UNIVERSIDAD EAFIT

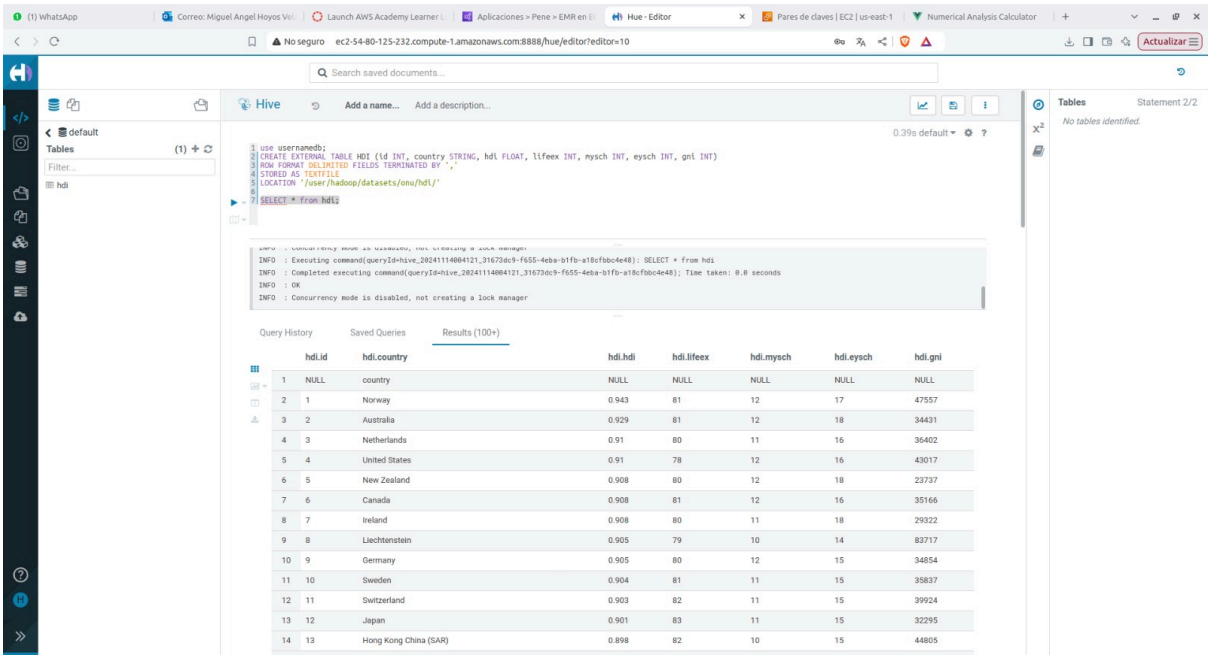
Tópicos Especiales En Telemática

Laboratorio 3-2 HIVE

Miguel Angel Hoyos Velasquez

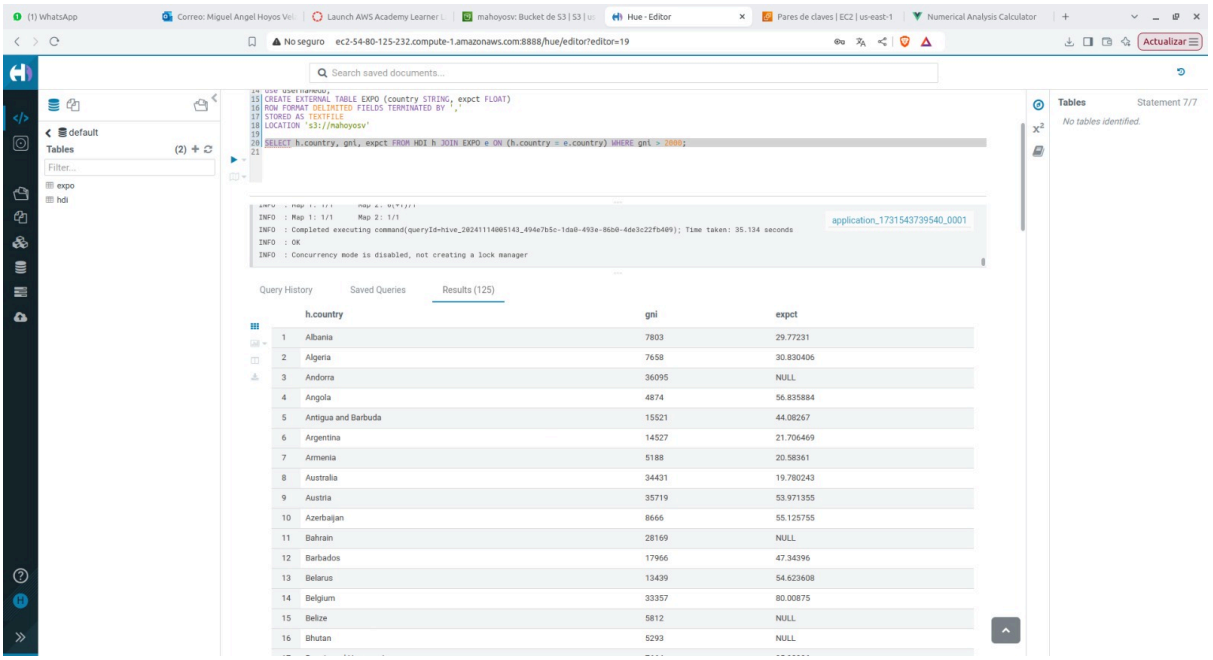
Medellín, 18 de Noviembre 2024

Aquí se aprecia la ingesta de datos de manera manual, sobre el mismo cluster.



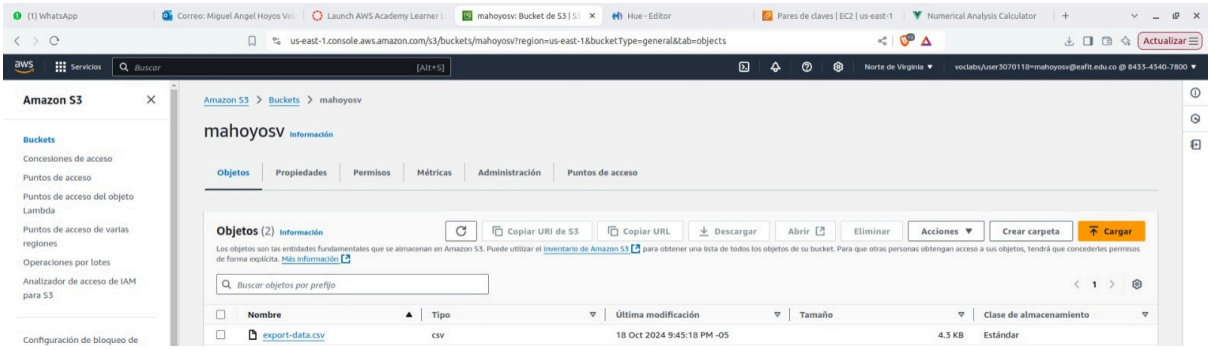
	hdi.id	hdi.country	hdi.hdi	hdi.lifeex	hdi.mysch	hdi.eysch	hdi.gni
1	NULL	country	NULL	NULL	NULL	NULL	NULL
2	1	Norway	0.943	81	12	17	47557
3	2	Australia	0.929	81	12	18	34431
4	3	Netherlands	0.91	80	11	16	36402
5	4	United States	0.91	78	12	16	43017
6	5	New Zealand	0.908	80	12	18	23737
7	6	Canada	0.908	81	12	16	35166
8	7	Ireland	0.908	80	11	18	29322
9	8	Liechtenstein	0.905	79	10	14	83717
10	9	Germany	0.905	80	12	15	34854
11	10	Sweden	0.904	81	11	15	35837
12	11	Switzerland	0.903	82	11	15	39924
13	12	Japan	0.901	83	11	15	32295
14	13	Hong Kong China (SAR)	0.898	82	10	15	44805

Posteriormente la definición de la tabla. Esta definirá la estructura de los datos con los que se trabajarán. Se hizo una consulta para validar el correcto funcionamiento.

