Pada tahap ini, setelah data diambil dari sumber tertentu, ia dimasukkan pada suatu environment. Lantas data diproses agar bisa diolah oleh model machine learning. Menjalankan proses machine learning sama seperti mengajari seorang anak kecil. Ketika mengajari anak kecil membedakan antara buah apel dan buah jeruk, kita tinggal memperlihatkan buahnya dan memberi tahu mana apel dan mana jeruk. Namun demikian, komputer saat ini belum secanggih dan seintuitif itu sehingga kita perlu mempersiapkan data dengan data processing agar bisa dimengerti komputer.

Salah satu library yang paling populer untuk pengolahan data dalam machine learning adalah Pandas Library. Pandas Library adalah salah satu library yang wajib Anda kuasai dalam bidang machine learning.

**Pandas Library**

Pandas adalah sebuah library open source yang dipakai untuk menganalisis dan memanipulasi data. Pandas dibangun menggunakan bahasa pemrograman Python yang menawarkan struktur data dan operasi untuk manipulasi tabel numerik dan time series. Tabel numerik adalah tabel yang berisi bilangan numerik dan Tabel time series adalah tabel yang berubah seiring waktu, misalnya tabel yang memuat perubahan nilai pasar saham untuk setiap menitnya.

Berbagai jenis data yang umum dipakai dalam ML seperti CSV dan SQL dapat diubah menjadi dataframe pandas. Dataframe adalah sebuah tabel yang terdiri dari kolom dan baris dengan banyak tipe data di dalamnya. Pandas terintegrasi dengan library machine learning yang populer seperti Scikit Learn (SKLearn) dan Numpy.

Pandas mendukung banyak jenis data yang dapat dipakai dalam sebuah project machine learning. Berikut adalah beberapa contoh data yang dapat diolah dengan pandas.

* CSV  
  CSV adalah sebuah format data di mana elemen dari setiap baris dipisahkan dengan koma. CSV sendiri adalah singkatan dari Comma Separated Value.
* SQL  
  Structured Query Language adalah sebuah data yang berasal dari sebuah relational database. Format data ini berisi sebuah tabel yang memiliki format data seperti integer, string, float, dan biner.
* EXCEL  
  Excel adalah berkas yang didapat dari spreadsheet seperti Microsoft Excel atau Google Spreadsheet. File Excel biasanya memuat data numerik.
* SPSS  
  SPSS atau Statistical Package for the Social Science adalah sebuah berkas dari perangkat lunak yang biasa dipakai untuk statistik dan pengolahan data. Berkas SPSS yang disimpan memiliki ekstensi .sav.
* JSON  
  JSON atau Javascript Object Notation adalah salah satu format data yang menggunakan sistem Key - Value di mana sebuah nilai disimpan dengan key tertentu untuk memudahkan mencari data.

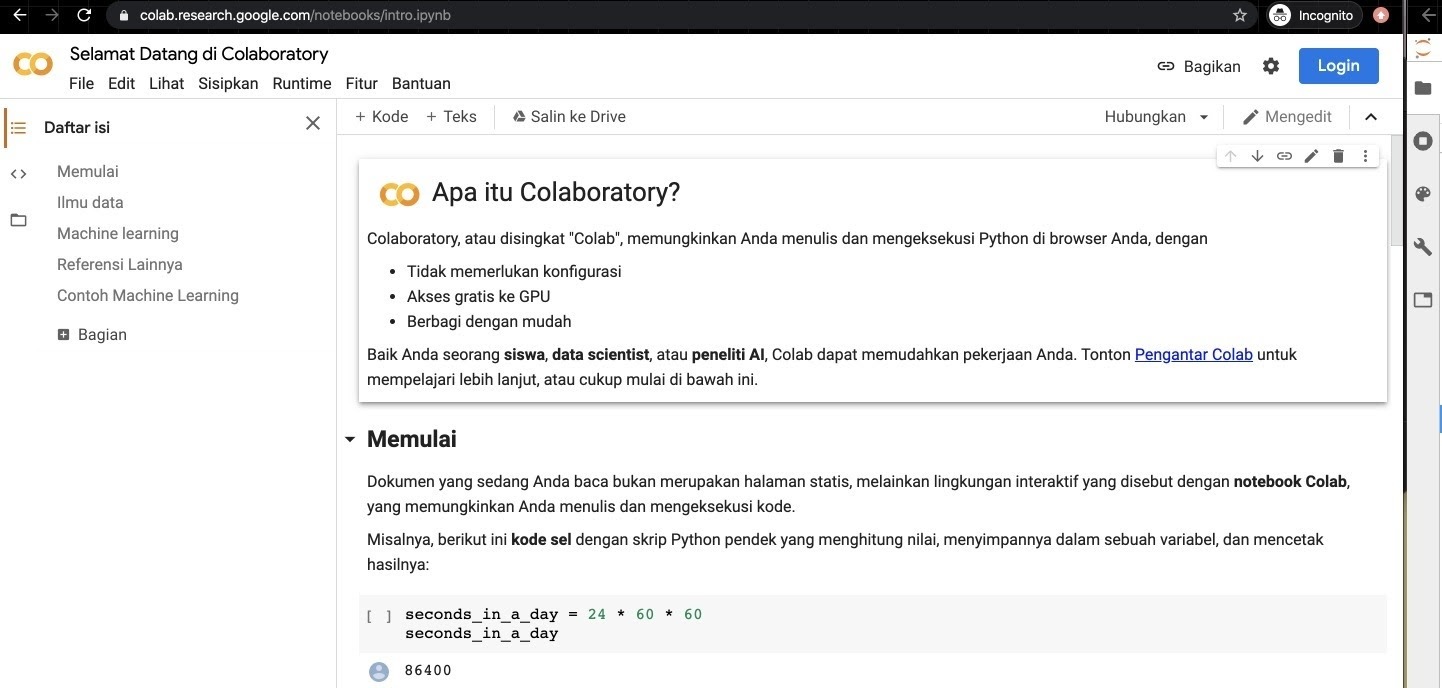
Pada kelas ini kita hanya akan menggunakan data berjenis csv karena data tipe ini mudah didapat di situs penyedia data seperti yang telah dibahas di sub-modul sebelumnya.

Untuk tahapan selanjutnya Anda dapat menulis dan menjalankan kode pada Google Colaboratory.

**Google Colaboratory**

Google Colaboratory atau sering juga disebut “Colab” adalah sebuah peranti  dari Google yang dibuat untuk keperluan mengolah data, belajar, dan bereksperimen khususnya dalam bidang Machine Learning. Colab berjalan sepenuhnya pada Cloud dengan memanfaatkan media penyimpanan Google Drive.

Untuk membuka Google Colab, Anda dapat membuka tautan ini [colab.research.google.com](http://colab.research.google.com/). Berikut adalah tampilan antarmuka dari Google Colaboratory. Untuk membuat dan menyimpan proyek pada Google Colab, Anda harus login terlebih dahulu menggunakan akun Google Anda.

[](https://www.dicoding.com/academies/184/tutorials/8330?from=8342)

Untuk mengetahui lebih lanjut tentang Colab, kunjungi [tautan](https://imam.digmi.id/post/google-colab-gratis-untuk-belajar-deep-learning/) berikut.