**Assignment Day 13 - Big Data Analytics with PySpark Part 2**

**Mochamad Ananda Radya Mikola**

1. **Tahapan dan Penjelasan dalam Menjalankan Script**

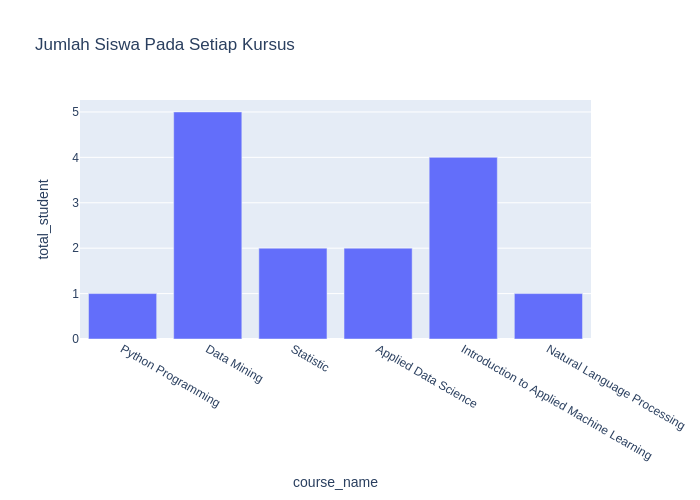
Saya mengerjakan atau membuat script ini menggunakan google colab. Untuk menjalankan script saya, bisa di run secara berurutan dari atas yaitu dari Data Preparation hingga Data Load. Namun, untuk urutan dan penjelasan yang lebih rinci terkait menjalankan scriptnya, saya jelaskan sebagai berikut.

1. Section Data Preaparation (Extract)
2. Menjalankan kode untuk menginstall pyspark dan juga kaleido
3. Melakukan import library yang dibutuhkan untuk sepanjang proses cleaning, transform, load serta anilisisnya.
4. Melakukan mounting ke drive saya, dimana file yang diberikan oleh Mas Albert saya simpan di drive saya.
5. Melakukan pembentukan path directory, dimana file-file yang digunakan nanti akan di read menggunakan SparkSession.
6. Pembuatan nama aplikasi
7. Melakukan pembuatan dataframe sesuai tabel dari data yang ada.
8. Section Data Cleaning dan Transform
9. Melakukan pembersihan dan transform pada data tabel **course\_attendance** yang meliputi pengubahan nama kolom dan mengubah format kolom attend\_date menjadi format date yang benar.
10. Melakukan pembersihan pada data tabel **course** hanya pengubahan nama kolom saja.
11. Melakukan pembersihan dan transform pada data tabel **enrollment** yang meliputi pengubahan nama kolom dan pengubahan format kolom enroll\_date menjadi format date yang benar.
12. Melakukan pembersihan dan transform pada data tabel **schedule** yang meliputi pengubahan nama kolom, pengubahan format kolom start\_date dan end\_date menjadi format date yang benar, dan mengubah data pada kolom course\_days yang sebelumnya berupa angka menjadi nama hari sesuai dengan angka urutan hari di dalam seminggu.
13. Section Data Analysis dan Visualization
    1. Pada analisis 1, melakukan join tabel yang dibutuhkan terlebih dahulu dan select sesuai nama kolom yang dibutuhkan agar lebih rapih saat di show(). Kemudian melakukan analisis dengan groupBy kolom course\_name dan menghitung jumlah siswa pada setiap kursus yang dipilih. Setelah itu, data analisis tersebut diubah ke pandas untuk bisa dilakukan visiualisasi. Sehingga pada visualisasinya, terlihat grafik bar terkait jumlah siswa pada setiap kursus.
    2. Pada analisis 2, melakukan join tabel yang dibutuhkan terlebih dahulu dan select sesuai nama kolom yang dibutuhkan agar lebih rapih saat di show(). Kemudian dilakukan filtering data terlebih dahulu sebelum dilakukan analisis, dimana tujuan filter yang dimaksud adalah untuk memisahkan data pada tahun 2019 dengan 2020. Setelah itu, dilakukan anilsis dengan groupBy kolom course\_name dan menghitung jumlah siswa yang hadir pada kursus yang ada pada masing-masing data yaitu tahun 2019 dan tahun 2020. Setelah itu, data analisis tersebut diubah ke pandas untuk bisa dilakukan visiualisasi. Sehingga pada visualisasinya, terlihat grafik bar terkait total kehadiran siswa tahun 2019 dan tahun 2020.
    3. Pada analisis 3, melakukan join tabel yang dibutuhkan terlebih dahulu dan select sesuai nama kolom yang dibutuhkan agar lebih rapih saat di show(). Kemudian dilakukan filtering data terlebih dahulu sebelum dilakukan analisis, dimana tujuan filter yang dimaksud adalah untuk memisahkan data pada tahun 2019 dengan 2020. Selain filtering, juga dilakukan pemisahan data yaitu split yang terdiri dari multiple data pada kolom course\_days. Setelah itu, dilakukan analisis dengan groupBy pada kolom day yang baru dibuat saat dilakukan pemisahan, menghitung jumlah kursus yang ada pada masing-masing hari dalam seminggunya, dan disortir dengan orderBy sesuai dengan urutan hari dalam seminggu.

