needs to be added along with a ORDER BY to make it visually pleasing .

Actual:

select appearances.yearID, appearances.teamID, appearances.playerID, teams.name, people.nameGiven

from teams join appearances

on (teams.teamID = appearances.teamID) and (teams.yearID = appearances.yearID)

join people

on (people.playerID = appearances.playerID)

order by appearances.yearID, teams.name;

A screenshot of a social media post

Description automatically generated