University of Abdelhamid Mehri-Constantine2
Faculty: NTIC
Department: IFA
Master’s second year SDIA 2024/2025

Lab 3 BDPA
Spark DataFrames

Apache Spark DataFrames are a powerful abstraction for working with structured and semi-structured data. Built on top of Spark’s RDDs, DataFrames provide a higher-level, SQL-like API that makes data manipulation more intuitive and efficient. In this Lab, We will explore that using the **EPL-Historical-Data** dataset, which contains information about matches played in the English Premier League during the 21st century. For more details, refer to this link.

Import the epl-training.csv file in your google Colab environment (You can mount your drive repertory) and read it as follows :

```
df = spark.read.csv(<FILEPATH>/epl-training.csv, header=True)
```

Write Pyspark codes that shows:

1. The list of all the involving teams
2. Top 5 teams with best attack at home
3. Top 5 teams with the worst defense away
4. Top 3 teams with the most goals scored
5. The team that played the most matches
6. Top 5 teams with best ratio goal scored/goal conceded in the last decade
7. Top 3 Best/worst team every year
8. The average goal per match in 3 differents periods (september to november, december to mars and avril to june)