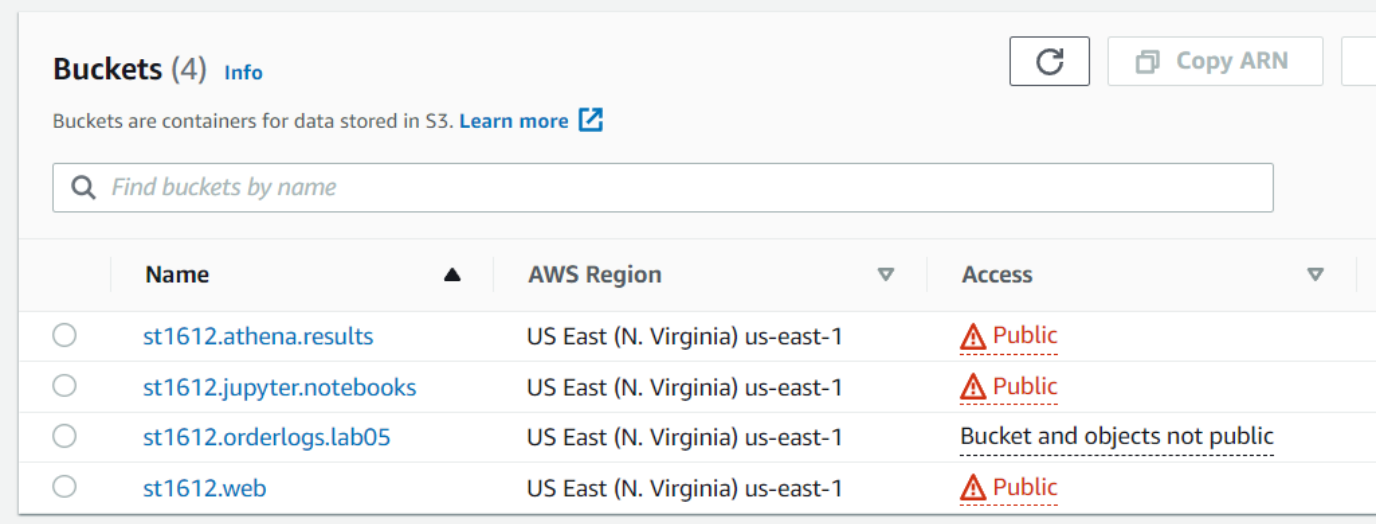
Lab 02_03 - EMR

Lab 02_03 - EMR

The following resources were created with Terraform. The code can be found in the infrastructure folder.

