**LSC Final Project a.y. 2020/21**

**Group Name: “people united”**

**Members: Gian Paolo Bernardini, Muhammad Umair Yousuf, Mehrpuya Fathi Bonyadabad**

**Title: RDD on NBA stats**

**Category (A/B/C): A**

1. APIs of the Spark ecosystems: advanced aspects of RDDs/DataFrames&SparkSQL/Spark Streaming, and other components, i.e., GraphX&GraphFrames and MLib
2. Optimisations such as custom partitioners, parameter tuning, performance monitoring, etc
3. Integration with data ingestion systems such as Kafka, other visual tools such as Nifi, cloud-based services such as Dataproc and Dataflow, etc

**Description:**

* Motivations and Goals: the motivations is to work on a field at the basis of the Map/Reduce pattern. Although RDD syntax is a low level one, it is the foundation for all the subsequent high level API like DataFrames and SparkSQL.

The queries are run on stats of the NBA, thus dealing with an interesting sport.

* Technical Description: the queries will be basically in RDD syntax.
* Programming Exercises: write queries in RDD syntax.
* Educational Goals: the goal is to get acquainted with the syntax of RDD.

**Programming Exercises (8 Questions) with Solutions:**

queries\_1.ipynb - Gian Paolo Bernardini

Q1 which is the most winning team in all the decade, based on WIN% (W/GP)?

Q2 is scoring a lot of points important for the winning of a title?

Q3 how the FGM (Field Goals Made) percentage influences the winning of a title? (Compared to the mean of those who have won a title)

Q4 how the mean of FGM for the winners is near to the mean of the team who has the best FGM percentage?

queries\_2.ipynb - Mehrpuya Fathi Bonyadabad

Q1 Which team has the most field GOAL percentage? (and its win%) (gathering part)

Q2 Which team has the most field GOAL percentage? (and its win%) (reduce part)

Q3 Can be (the best most field GOAL) guarantee to be (the best winner)?

Q4 How about be less than average (to be in top 10)?

Q5 Group by data on their seasons.

Q6 Whitch are faster (RDD,SQL,Data frame) in a (SELECT , SORT) problem?

Q7 Whitch are faster (RDD,SQL,Data frame) in a (SELECT , FILTER) problem?

queries\_3.ipynb - Muhammad Umair Yousuf

Q1 Each Team Win and Lost Count?

Q2 Select top 10 teams with its wining percentage and with total points?

Q3 In this Query group data by wining percentage?

Q4 In this Query group data by Team Name?

Q5 In this Query Reduce by key and calculate the average by the key?

**Links**

* Material: https://github.com/gianpaolobernardini/rdd\_nba