**FINAL PROJECT REPORT**

**Data Set:**

I tried to use most of the topics covered in the course. I have taken a football data set from Kaggle which consists of all the international football results from 1872-2021. The volume of the data set is huge and has around 40000 records.

**Data description::**

* date - date of the match
* home\_team - the name of the home team
* away\_team - the name of the away team
* home\_score - full-time home team score including extra time, not including penalty-shootouts
* away\_score - full-time away team score including extra time, not including penalty-shootouts
* tournament - the name of the tournament
* city - the name of the city/town/administrative unit where the match was played
* country - the name of the country where the match was played
* neutral - TRUE/FALSE column indicating whether the match was played at a neutral venue

Link: [Dataset](https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017)

**Contents:**

● Introduction

● Background

● Methodologies & Results

● Discussion

● Conclusion

● References

**Introduction ::**

* In this project, I am working on to analyze the football data over the years. I would like to get my hands dirty on how does the big football teams makes use of the data available to them.
* This dataset includes **42,899** results of international football matches starting from the very first official match in 1972 up to 2019.
* This data set contains wide range of international football matches ranging from FIFA world cup to friendly matches
* The matches are strictly men's full international and the data does not include Olympic Games or matches where at least one of the teams was the nation's B-team, U-23 or a league select team.

**Background ::**

* Big data has truly transformed sports industry. It has lifted up the lid on sports being physical game . From sports like soccer, baseball as well as the fields like fantasy sports, each of this has begun depending on big data for enhancing the efficiency of its players and to work towards predicting future performances.
* Big data allows teams and companies to stay updated on the performances, carry out predictions and be resolute when it comes to sports field , In the competitive sports industry , collective knowledge of the player stats, abilities and competitive performance skills are the elements that propel the results
* With that motivation , I have selected this data set to look for one or more insights from the following:
* What trends have there been in international football throughout the ages - home advantage, total goals scored, distribution of teams' strength so on.
* Can we say anything about geopolitics from football fixtures - how has the number of countries changed, which teams like to play each other.
* What type of tournaments are most played by the teams.

**Methodologies and results:**

* I have many resources and technologies for this project. My data pipe line has essentially the following steps:
* Exploratory Data Analysis
* virtualization
* Data Transformation
* Data Loading
* Analysis
* Results and Visulaizations
* I have used the following tools and technologies:
* GCP
* Python
* Parallel processing
* PySpark
* Jupyter notebook
* Jetstream
* **Exploratory data analysis in PySpark**: Even though, I can do this by creating a Jupyter notebook in GCP, I have used Jetstream, I have the other modules that can be done in GCP. So in order to cover both of them I have engendered this step. The main motive of this step is to understand the data better before making any transformations.
* Ss:1, I have initially loaded the data into RDD by using the RDD. Printed the first 10 rows to see the format of the data.
* Ss:2, Since this being football matches data, I wanted to get a rough idea on how many international matches played by top teams, I have done this using a filter().
* Ss:3 In the third snapshot, I have displayed the above data in bar graph.
* Ss:4 &5, Since they are many different types of football matches like world cup, Friendly, Euro cup, I wanted to know what types of matches are more frequent than others. To get this, I have used the map reduce function and found the wordcounts. Once I got the word Counts, I can use a filter to get the count of different types of football matches played

Ss1: Loading into RDD

Graphical user interface, text, application, email

Description automatically generated

Ss2: International palyed by top teams like Germany, France, portugal etc

Text

Description automatically generated

Ss3: The above outputs in bar graph.

Graphical user interface, application

Description automatically generated

Ss:4

How many matches played by Tournament

Graphical user interface, text, application, email

Description automatically generated

Ss5: it is Clear that countries play more friendlies than any big events

Graphical user interface, text, application, email

Description automatically generated

* **Virtualization**: For the basic EDA, I have used pyspark. From this point I have made pure transformation to GCP hence I have called it Virtualization.
* Firstly, I have created a bucket under my project and loaded the csv file to the bucket.
* Post that I enabled the notebook API. I have loaded all the data from csv in the bucket to the data frame.

Graphical user interface, text, application

Description automatically generated

* Once I have everything in data frames, I have removed NAN values from the columns using simple commands as follows:

Chart

Description automatically generated with medium confidence

* After the above step, I have covereted true or false values to 1 and 0 respectively for easier calculations

Graphical user interface, application

Description automatically generated with medium confidence

* After this cleaning step, I have put this data frame back to csv in my bucket, to make it ready for transformation.

