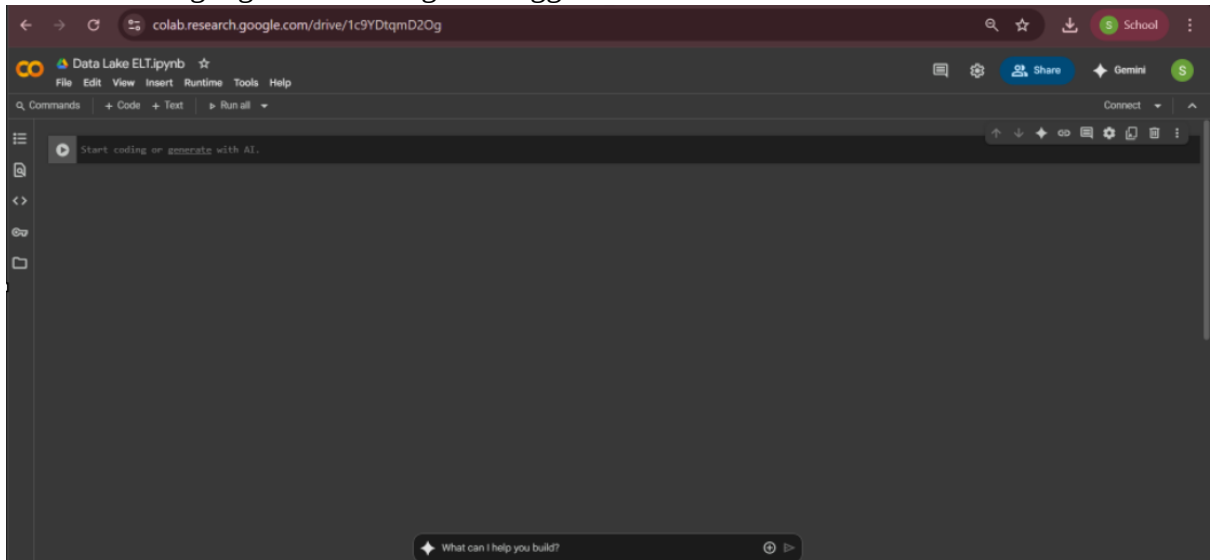


CARA MENJALANKAN PROGRAM

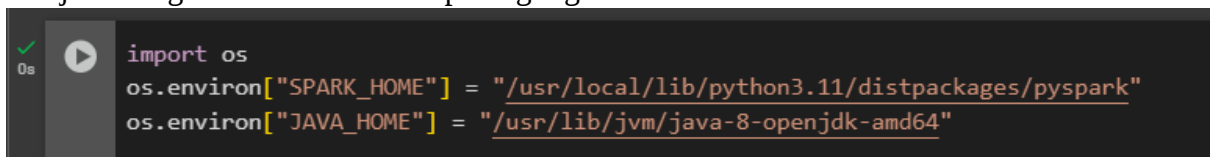
1. Program ini bertujuan untuk mencoba mesimulasikan Data Lake dengan menggunakan google drive dan google colabs. Langkah yang pertama kalian masuk ke dalam environement google colabs dengan menggunakan akun email dari UTDI.



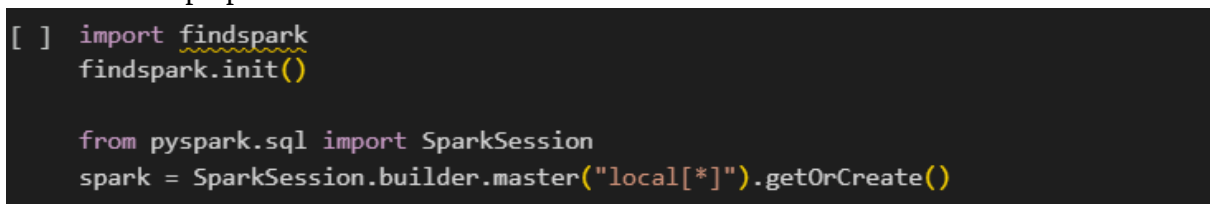
2. Lalu inputkan code untuk menginstall openjdk 8, Hadoop v.3.2 (versi terbaru), dan Apache spark sebagai integrasi.



3. Setelah instalasi telah dieksekusi kalian lanjutkan untuk mengkonfigurasi apache spark dan java dengan memasukkan script ke google colabs.



4. Pastikan apache spark berada dengan melakukan inisialisasi dan melakukan session di dalam local laptop/PC.



5. Lakukan ekstrak data dari google colab untuk melakukan proses ini yang melibatkan autentikasi dan otorisasi pengguna demi memberikan akses ke Google Drive. Setelah

berhasil dipasang, pengguna dapat mengakses, membaca, menulis, dan memodifikasi file dalam Google Drive melalui direktori `/content/drive/MyDrive/`.

```
✓ 20s [9] from google.colab import drive
      drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

6. Selanjutnya lakukan proses load data dari google drive untuk menampilkan isi dari file `tempat_wisata.csv`.

```
[ ] data_path = "/content/drive/MyDrive/Colab Notebooks/tempat_wisata.csv"
    df = spark.read.csv(data_path, header=True, inferSchema=True)

    df.show()
```

```
┌-----┐┌-----┐┌-----┐┌-----┐
| Nama_Tempat | Provinsi | Kode_Pos | Jumlah_Pengunjung |
├-----┴┴-----┴┴-----┴┴-----┴┴-----┘
| Raja Ampat | Papua | 98482 | 79079 |
| Pantai Kuta | Bali | 80361 | 304819 |
| Candi Borobudur | Yogyakarta | 56553 | 565530 |
| Pulau Komodo | NTT | 86554 | 202184 |
| Pantai Bunaken | Sulawesi Utara | 95242 | 162346 |
└-----┴┴-----┴┴-----┴┴-----┴┴-----┘
```

7. Setelah proses load data berhasil dilanjutkan untuk melakukan proses transformasi dataframe yang ditampilkan dari jumlah pengunjung yang akan datang ke 4 lokasi tempat wisata.

```
[ ] transformed_df = df[df[" Jumlah_Pengunjung"] > 100000][["Nama_Tempat", " Jumlah_Pengunjung"]]
    transformed_df.show()
```

```
┌-----┐┌-----┐
| Nama_Tempat | Jumlah_Pengunjung |
├-----┴┴-----┘
| Pantai Kuta | 304819 |
| Candi Borobudur | 565530 |
| Pulau Komodo | 202184 |
| Pantai Bunaken | 162346 |
└-----┴┴-----┘
```

8. Setelah proses transformasi data selesai, kita lanjutkan proses akhir dari program ini dengan menulis kembali dalam bentuk CSV dengan format bebas kalian tentukan.

```
[ ] transformed_df.write.csv("/content/drive/MyDrive/Colab Notebooks/transformed_data.csv", header=True)
```

- ▼  Colab Notebooks
 - ▶  .ipynb_checkpoints
 - ▼  transformed_data.csv
 -  ._SUCCESS.crc
 -  .part-00000-b22583a5-67...
 -  _SUCCESS
 -  part-00000-b22583a5-671...