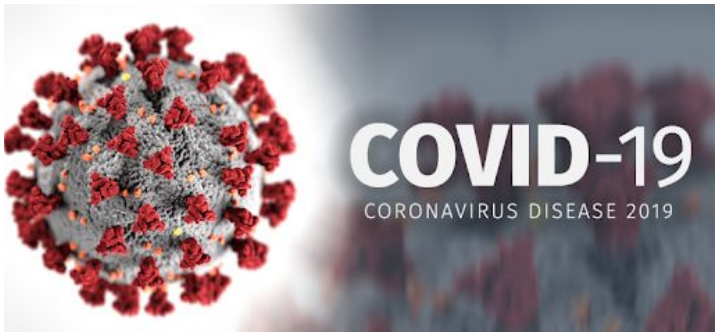
AWS기반
COVID 19 데이터
분석과 시각화

발표 날짜	2021.09.27
팀명	N조
팀원 이름	김현수 나혜주 레이첼 간

Table of Contents

Overview



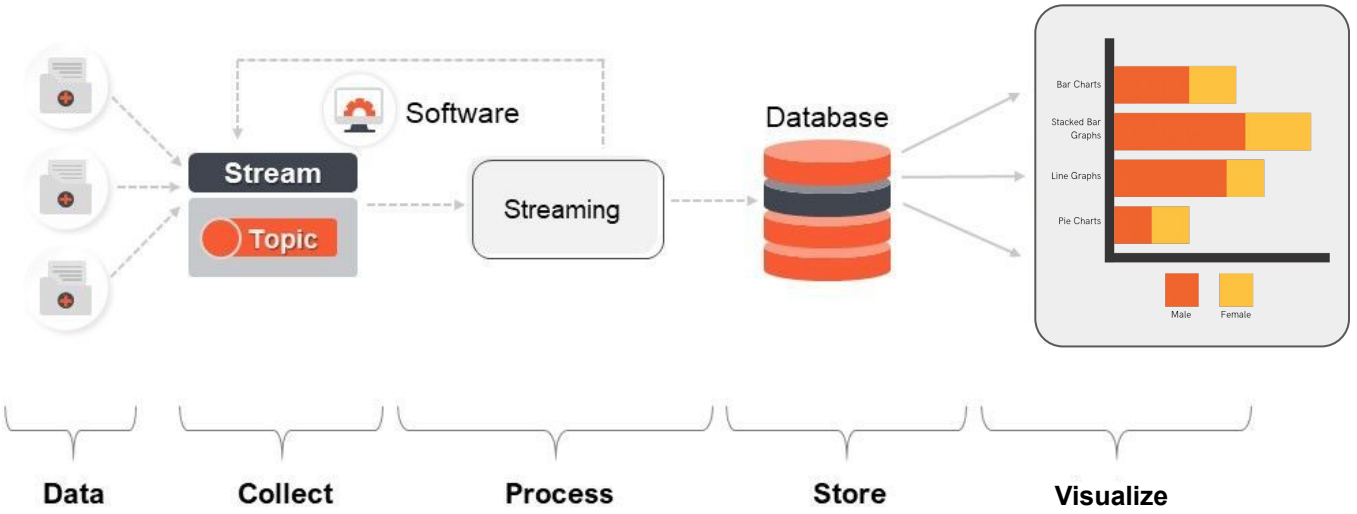
프로젝트 내용

- COVID 19 관련 다양한 공개 데이터와 소셜 미디어 자료를 활용해 데이터 수집과 저장,
- 데이터 전처리와 분석, 그리고 시각화까지 데이터 분석 플랫폼을 구축한다.

교육/훈련 효과

- 데이터 분석의 전반적인 과정과 그에 필요한 지식 습득, 데이터 분석 오픈 소스 활용 경험,
- 데이터 시각화를 통한 인사이트 도출

Goal



COVID-19의 다양한 데이터 셋을 활용해 데이터 수집과 저장, 데이터 전처리와 분석, 그리고 시각화까지 데이터 분석 플랫폼을 구축한다.

