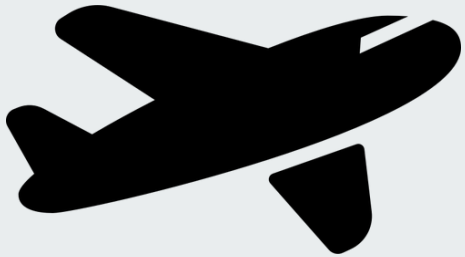
Análisis de datos de OpenSky



Con Spark y Python

Datos de OpenSky

- API
- Impala Shell (Hadoop)

API


```
{
   "time": 1621781094,
   "states": [
       [
           "4952c3",
           "TAP557E ",
           "Portugal",
           1621781094,
           1621781094,
           -9.36,
]
```

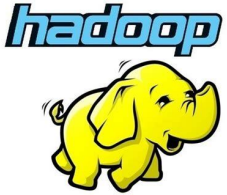

All State Vectors

The following API call can be used to retrieve any state vector of the OpenSky. Please note that rate limits apply for this call (see Limitations). For API calls without rate limitation, see Own State Vectors.

Operation

`GET /states/all`

Impala Shell




```
[hadoop-29:21000] > show tables;
+--------------------------+
| name                     |
+--------------------------+
| acas_data4               |
| allcall_replies_data4    |
| flarm_raw                |
| flights                  |
| flights_data4            |
| identification_data4     |
| operational_status_data4 |
| position_data4           |
| rollcall_replies_data4   |
| sensor_visibility        |
| sensor_visibility_data3  |
| state_vectors            |
| state_vectors_data3      |
| state_vectors_data4      |
| velocity_data4           |
+--------------------------+
[hadoop-29:21000] >
```

SELECT icao24,estarrivalairport,estdepartureairport,callsign FROM flights_data4 WHERE day=unix_timestamp('2021-05-17 00:00:00') ORDER BY rand() LIMIT 1;

```
[hadoop-29:21000] > SELECT icao24,estarrivalairport,estdepartureai
+--------+--------------------+----------------------+----------+
| icao24 | estarrivalairport | estdepartureairport | callsign |
+--------+--------------------+----------------------+----------+
| ac32ed | KLWM               | KHYA                 | N8851K   |
+--------+--------------------+----------------------+----------+
[hadoop-29:21000] >
```

HiveQL

[Cláusula EXPLAIN]

```
> EXPLAIN SELECT COUNT(DISTINCT icao24)
> FROM state_vectors_data4
> WHERE lat<=43.74 AND lat>=35.94
> AND lon<=3.03 AND lon>=-9.39
> AND hour>=unix_timestamp('2021-05-14 01:00:00')
> AND hour<=unix_timestamp('2021-05-16 23:00:00');
```

Importancia de usar datos particionados y buenas queries

```
00:SCAN HDFS [opensky.state_vectors_data4]
   partitions=71/46763 files=71 size=26.01GB
   predicates: lat <= 43.74, lat >= 35.94, lon <= 3.03
```

26**GB**(0,026TB) vs 13**TB**

```
00:SCAN HDFS [opensky.state_vectors_data4]
   partitions=46763/46763 files=47675 size=13.33TB
```

500 veces más datos

Obtención y Filtrado de Datos

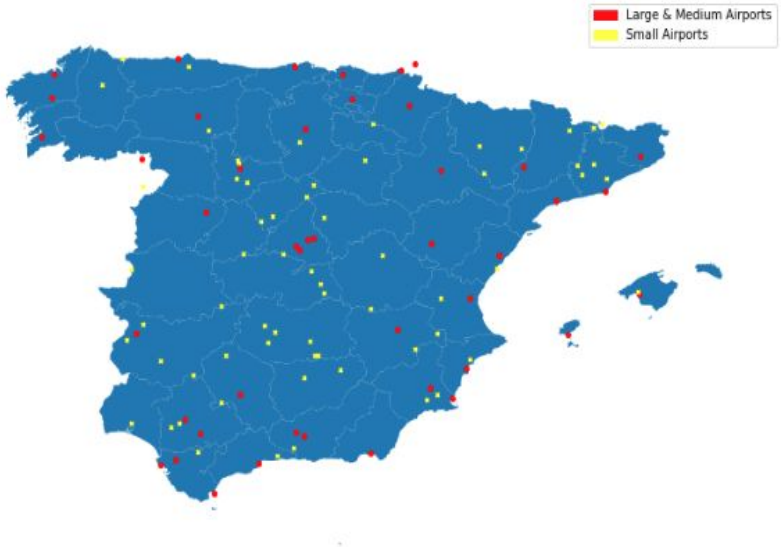
Obtenidos de:

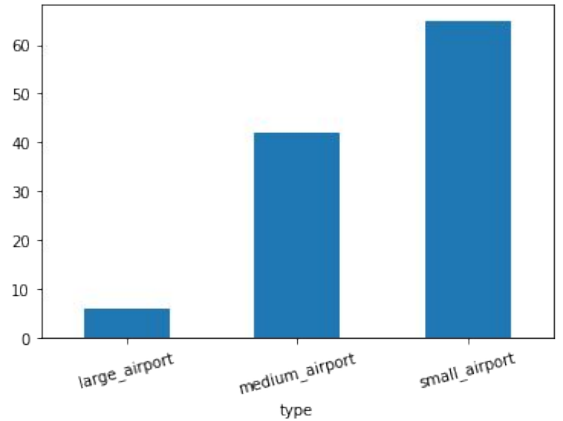
- API sin filtrar las peticiones (poco volumen de datos)
- Hadoop filtrando con queries (mucho volumen de datos)

Filtrado:

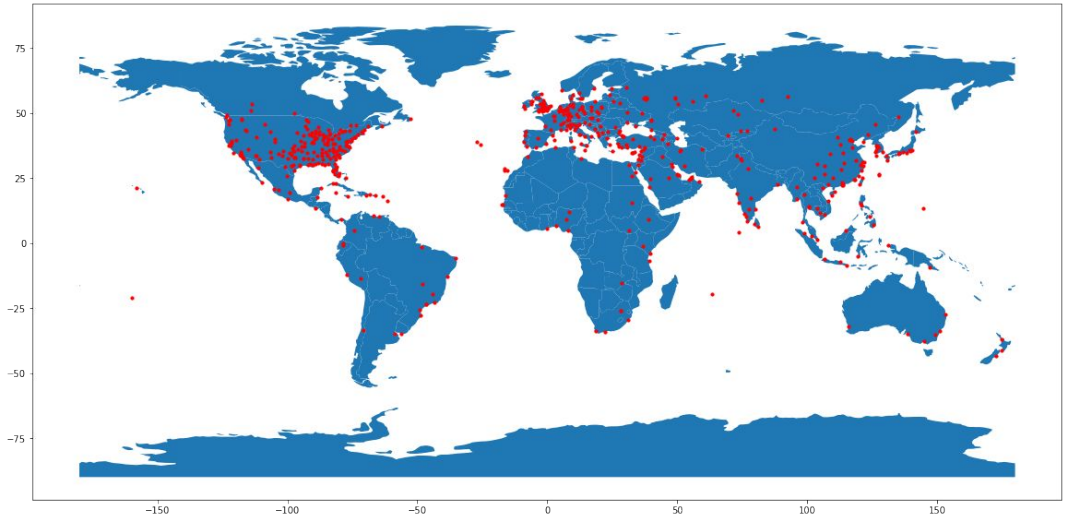
- Python Pandas (Para la API)
- PySpark (Para los CSV obtenidos de Hadoop)

Aeropuertos (España)





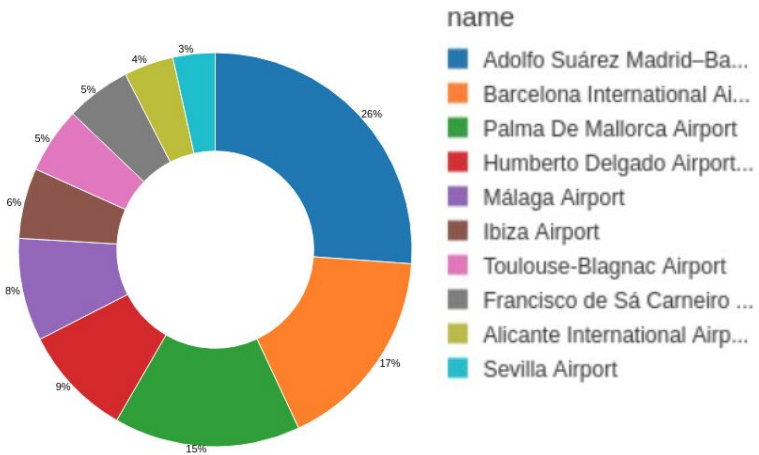
Aeropuertos (Global)

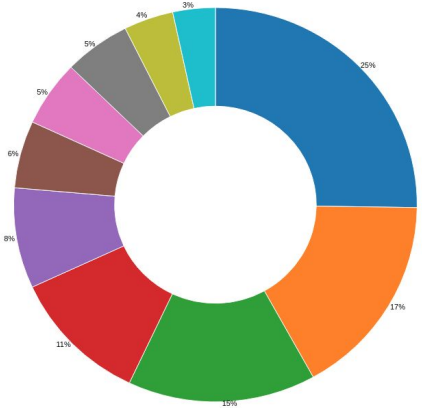


Flight Data


count
1200
1100
1000
900
800
700
May 1
2021
May 3
May 5
May 7
May 9
May 11
May 13
May 15
May 17
May 19
window.end
count
30
25
20
15
Apr 30
2021
May 1
May 2
May 3
May 4
May 5
May 6
May 7
May 8
May 9
May 10
May 11
May 12
May 13
May 14
May 15
May 16
May 17
window.end