Sehingga pada visualisasinya, terlihat grafik bar terkait total kursus yang aktif dalam seminggu pada tahun 2019 dan tahun 2020.

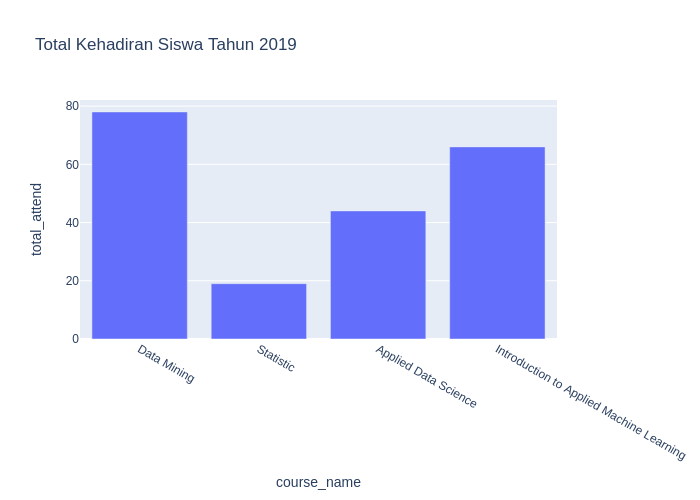
* 1. Pada analisis 4, menggunakan dataframe yang telah di join pada analisis 3. Kemudian dilakukan pemisahan data terlebih dahulu pada kolom course\_days yang memiliki beberapa data didalamnya. Setelah itu, melakukan analisis dengan groupBy pada kolom lecturer\_id, menghitung jumlah hari mengajar oleh setiap lectur pada tahun ajaran 2019/2020, dan disortir dengan orderBy pada kolom lecturer\_id. Saya juga mengubah format data pada kolom lecturer\_id dari integer menjadi string. Sehingga pada visualisasinya akan menjadi lebih rapih dan memperlihatkan grafik bar terkait total hari mengajar oleh setiap lectur pada tahun ajaran 2019/2020.
  2. Pada analisis 5, mencari tahu nilai korelasi dengan menggunakan dataframe yang digunakan untuk visualisasi pada analisis ke 1 dan 2. Adapun cara mendapatkan nilai korelasi dengan menggunakan fungsi corr terhadap dataframe analisis 1 yang telah di join dengan dataframe analisis 2.

1. Section Data Load
   1. Melakukan export grafik ke image dengan menggunakan kaleido dengan command write\_image.
   2. Melakukan export dataframe analisis ke csv dengan command write.format dan di save ke directory folder yang diinginkan.
2. **Penjelasan Hasil Analisis**
3. Analisis 1

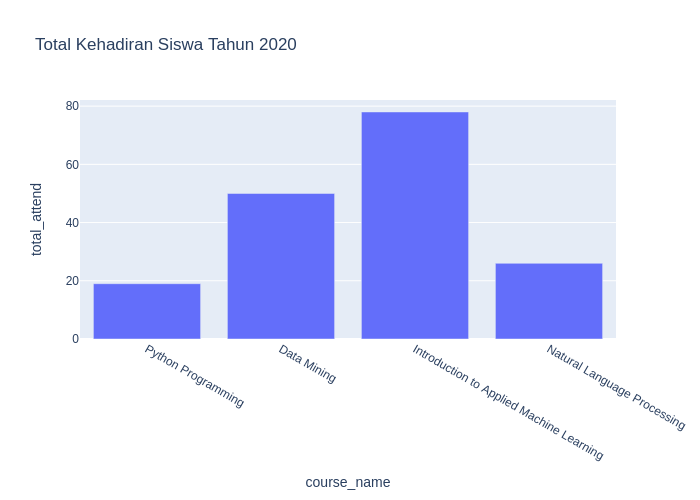
Mengetahui berapa banyak siswa pada masing-masing kursus dan kursus apa yang paling diminati oleh siswa tersebut.

