UJIAN TENGAH SEMESTER GANJIL TA. 2022 / 2023

Fakultas : Ilmu Komputer

Program Studi : Informatika

Mata Kuliah dan kode mk : BIGDATA & PREDICTIVE ANALYTICS LANJUT (ST153)

Sifat Ujian : Buku terbuka

Tanggal Ujian : 26 November 2022, 08.00

Dosen Pengampu : Arif Dwi Laksito, M.Kom

**NAMA : FATHURRAHMAN NUR AZIZ**

**NIM : 20.11.3694**

**KELAS : 20 IF 07**

Dengan dataset yang didapat dari <https://www.kaggle.com/datasets/artimous/complete-fifa-2017-player-dataset-global/code?select=FullData.csv>, lakukan:

1. Setup PySpark pada Google Colab (5, SCPMK 1534002)
2. Baca data dari FullData.csv, kemudian pecah setidaknya menjadi 2 RDD (players, clubs) (10, SCPMK 1534104)
3. Ubah RDD di atas menjadi PySpark Dataframe (5, SCPMK 1532205)
4. Lakukan agregasi untuk mendapatkan jumlah pemain untuk masing-masing kelompok berikut: (25, SCPMK 1534104)
   1. Berdasarkan Nationality
   2. Berdasarkan Club
5. Lakukan agregasi untuk mendapatkan nilai: (25, SCPMK 1534104)
   1. Rata-rata Rating berdasarkan Club\_Position
   2. Maximum Rating berdasarkan National\_Position
6. Lakukan setidaknya 3 operasi transform RDD pada data yang kalian punya, contoh reduceByKey, sortByKey, flatMap, etc. (30, SCPMK 1532206)

Setiap codeblock harus disertakan penjelasan proses yang terjadi/dilakukan.

Kumpulkan hasil pekerjaan notebook dalam format PDF.

**Jawab**

|  |  |
| --- | --- |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 | TRANSFORM 1 DENGAN CountByKey()    TRANSFORM 2 DENGAN filter()    TRANSFORM 3 DENGAN CountByValue() |