Graphical user interface, text, application, email

Description automatically generated

* **Data Transformation**: Once I have everything in bucket after cleaning, I am adding one more column which compares if the home team has won against the away team. If so, it will populated as won, other way, Loose and if the scores are equal, that will populate draw.

Graphical user interface, text, application, email

Description automatically generated

* **Data Loading:** Once you have transformed the data or added a new features we will load the data into big query or data lake.
* The below snapshot shows that we are taking the input from csv and bucket and loading into data lake football table. First arrow is input

Text

Description automatically generated

* The following snapshot is where we are initiating the pipeline

Text

Description automatically generated

* From the below snapshot, it is clear that we are getting each row from csv and loading it into big query table.

Text

Description automatically generated

* Once we run the above script there will be a table called football in the mentioned data lake: refer the ss below.

Graphical user interface, text, application, email

Description automatically generated

Football table.

* If you see the football table populated, you can see the status of the pipe line in green. Refer the ss below.

Graphical user interface

Description automatically generated with medium confidence

* **Analysis and Visualization:** Once we have the data in the big table, we can play with the data and get the part of data we want to analyze . In my Football data, I am only interested to analyze the big teams. And so I a using the following query.

Graphical user interface, text, application

Description automatically generated

* The output of the above query is only related to the top soccer teams

Table

Description automatically generated

* Now I have all the data in numbers, In order to make an impact we will create visualizations and let see what we can infer from it. I have used the inbuilt data studio to create visualizations.
* **Results and Visualizations:**
* **VIZ1**

Map

Description automatically generated

Inference1:Playing at home, most number of goals I scored by Spain compared to any other big teams. In the above viz, the size of the circle is proportional to number of goals. Portugal and Germany are the close followers of Spain and Fin land has scored very less compared to other nations when they are playing at home.

* **Viz2**

**Filter1**

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Taking all the matches where home team has lost

Inference2: from the above figure if we take Hungary as an example

In the matches Hungary lost at home, they have scored only 12 goals and away team has scored around 40 goals, hence their loss is quite evident.

* **Viz3:**

Inference: From the below viz, We can say that most international matches has happened in Budapest followed by Moscow

Chart, bar chart

Description automatically generated

**Viz4**

**Inference: if we consider teams playing at home and see the away goals and home goals, it is evident the dominance of the home teams and there is clearly a home advantage for big teams. In order to mitigate this factor, major tournaments are happening away from home stadiums.**Table

Description automatically generated

* **Discussion:**

I am able to put all the skills I have learned in this course like, Distributed computing, virtualization, analysis with pyspark, GCP and also working with pipelines(ETL) has made my project more easier. Every part of data life cycle is covered in my project starting from EDA, Data cleaning, Data Extraction, Data Transformation, Loading, analysis and visualization. The best thing with the cloud is that, I did not worry about any other competencies let it be memory or speed. My main motivation beyond taking this project is to show how easy and quick the analysis on sports has become. No wonder that Germany has used big data to win the FIFA World Cup in the year 2014.

* **Conclusion:**

These types of analytics are not just in Soccer, but widespread around sports industry, ranging from international matches to intra-school competitions. This also has an impact on virtual gaming such as fantasy sports. We can do much more taking the reference as my project such as Player Analysis, Team Analysis, Fan Analysis.

Data Science is both the present and future of sports analytics. Enhanced Artificial intelligence and improved machine learning models stood as a strong support for updating performance of the teams. In the present world where data is considered as the most valuable resource, categories like data science has emerged as a weapon that can be applied to enhance the winning chances of the teams in different circumstances.

After Completing this project, I am sure that I will be handling the resources more efficiently than before, this project made me explore the Cloud platforms to my max limit. Which in other case was just a storage bucket to me. I can purely concentrate on my analysis having everything at one place.

* References

[XSEDE User Portal | IU/TACC Jetstream User Guide](https://portal.xsede.org/jetstream)

[ETL Processing on Google Cloud Using Dataflow and BigQuery | Qwiklabs](https://www.qwiklabs.com/focuses/3460?locale=pl&parent=catalog)

[Ingesting New Datasets into BigQuery | Qwiklabs](https://www.qwiklabs.com/focuses/3692?catalog_rank=%7B%22rank%22%3A5%2C%22num_filters%22%3A0%2C%22has_search%22%3Atrue%7D&parent=catalog&search_id=14163071)

[International football results from 1872 to 2021 | Kaggle](https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017)