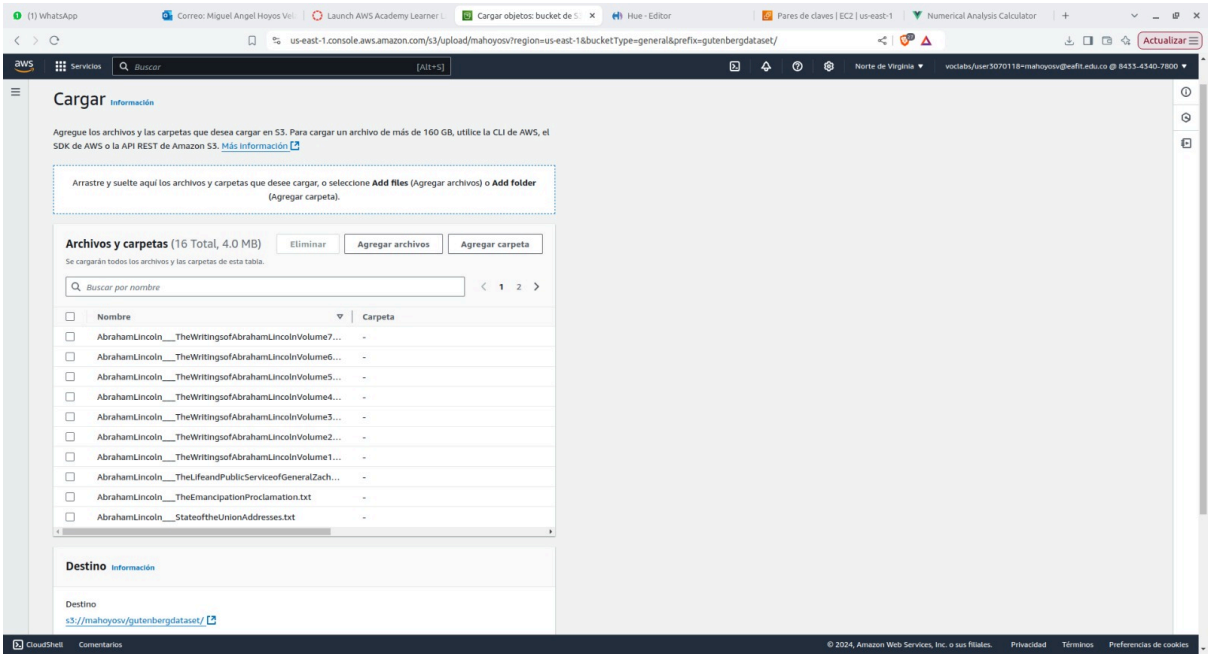


	h.country	gni	expct
1	Albania	7803	29.77231
2	Algeria	7658	30.830406
3	Andorra	36095	NULL
4	Angola	4874	56.835884
5	Antigua and Barbuda	15521	44.08267
6	Argentina	14527	21.706469
7	Armenia	5188	20.58361
8	Australia	34431	19.780243
9	Austria	35719	53.971355
10	Azerbaijan	8666	55.125755
11	Bahrain	28169	NULL
12	Barbados	17966	47.34396
13	Belarus	13439	54.623608
14	Belgium	33357	80.00875
15	Belize	5812	NULL
16	Bhutan	5293	NULL

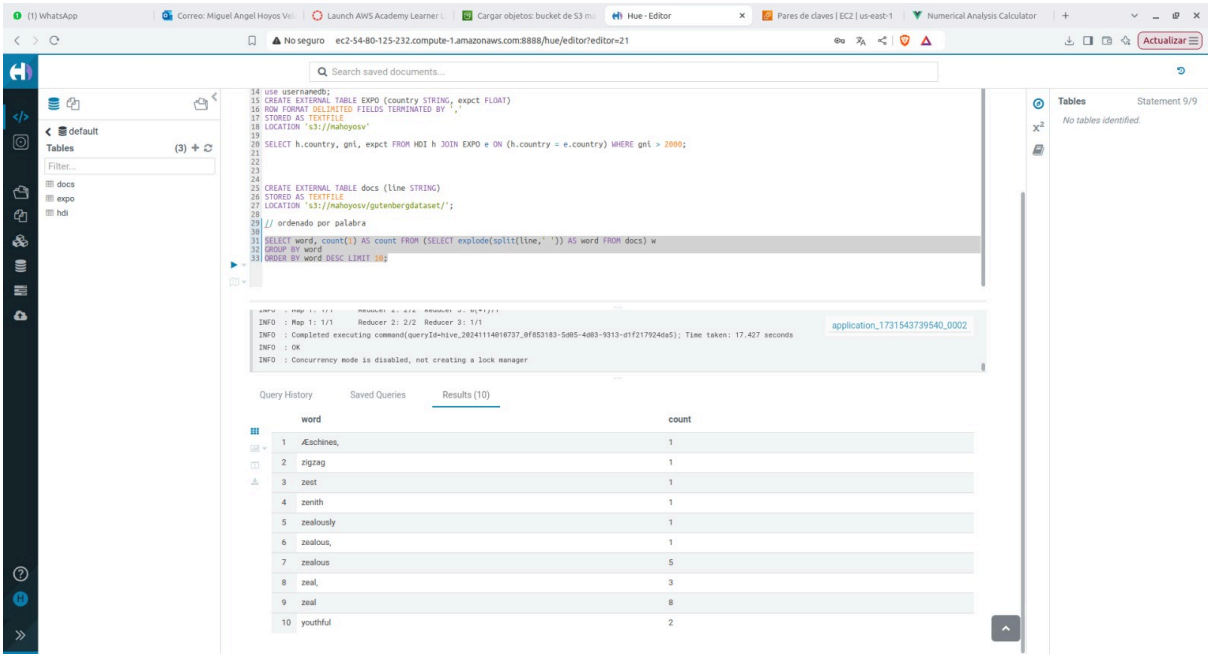
Ahora vemos una consulta, pero en este caso cargando los datos directamente desde S3



Posteriormente se trabajó con el dataset de gutenberg small, para realizar el wordcounter.

