Data Engineering Project

Building a pipeline of Catalonian cell network API data

Covering Coverage

Background & Motivation

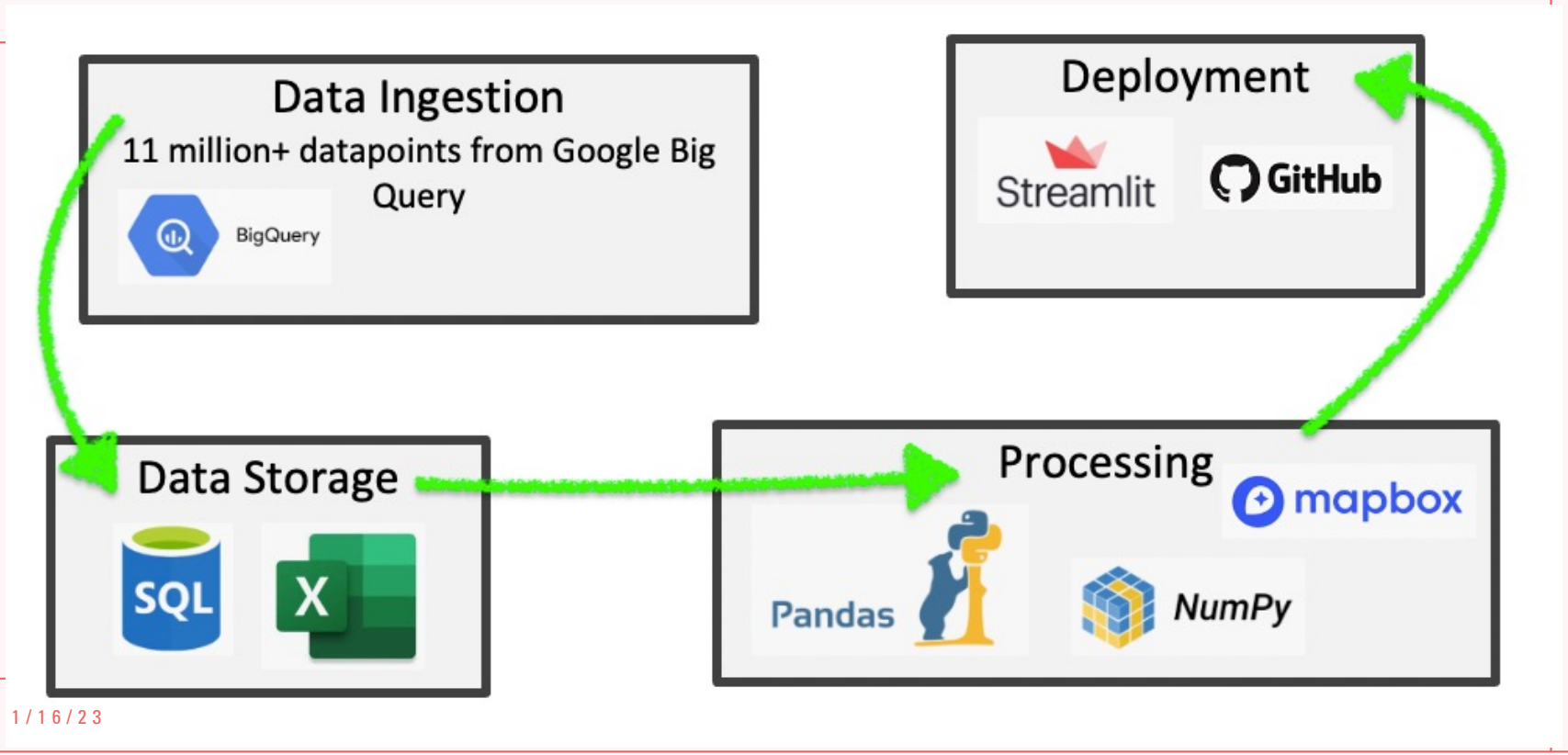
- Popular region with lots of tourism
- Majority of population is connected to the network
- Need a site to monitor outages, performance, and changes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10683528 entries, 0 to 1068352
Data columns (total 21 columns):
 #   Column           Dtype
---  ------           -----
 0   Date             object
 1   Hour             object
 2   lat              float64
 3   lon              float64
 4   Signal           int64
 5   network          object
 6   Operator         object
 7   status           int64
 8   description      object
 9   net              object
 10  Speed            float64
 11  Satellites       int64
 12  Precision        int64
 13  Provider         object
 14  activity         object
 15  postal_code      float64
 16  town_name        object
 17  position_geom    object
 18  Cleaned_Network  object
 19  Top_Networks     object
 20  Time Bucket      object
dtypes: float64(4), int64(4), object(13)
memory usage: 1.7+ GB
```

Data & Methods

- Catalonian Cell Coverage from Google BigQuery's API
- 11,000,000+ rows,
- each representing a carrier's coverage status at a specific date, time, and place.
- The categorical groups examined will be for:
  1. Network operators
  2. Activity

Pipeline


Data Ingestion
11 million+ datapoints from Google Big Query
BigQuery
Deployment
Streamlit
GitHub
Data Storage
SQL
X
Processing
mapbox
Pandas
NumPy

Results

- Streamlit App Link:
- **http://192.168.1.239:8501**

Conclusion & Future Work

- Try Spark to process data more quickly.
- Find API with live updates & create connection.