Resources

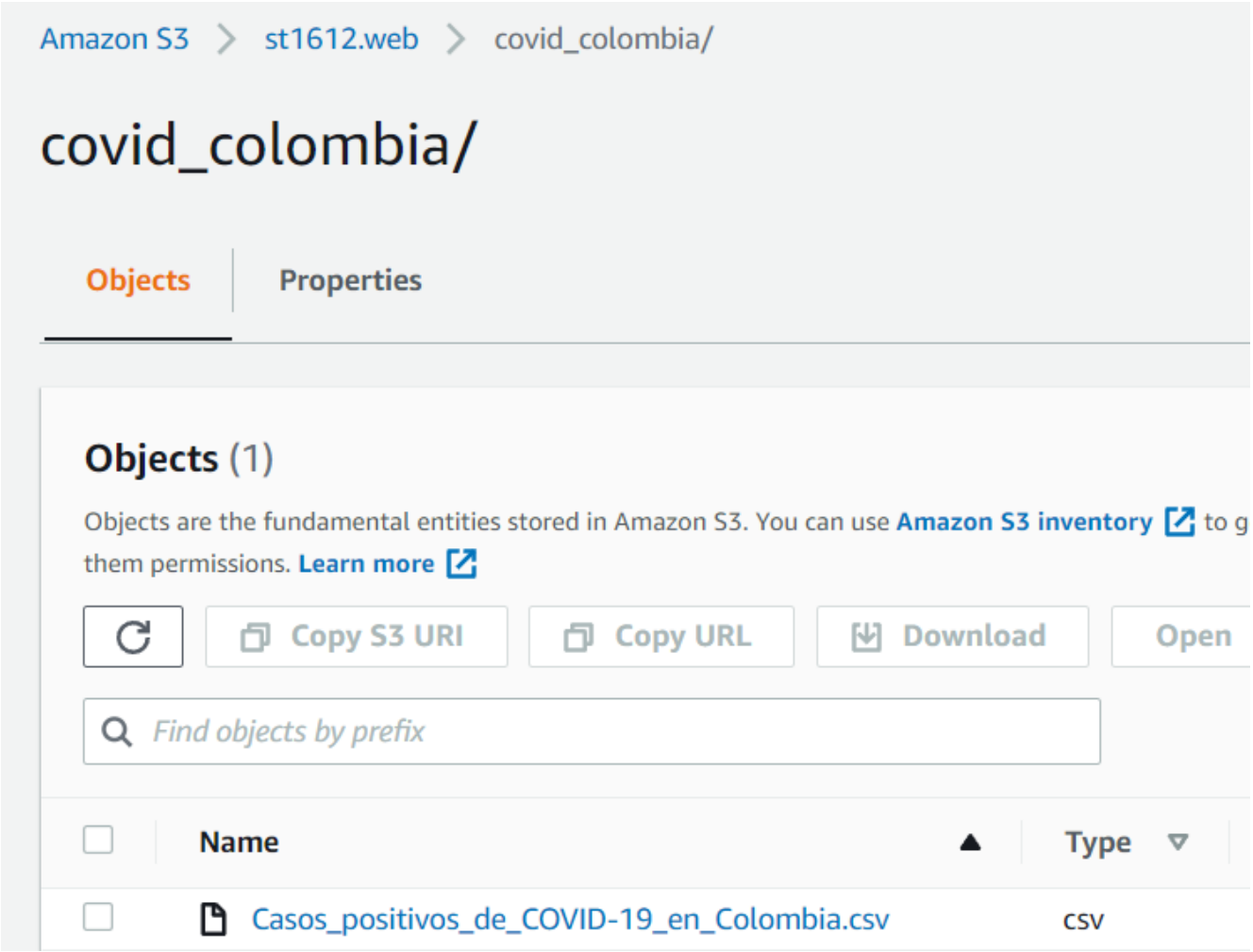
S3 Buckets



Datalake

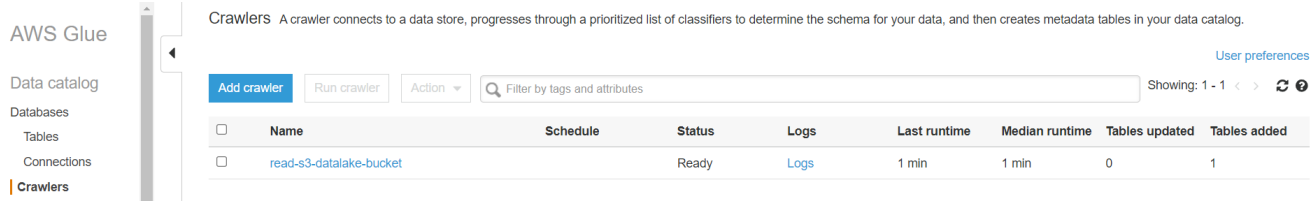
> Taken from: https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data

- Bucket with Colombia Covid-19 data.