Pada grafik diatas dapat disimpulkan bahwa dari 10 kursus yang tersedia pada tahun ajaran 2019/2020, hanya 6 kursus yang aktif. Sedangkan 4 kursus lainnya tidak muncul pada grafik atau tidak aktif karena tidak memiliki siswa. Untuk kursus yang paling diminati oleh para siswa adalah kursus Data Mining yang memiliki 5 siswa, kemudian diikuti oleh kursus Introduction to Applied Machine Learning yang memiliki 4 siswa. Terdapat masing-masing 2 siswa pada kursus Statistic dan Applied Data Science. Terdapat masing-masing 1 siswa pada kursus Python Programming dan Natural Languange Processing.

1. Analisis 2

Mengetahui total kehadiran siswa pada setiap kursus di tahun 2019 dan 2020.

Pada grafik diatas dapat disimpulkan bahwa total kehadiran siswa paling tinggi pada kursus yang aktif khususnya pada tahun 2019 merupakan kursus Data Mining. Kemudian diikuti secara berurutan oleh kursus Introduction to Applied Machine Learning dan Applied Data Science. Sedangkan untuk total kehadiran siswa terendahnya adalah kursus Statistic.

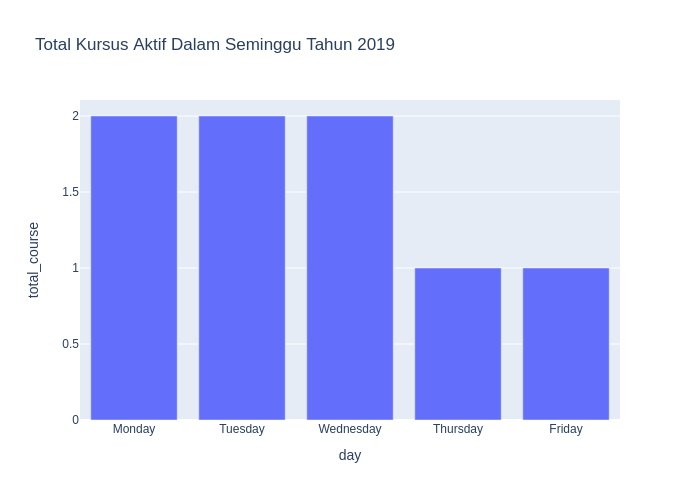
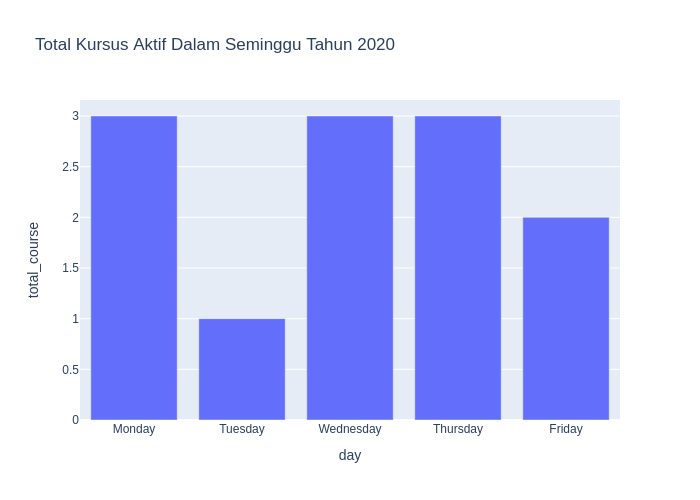


Pada grafik diatas dapat disimpulkan bahwa total kehadiran siswa paling tinggi pada kursus yang aktif khususnya pada tahun 2020 merupakan kursus Introduction to Applied Machine Learning. Kemudian diikuti secara berurutan oleh kursus Data Mining dan Natural Languange Processing. Sedangkan untuk total kehadiran siswa terendahnya adalah kursus Python.