Salidas/Llegadas

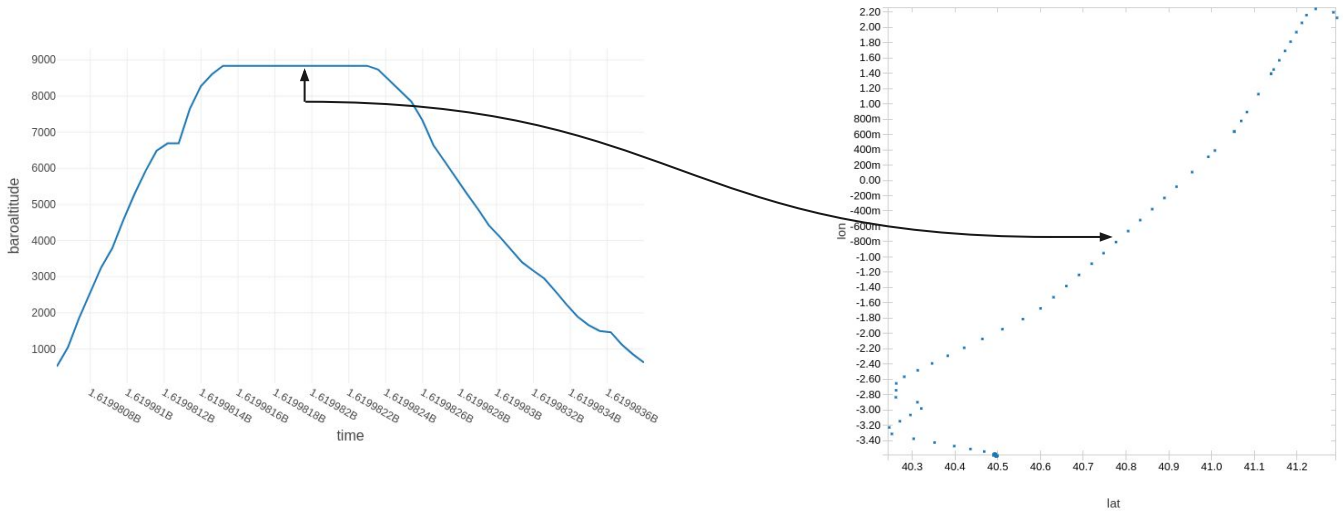




Salidas (Madrid)

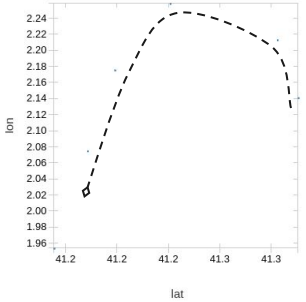
num_flights	name	departure	arrival
234	Palma De Mallorca Airport	LEMD	LEPA
232	Barcelona International Airport	LEMD	LEBL
97	Humberto Delgado Airport (Lisbon Portela Airport)	LEMD	LPPT
94	Ibiza Airport	LEMD	LEIB
75	Málaga Airport	LEMD	LEMG

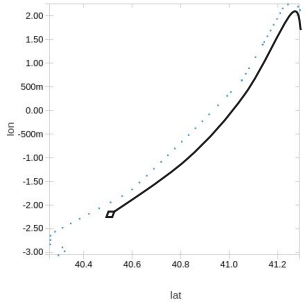
Visualización datos de Vuelo (IB2031)

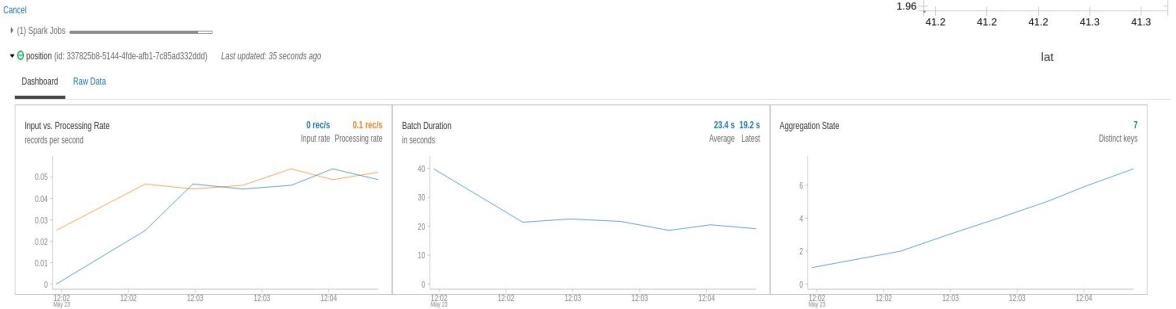


Streaming con Spark

- Datos del mismo vuelo (BCN-MAD)
- Emulado con ficheros con distinta fecha de creación







Colaboración con OpenSky

- Obtención de datos con un receptor propio (en una Raspberry Pi 30€ aprox)
- Petición para instalar un receptor en la Facultad de Informática?

- Get A Receiver