Glue

- Crawler



- Schema/Database

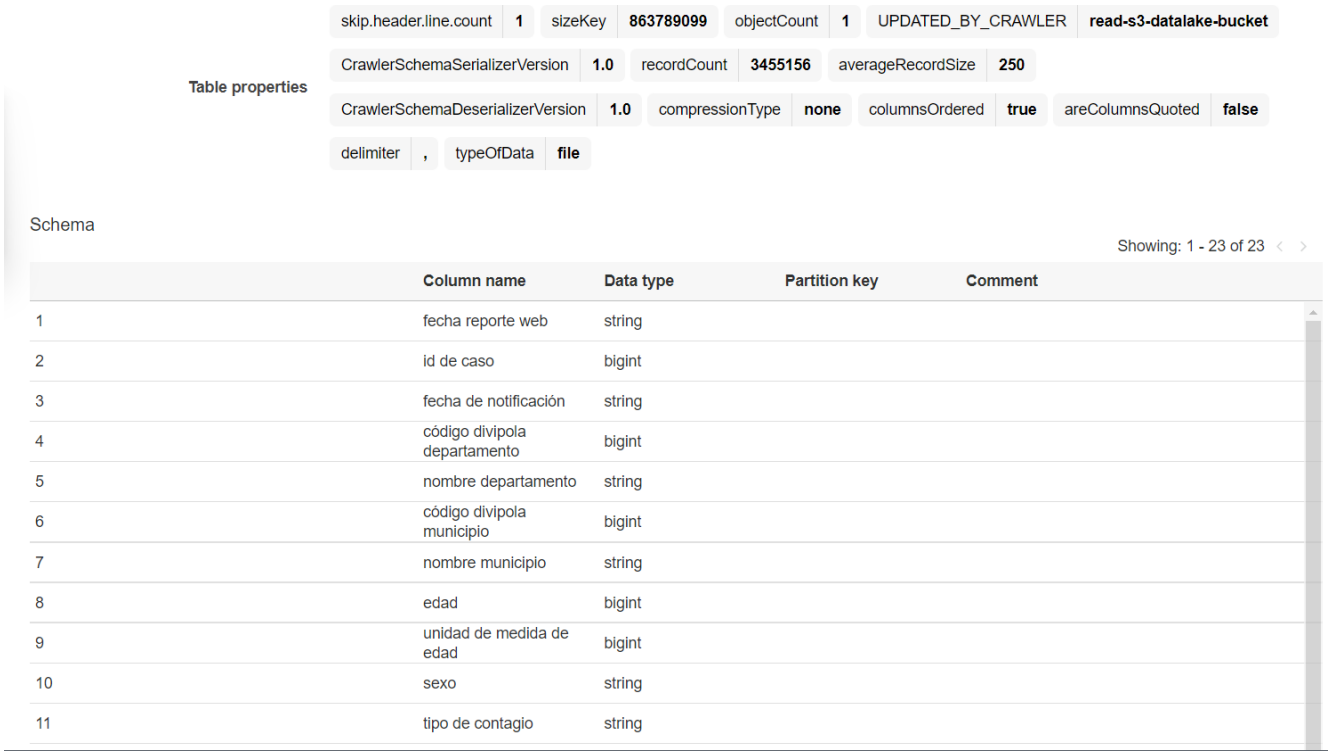


Table properties

skip.header.line.count	1	sizeKey	863789099	objectCount	1	UPDATED_BY_CRAWLER	read-s3-datalake-bucket
CrawlerSchemaSerializerVersion	1.0	recordCount	3455156	averageRecordSize	250		
CrawlerSchemaDeserializerVersion	1.0	compressionType	none	columnsOrdered	true	areColumnsQuoted	false
delimiter	,	typeOfData	file				