Kedua grafik diatas menunjukkan bahwa total kehadiran siswa pada setiap tahunnya dapat berbeda antara tahun 2019 dan 2020.

1. Analisis 3

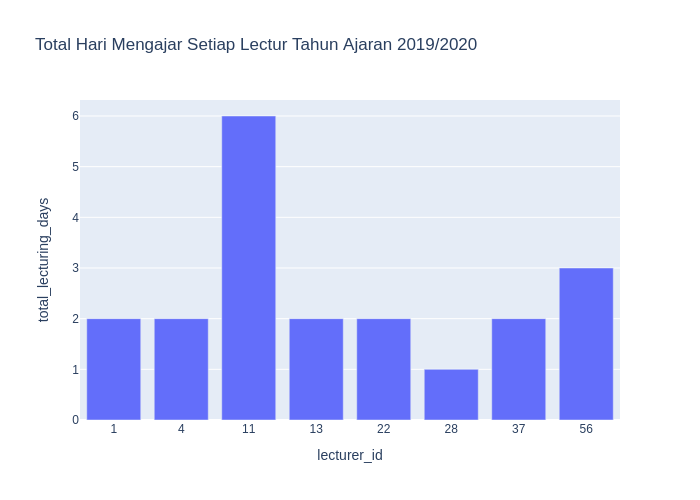
Mengetahui pola hari dalam seminggu di mana kursus yang paling sering dijadwalkan pada tahun 2019 dan 2020.

Pada grafik diatas dapat disimpulkan bahwa hari yang paling sering dijadwalkan dari total kursus yang aktif dalam seminggu khususnya pada tahun 2019 adalah hari senin, selasa, dan rabu yang terdiri dari 2 kursus. Sedangkan hari kamis dan jumat hanya terdiri dari 1 kursus saja.

Pada grafik diatas dapat disimpulkan bahwa hari yang paling sering dijadwalkan dari total kursus yang aktif dalam seminggu khususnya pada tahun 2020 adalah hari senin, rabu, dan kamis yang terdiri dari 3 kursus. Sedangkan hari jumat terdiri dari 2 kursus dan hari selasa terdiri dari 1 kursus.

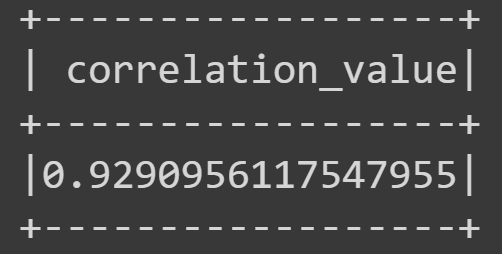
Kedua grafik diatas, dapat disimpulkan kembali bahwa hari yang paling sering dijadwalkan untuk kursus yang aktif di setiap tahun 2019 dan 2020 bisa berbeda.

1. Analisis 4

Mengetahui total hari mengajar kursus oleh setiap dosen pada tahun ajaran 2019/2020.

Pada grafik diatas dapat disimpulkan bahwa dosen yang paling aktif mengajar dalam tahun ajaran 2019/202 adalah dosen dengan id nomor 11 yang memiliki jumlah total mengajarnya sebanyak 6 hari. Kemudian, diikuti oleh dosen dengan id nomor 56 dengan total mengajar sebanyak 3 hari. Untuk sisanya memiliki jumlah total mengajar sebanyak 2 hari dan 1 hari.

1. Analisis 5

Pada analisis ke 5 ini adalah untuk mengetahui nilai korelasi antara hasil analisis 1 dengan analisis 2.

Pada gambar diatas, antara analisis 1 dengan analisis 2 menunjukan angka hasil korelasi 0,93 (saya bulatkan) yang artinya mendekati 1 menunjukkan bahwa terdapat korelasi positif yang kuat antara jumlah siswa yang mengikuti kursus dan tingkat kehadiran siswa pada kursus tersebut. Dalam hal tersebut, berarti:

* Ketika jumlah siswa yang mengikuti kursus meningkat, cenderung meningkatkan juga tingkat kehadiran siswa pada kursus tersebut.
* Ketika jumlah siswa yang mengikuti kursus berkurang, cenderung juga mengurangi tingkat kehadiran siswa pada kursus tersebut.