Final Project

Ewan B-W

CS 178

5/10/17

**Part 1:**

Amazon Web Services (AWS) and Google Cloud Platform (GCP) are both cloud computing services. Both AWS and CGP offer Infrastructure as a Service (IAAS) and Platform as a Service (PAAS). From Seamgen.com, “The IAAS offerings cover servers, the associated operating systems, and storage. The PAAS offerings include email, message queues, databases, and web service hosts, to name just a few.” These are services offered by AWS. There are a broad and wide number of products that AWS offers compared to GCP. Google Cloud Platform’s products include computing and hosting services, storage services, networking services, and big data services. An advantage of GCP is that it is able to handle Big Data better than AWS. In Google Cloud Platform you can query large sets of data using a service called BigQuery. BigQuery is much faster at getting results than AWS’s data search service, EC2.

One disadvantage AWS has compared to to GCP is that AWS bills users by the hour, while GCP only bills people per minute. This makes GCP much cheaper since the user only has to pay for the time they need on the server, instead of getting charged even when you aren’t using the server. Another great thing about CGP according to the website Cloud Cruiser, “Google’s overall speed of I/O and ability to startup instances very quickly is an advantage”. From experience with AWS instances, it takes like 12 minutes just to start an instance and there are so many steps involved in creating one. However, SAWS offers the highest number of instances and the broadcast selection of technologies and databases. I only used EC2, S3, and EMR in AWS, but from what I saw when I clicked on the services tab, it was very extensive.

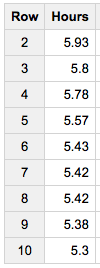
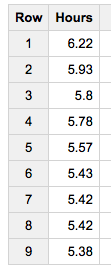
AWS is also one of the oldest cloud computing services around, and because of this, most people have been using AWS and are more comfortable with it. GCP is becoming more popular and is newer, so Google could possibly add more services in the future. Again, since AWS has been around longer than GCP, Amazon has more servers worldwide than Google does. Google mostly has servers in the US, with only a couple in other parts of the world.

If you wanted to crate your own app Google Cloud Platform would be one of your best options. Within GCP is a service called Google App Engine. This is an example of PAAS and from the Google Cloud Platform website, “if your application requires more computing resources because traffic to your website increases, Google automatically scales the system to provide those resources. If the system software needs a security update, that's handled for you, too.”

If you only wanted to store large amounts of data, then I would use AWS’s S3 service. Netflix is a company that uses this great service.

**Part 2:**

Using Google Cloud Platform I downloaded data sets and queried them. The data set I used was called ‘baseball’ and it has information about games from the 2016 MLB season. The reason I used this data set is because I love baseball and because I was intrigued with the data. One of the first things that came to mind was the fact that MLB games are so long and sometimes boring. I’ve always thought this, even though I still love watching the games. I also know that the MLB commissioner has tried coming up with ideas to speed games up and Bryce Harper even wore a hat that said “Make Baseball Fun Again”. From the Fox Sports website, “According to ESPN, the length of average MLB games in 2016 has stretched three hours and 26 seconds.” MLB games are some of the longest major sports games in the US.



Here is one of the queries I used:

#standardSQL

select round((duration\_minutes/60),2) as Hours

from `bigquery-public-data.baseball.schedules`

order by Hours desc

limit 10;

Here are the results: The results only show the top 9 times so I took a screenshot of the top 9 and then took a screen shot with the 10th time.

As you can see from the results, all the games were over 5 hours, when the average game lasts about 3 hours. That’s 2 more, long, and potentially hot hours that you have to sit through. MLB Commissioner Rob Manfred has tried implementing new pace-of-game rules to speed up the game, which have helped a bit, but long reviews have not helped. In 2014, MLB tested new pace-of-game rules in the Arizona Fall League. A couple of the rules tested include the “20-Second Rule” where “A pitcher shall be allowed 20 seconds to throw each pitch.” Another rule, called the “2:30 Inning Break Clock”, states that “There shall be a maximum 2:30 break between innings.” These along with 4 others could speed up the game a ton. I really hope they start using the pitch clock, since that would force players to pitch the ball faster and might even make the game more interesting.

Another thing I was interested in was how many teams played double headers last year. Currently, my hometown team, the St. Louis Cardinals, has already played one double header.

My query for this was:

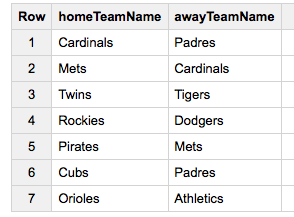
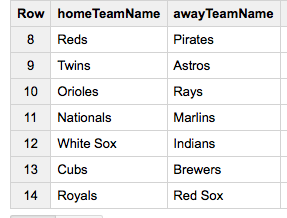
#standardSQL

select distinct homeTeamName, awayTeamName

from `bigquery-public-data.baseball.games\_wide` inner join `bigquery-public-data.baseball.schedules` using(homeTeamName, awayTeamName)

where gameNumber = 2;

The results were:



These 14 team matchups played double headers in 2016. The teams that played more than 2 double headers throughout the 2016 season include: Cardinals, Cubs, Pirates, Orioles, Mets, Twins, and Padres. All teams but Twins and Orioles are in the National League. The Twins and Cubs played both games at home, while the Padres played both games away. The MLB season has only been going on for about a month so maybe the Cardinals will play more double headers than last year.

The last thing I wanted to know was who was the fastest pitcher in 2016. A pitcher with a fast pitch is deadly, especially if the pitcher is accurate. Many pitchers are now starting to throw more than 100 mph and the average pitch speed in the MLB is 90.9 mph.

In my queries, I started off by checking how many pitchers could throw over 90 mph. Those results gave me 693 pitchers. Then I amped it up by checking how many pitchers threw over 95 mph. Those results gave me 349, almost half the number of the first result. Then I checked how many pitchers threw over 100 mph. 18 pitchers can.

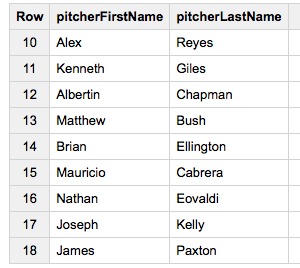
Here is the query used and here are the names of those 18 pitchers:

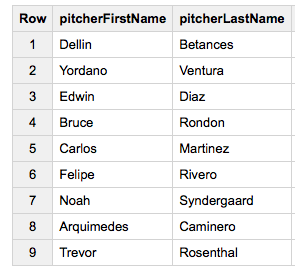
#standardSQL

select distinct pitcherFirstName, pitcherLastName

from `bigquery-public-data.baseball.games\_wide`

where pitchSpeed > 100;





The list became even smaller when I checked how many pitchers threw over 101 mph. There were only 5 pitchers who could throw that fast. The fastest pitcher in MLB history is Albertin Chapman (Aroldis Chapman). He once threw a 105.1 mph pitch.

**Sources:**

1. <http://www.foxsports.com/mlb/story/mlb-game-average-length-increase-how-to-fix-051716>
2. From Lab #3 I used the ROUND function to help me round the duration of games to 2 decimal digits.
3. <http://mlb.mlb.com/mlb/pace-of-game/>
4. <https://www.cloudcruiser.com/aws-azure-google-cloud/>
5. <http://stlouis.cardinals.mlb.com/schedule/index.jsp?c_id=stl#y=2017&m=4&calendar=DEFAULT>
6. <https://cloud.google.com/bigquery/docs/reference/legacy-sql#functions>
7. <http://m.mlb.com/statcast/leaderboard#pitch-velo>
8. <http://www.seamgen.com/blog/what-is-aws-amazon-web-services/>