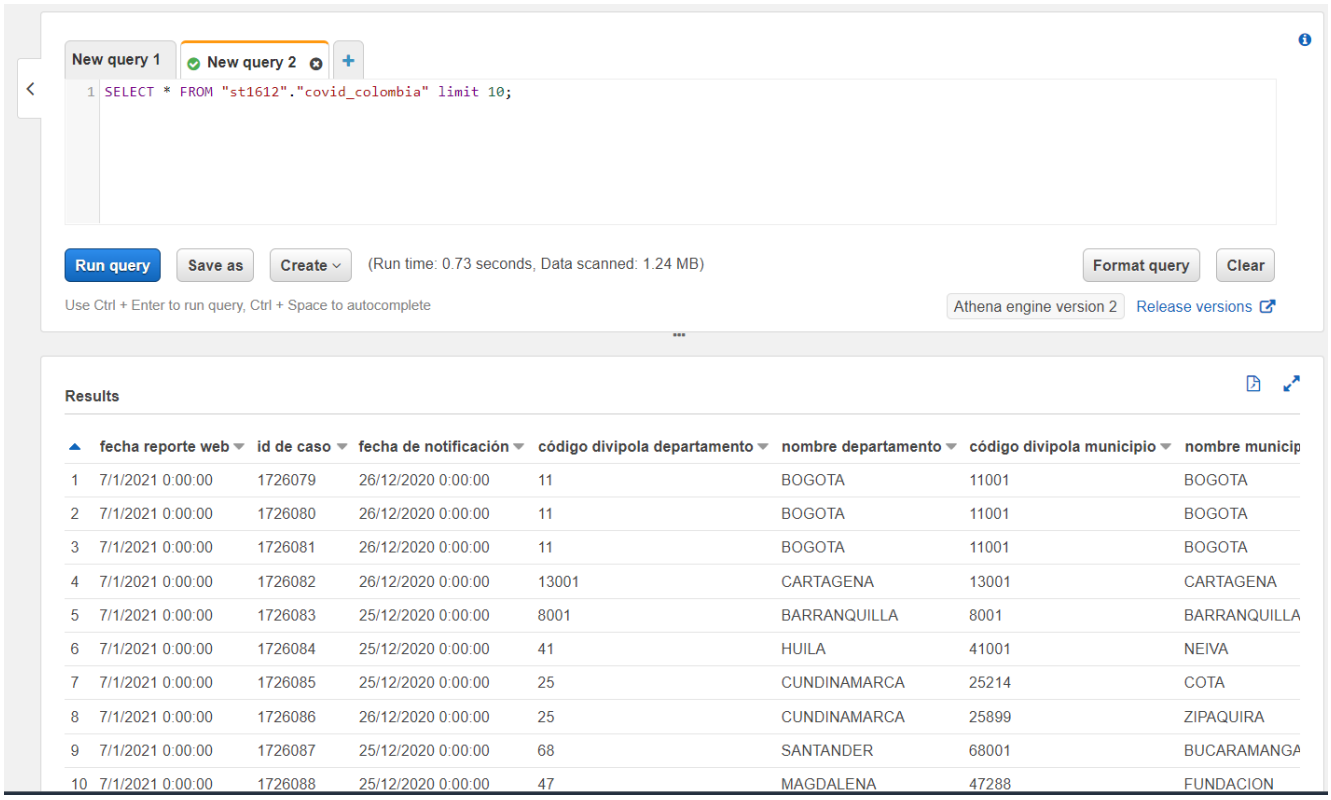
Schema

Showing: 1 - 23 of 23

	Column name	Data type	Partition key	Comment
1	fecha reporte web	string		
2	id de caso	bigint		
3	fecha de notificación	string		
4	código divipola departamento	bigint		
5	nombre departamento	string		
6	código divipola municipio	bigint		
7	nombre municipio	string		
8	edad	bigint		
9	unidad de medida de edad	bigint		
10	sexo	string		
11	tipo de contagio	string		

