

## Loading the Dataset

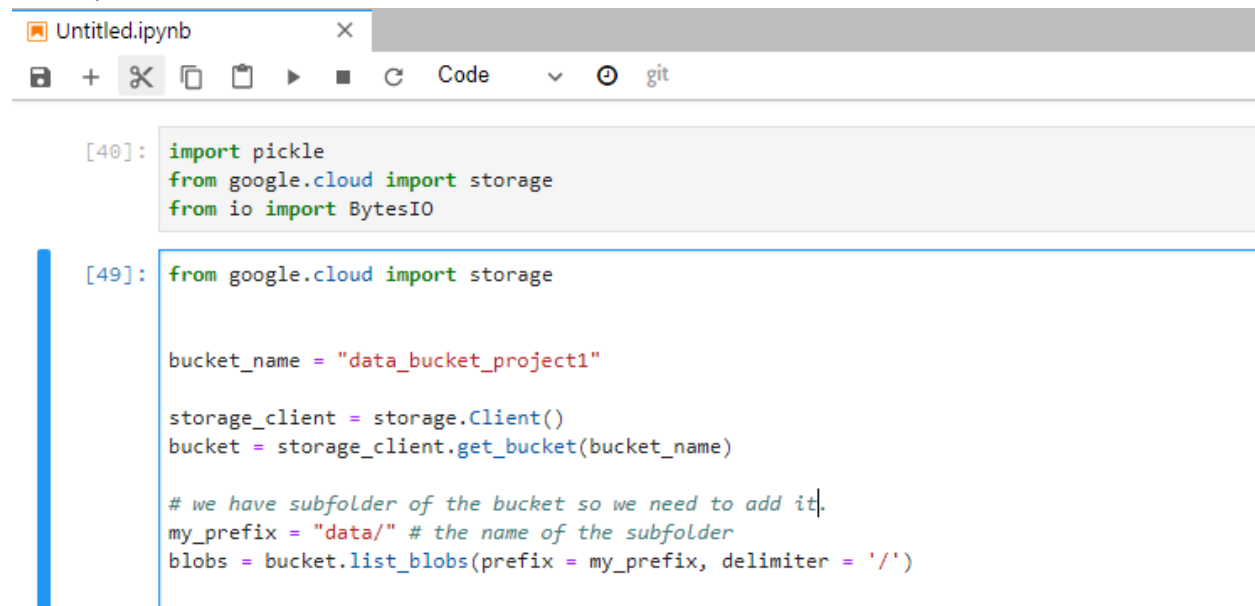
After acquiring the dataset, we decide to choose Google Cloud Platform to work on the project. We will need to load the data into a storage bucket, then try to import it on a local Jupyter Notebook in order to have some visualization and processing.

As the files are of the type '.pkl' we will have to use the python library pickle to use the actual stored data in those files.

## Visualizing the Data

Next, we boot up notebook, and start by importing the libraries needed, we will be using the google storage so it must be called.

Our imports look like this:

A screenshot of a Jupyter Notebook window titled 'Untitled.ipynb'. The interface includes a toolbar with icons for saving, adding, deleting, and running code, along with a 'Code' dropdown menu and a 'git' icon. Two code cells are visible. The first cell, labeled '[40]:', contains the following code: 

```
import pickle
from google.cloud import storage
from io import BytesIO
```

 The second cell, labeled '[49]:', contains the following code: 

```
from google.cloud import storage

bucket_name = "data_bucket_project1"

storage_client = storage.Client()
bucket = storage_client.get_bucket(bucket_name)

# we have subfolder of the bucket so we need to add it.
my_prefix = "data/" # the name of the subfolder
blobs = bucket.list_blobs(prefix = my_prefix, delimiter = '/')
```

We have also specified the bucket and folder in order to allow the loading.

Next up, we need to work on the specifics of those files, since there are 350 different files we will use a loop to load all of them into our session, to do that, we use the following code:

```
for blob in blobs:
    if(blob.name != my_prefix): # ignoring the subfolder itself
        file_name = blob.name.replace(my_prefix, "")
        blob.download_to_filename(file_name) # download the file to the machine
        with open(file_name, 'rb') as pickle_file:
            event = pickle.load(pickle_file)
```

After running this, the files started to load into the session, it takes up a few minutes since there are many of them.

Next up, a first visualization, for that we follow this procedure: (we base this on the documentation)

- specifying a file to use

- specifying the data and the target (as the 2 columns present in the files)

- using our preliminary info about the labels of the particles (thus, labeling the particle codes), see as follows:

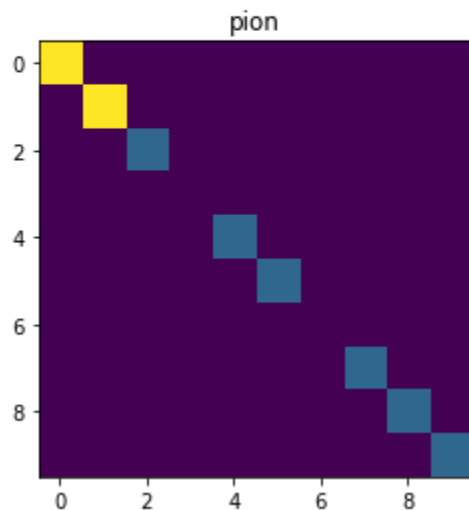
```
pkl_file = open('event1.pkl', 'rb')
event1 = pickle.load(pkl_file)
```

```
data,target=event1[0],event1[1]
target=target.astype('int')
```

```
dic_types={11: "electron", 13 : "muon", 211:"pion", 321:"kaon",2212 : "proton"}
```

So far we have only worked on reading the data, now we visualize it, using matplotlib:

```
[54]: import matplotlib.pyplot as plt
plt.title(dic_types[target[0]])
plt.imshow(data[0])
plt.show()
```



We can see the distinct values for the first particle found in the file column, which seems to be 'pion'.

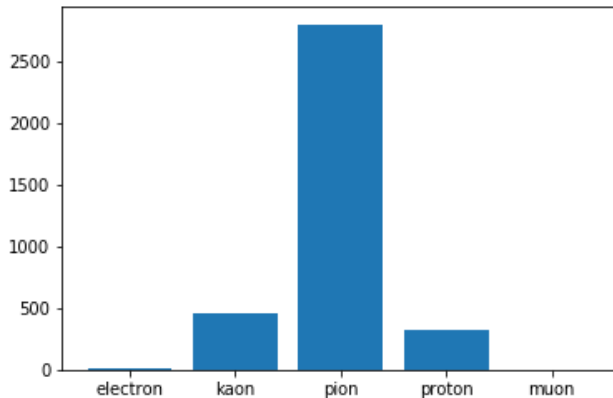
## Working on the Data

Now let us try to find the different number of particles present in this event.

```
from collections import Counter

plt.bar(range(len(dic_types)), list(Counter(target).values()))
plt.xticks(range(len(dic_types)), [dic_types[i] for i in list(Counter(target).keys())])

plt.show()
```



We use count to find the occurrences of the values specific to each labeled particle. This file seems to include pions more than any other particle. We should proceed on this to be able to predict the type of particle based on the events.

The type of the event file is `numpy.ndarray` and the shape is (2, 3598).