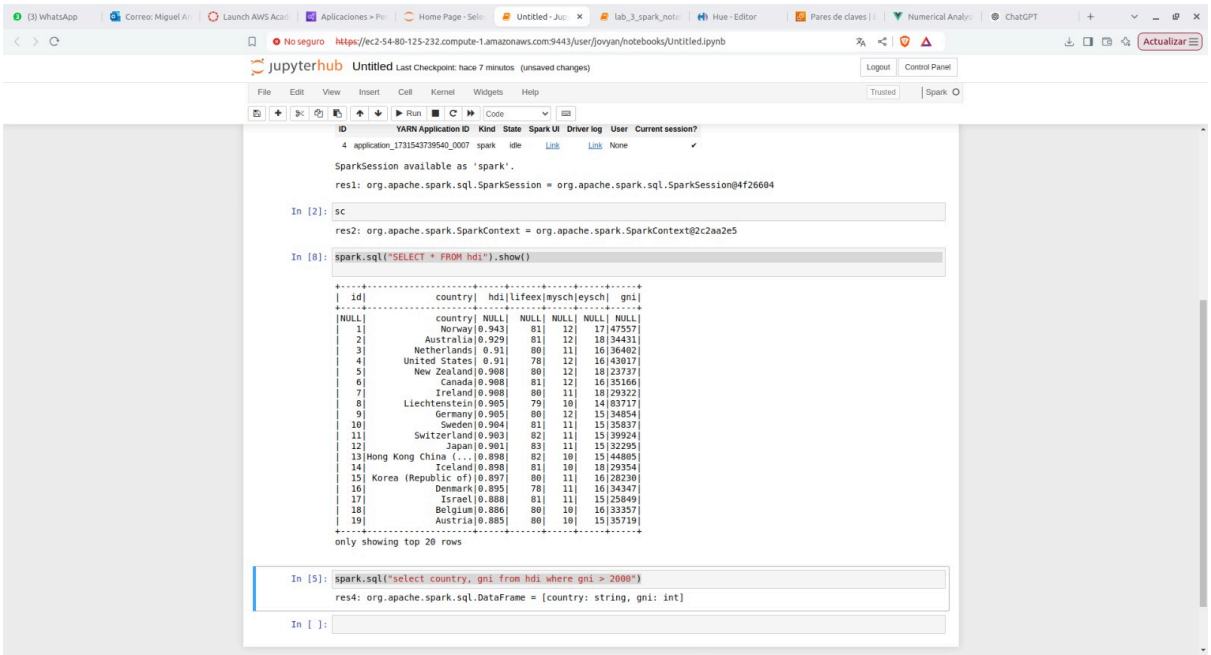


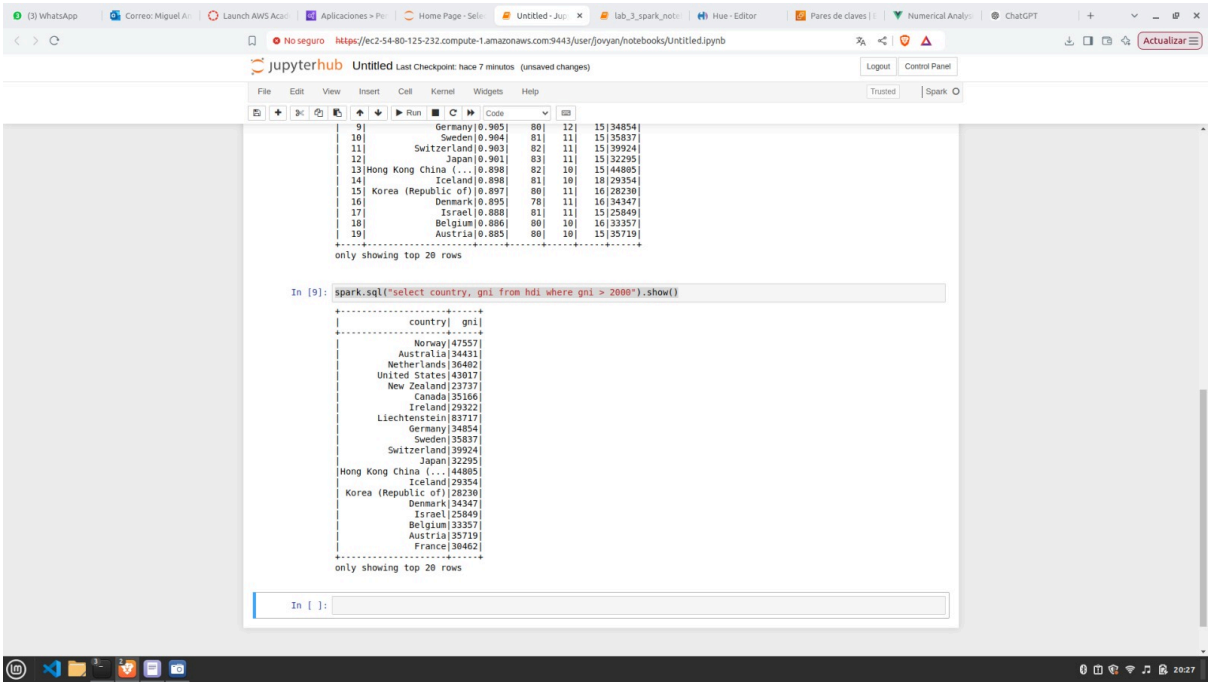
Aquí se aprecia la carga de los archivos en S3, para su posterior utiliización.



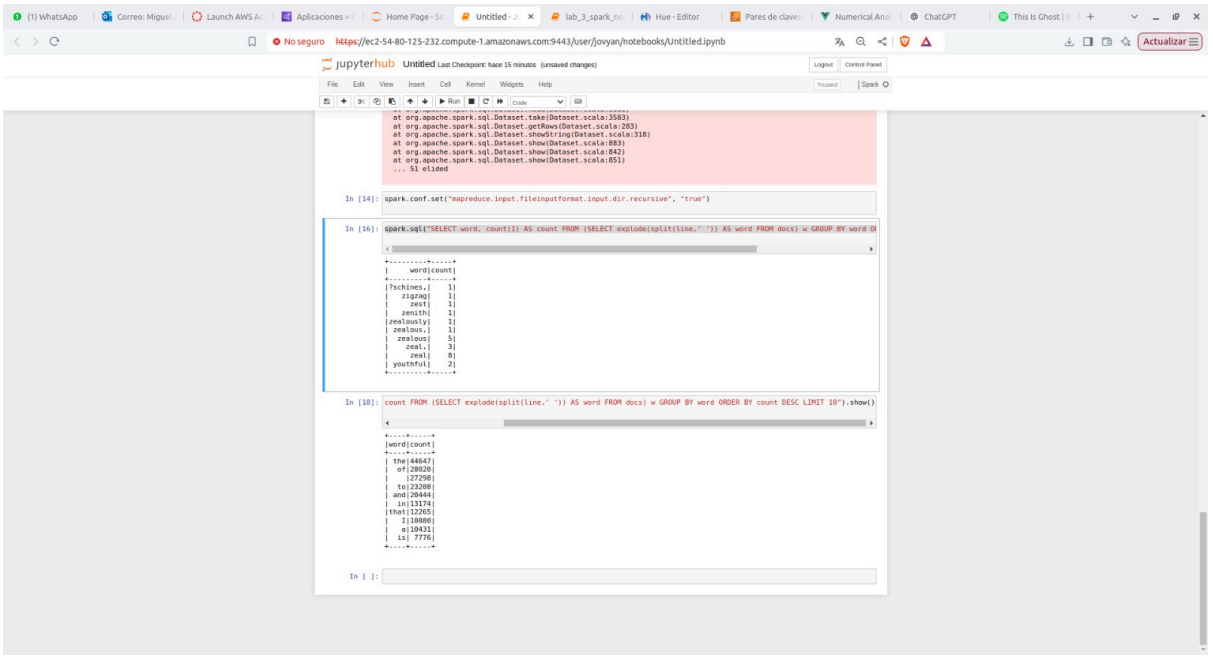
Creación de la tabla y ordenada por palabra.

Adicionalmente, también se trabajó desde jupyterhub con EMR.





Data filtrada con una query simple.



WordCounter. Adicionalmente se adjunta el notebook utilizado, por si se desea replicar.