Conner McNicholas

Data Engineering Career Track

Open-ended Capstone Step 2: Project Proposal

Problem statement formation:

This project aims to create a publicly accessible, user friendly cloud hosted database for exploration of historical severe weather events in the United States.

Context:

Currently, the only comprehensive dataset of severe weather events in the U.S. exists as a complex repository of tersely named compressed csv files, hidden away behind multiple layers of subdirectories in a rarely visited corner of the National Centers for Environmental Information's dated web portal. To download, decompress, and convert the data into a queryable format from which insights could be gained would require technical savvy far beyond the capability of most non-developers.

This project will provide tools from which data will be presented in a clean, inviting, user-friendly interface from which users can explore the dataset by filtering both rows and columns, submit queries, as well as export results locally in formats of their preference (sql, csv, json, etc). Later versions might offer tools to explore the data with interactive widgets offering geospatial and time-series views.

Criteria for success:

All severe weather data from 2010 until 2021 is available in singular database
Users can filter the data by row and column
Users can export the data to save locally

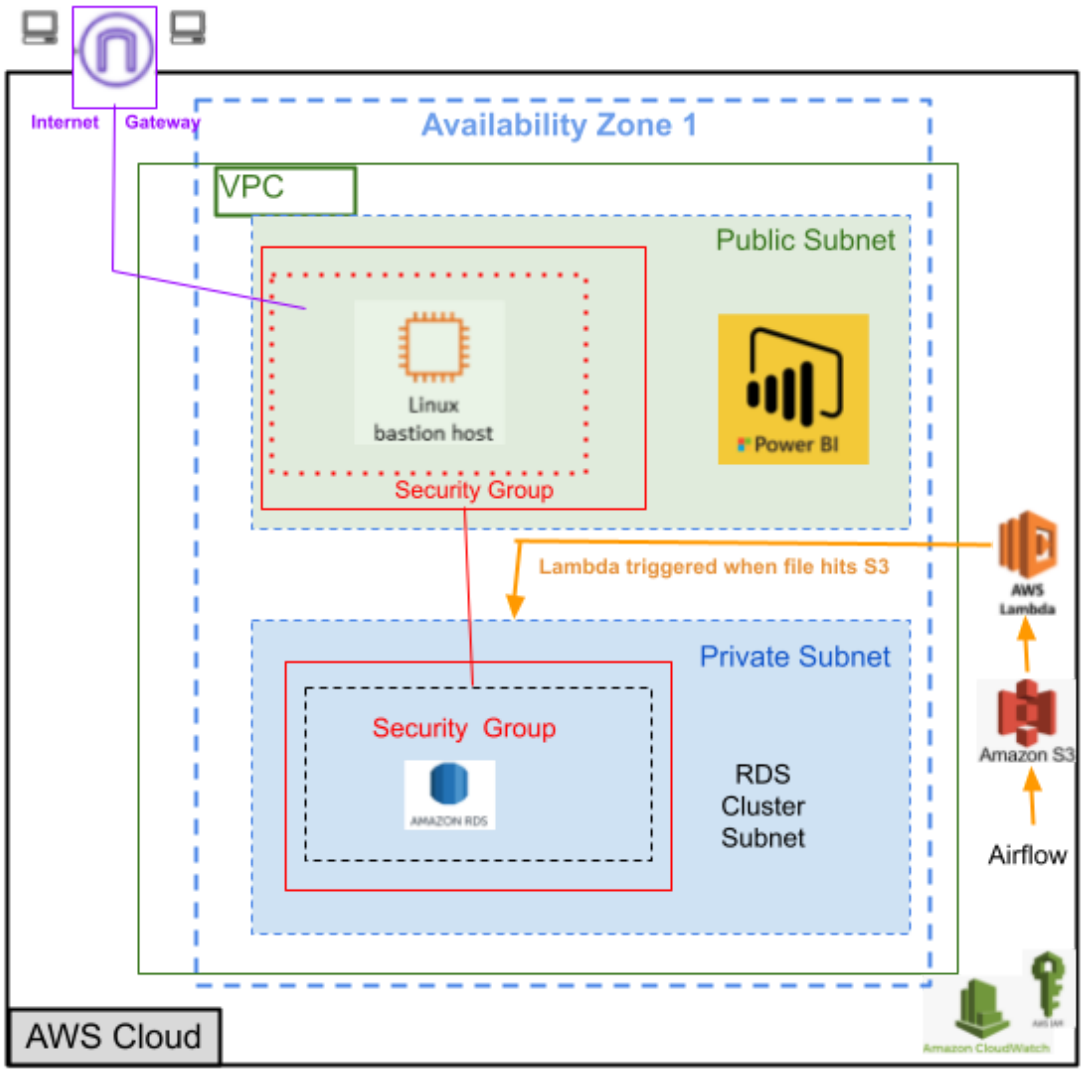
Scope of solution space:

Database with toolbar to offer filtering, exportation, and previewing of the dataset.

Data source(s):

https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/

Proposed architecture for the solution and rationale behind it:



Choice of technology for the solution and rationale:

- AWS
  - S3
  - Amazon RDS
  - Microsoft Power BI
  - Cloudtrail
  - IAM
  - Lambda
  - EC2
    - Python3
      - psycopg2
      - Sqlalchemy
    - cron
    - rsync
    - tar
    - Airflow