Athena

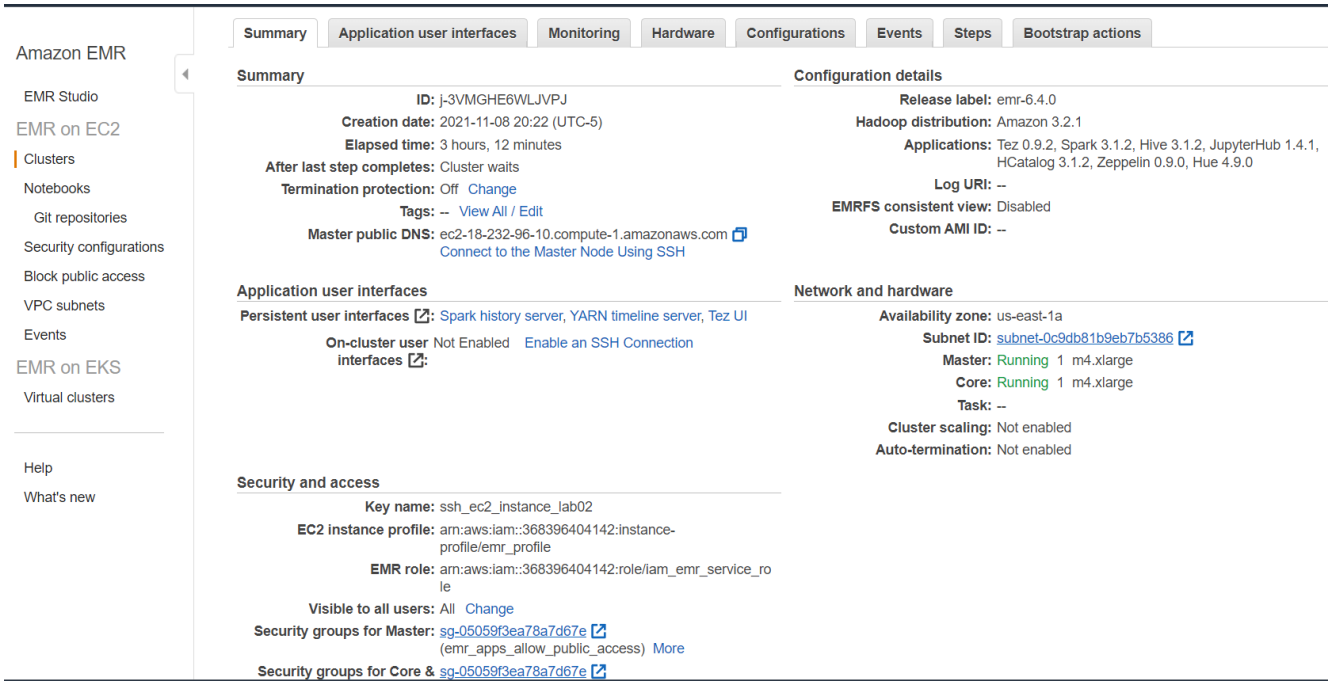
- Query Editor



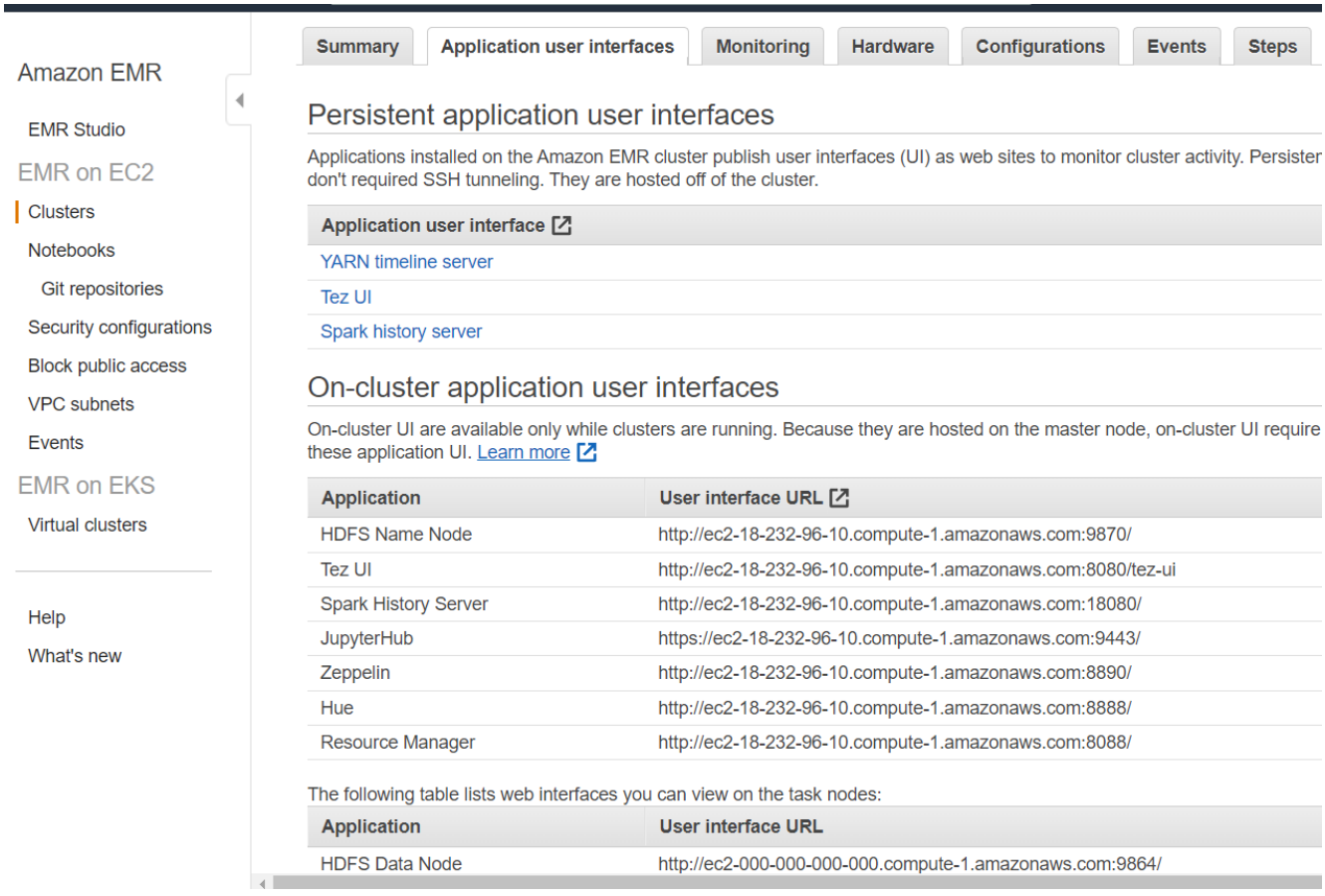
	fecha reporte web	id de caso	fecha de notificación	código divipola departamento	nombre departamento	código divipola municipio	nombre municip
1	7/1/2021 0:00:00	1726079	26/12/2020 0:00:00	11	BOGOTA	11001	BOGOTA
2	7/1/2021 0:00:00	1726080	26/12/2020 0:00:00	11	BOGOTA	11001	BOGOTA
3	7/1/2021 0:00:00	1726081	26/12/2020 0:00:00	11	BOGOTA	11001	BOGOTA
4	7/1/2021 0:00:00	1726082	26/12/2020 0:00:00	13001	CARTAGENA	13001	CARTAGENA
5	7/1/2021 0:00:00	1726083	25/12/2020 0:00:00	8001	BARRANQUILLA	8001	BARRANQUILLA
6	7/1/2021 0:00:00	1726084	25/12/2020 0:00:00	41	HUILA	41001	NEIVA
7	7/1/2021 0:00:00	1726085	25/12/2020 0:00:00	25	CUNDINAMARCA	25214	COTA
8	7/1/2021 0:00:00	1726086	26/12/2020 0:00:00	25	CUNDINAMARCA	25899	ZIPAQUIRA
9	7/1/2021 0:00:00	1726087	25/12/2020 0:00:00	68	SANTANDER	68001	BUCARAMANGA
10	7/1/2021 0:00:00	1726088	25/12/2020 0:00:00	47	MAGDALENA	47288	FUNDACION

