**Preprocesado de los datos**

El preproceso de los datos lo hago con un archivo en Python “1.2 GetCSVDataFRomSimulation.py”

Y los datos se cargan de las distintitas carpetas:

/ 2A1DD7CB-EAA3-4D50-944D-C6F77031592C

/ 0320BF40-5294-46A7-9CF8-A9DE41A66D3C

/ 389FE1BC-1058-4FF0-B428-42F92B19222A

/ A8266BA2-5937-45D5-90BD-79875E9E7255

/ D4147996-D704-4328-8891-3BE1E29E76BE

/ EE7F8FBF-3683-4435-9C3A-E4A5A3BD0BFD

que corresponde a los nombres de los JOBS en el cluster, en estas carpetas se encuentran los archivos solución de cada una de las simulaciones realizadas.

Estos archivos están en formato JSON se extrae la información haciendo un PARSER usando la librería de Python import json.

Destacar que el tarjet de salida se ha modificado respecto a la presentación inicial realizada el 15 de junio, donde se buscaban dos características A y B para calcular la diferencia presión y en la actualidad se busca ya directamente la diferencia de presión ya que existe una relación *P = A\*Q + B\*QQ*, por otro lado destacar que durante el pase se han incluido algunos atributos más de los que se contaban inicialmente en el powerpoint pero que se han incluido simplemente para determinar que los datos arrojados por las simulaciones eran “creibles”, porque en ocasiones las simulaciones fallaban o el proceso de mallado no era suficientemente bueno y se producían errores.

Me hubiera gustado contar con más filas para el dataset final, pero como he comentado antes se produjeron errores de simulaciones repetidas o anómalas y al ser computacionalmente tan costosos los procesos solo han podido ser 420 filas correctas.