EMR

- Cluster

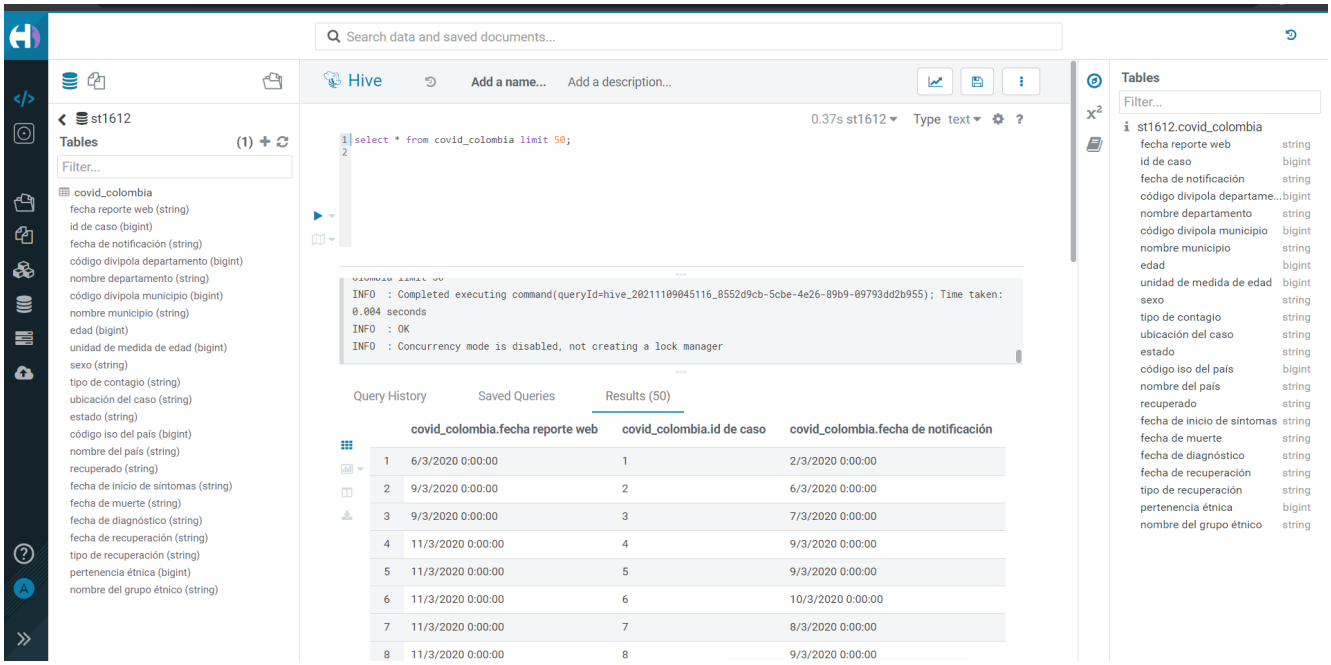


- Deployed applications



Hue

- Query data from Glue Database



Jupyter

- Notebook

```
In [4]: sqlDF = spark.sql("SELECT * FROM raw")
        df.show()

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…

In [5]: sqlDF.columns

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…

['fecha reporte web', 'id de caso', 'fecha de notificación', 'código divipola departamento', 'nombre departament
o', 'código divipola municipio', 'nombre municipio', 'edad', 'unidad de medida de edad', 'sexo', 'tipo de contagi
o', 'ubicación del caso', 'estado', 'código iso del país', 'nombre del país', 'recuperado', 'fecha de inicio de sí
ntomas', 'fecha de muerte', 'fecha de diagnóstico', 'fecha de recuperación', 'tipo de recuperación', 'pertenencia
étnica', 'nombre del grupo étnico', 'partition_0']

In [8]: sqlDF.select("fecha reporte web", "estado").show(5)

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…

+-----------------+------+
|fecha reporte web|estado|
+-----------------+------+
|fecha reporte web|Estado|
| 6/3/2020 0:00:00|  Leve|
| 9/3/2020 0:00:00|  Leve|
| 9/3/2020 0:00:00|  Leve|
|11/3/2020 0:00:00|  Leve|
+-----------------+------+
only showing top 5 rows
```