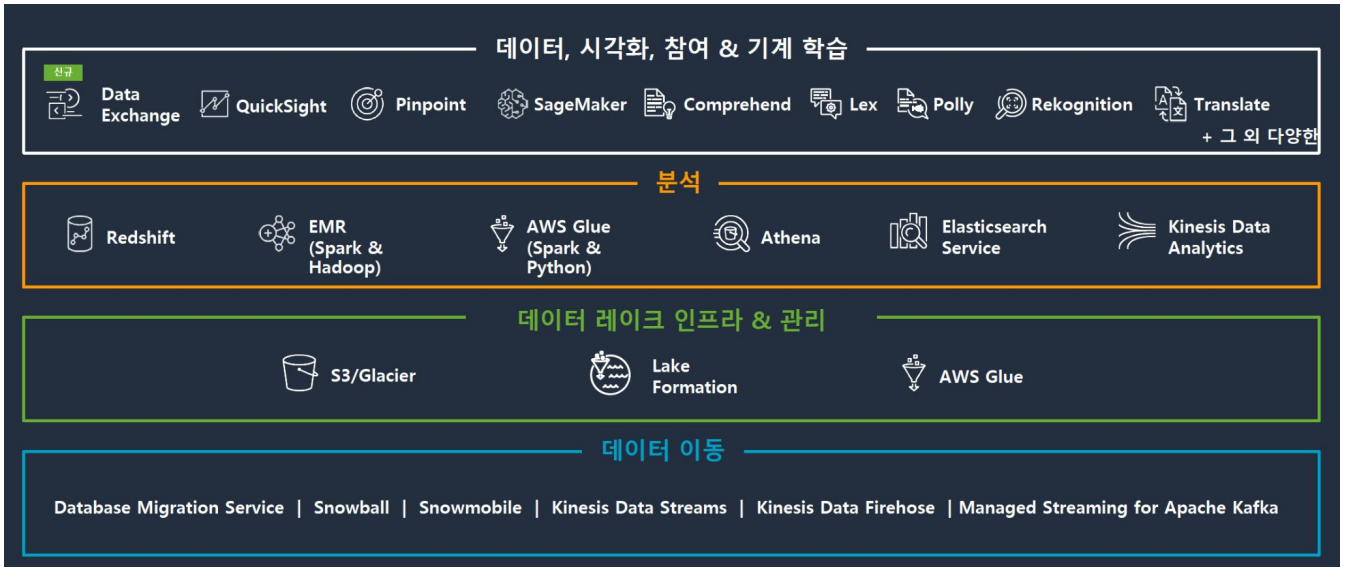
- 필요에 따라 AWS의 다양한 데이터 분석 서비스를 사용
- 사용자가 원하는 형태의 데이터 시각화 구현

구현 내용

- 데이터 수집기 : COVID-19 및 관련 자료 수집기
- 데이터 처리기: 원하는 정보로 가공 가능하도록 만들기
- 대시 보드: 다양한 COVID-19 트렌드 및 관련 소셜 미디어 데이터 연동

Requirement

- AWS 서비스와 기타 API, CLI를 사용해 데이터를 수집 및 이관하는 아키텍처를 3개 이상 구현
- 4개 이상의 데이터셋을 통해 적절한 데이터 전처리 과정을 구현
- 서로 다른 4개 이상의 AWS 서비스를 사용해 데이터 분석 아키텍처를 구현
- 원하는 형태의 대시보드 기능 구현이 가능하며 분석에 대한 기능 항목 3개 이상을 설정하고 구현



Approach 주제탐색

주제 선정 시 고려 사항

데이터 셋이

(1) 코로나 전과 후 각각에 대해 동일 조건으로 있어야 함

(2) 최근 데이터까지도 있어야 함

(3) 데이터의 주기가 1년 미만

데이터로 해보고 싶은 것

(1) 코로나가 경제, 환경, 생활 등 다양한 분야에 어떤 영향을 미쳤는지 알아보자

(2) 코로나 사태와 이전 전염병 사태, 나아가 경제 위기 등등을 비교해보자.

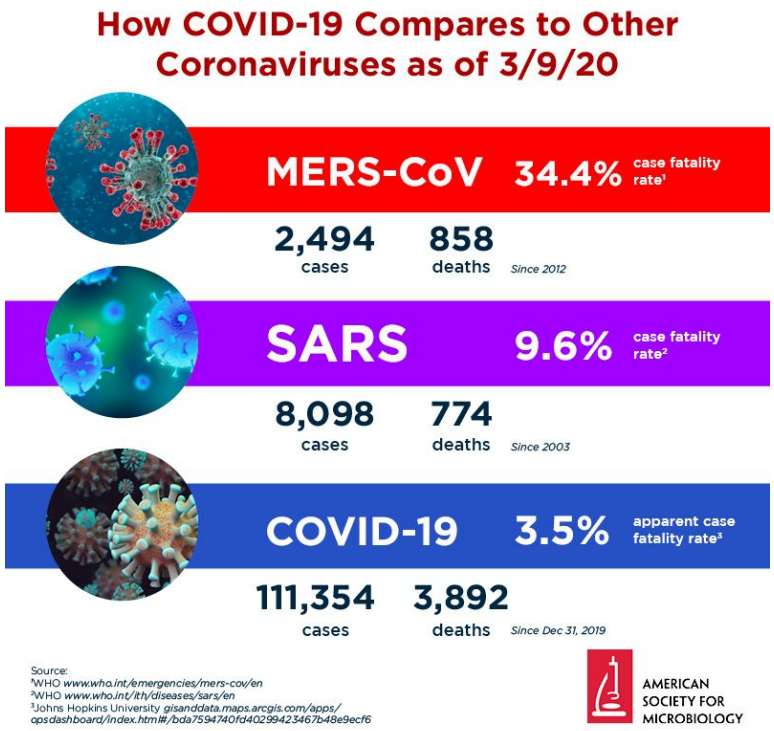
Approach 주제선정

(1)주가

(2)실업률

(3)대기오염도

- 다른 질병, 경제적 위기 vs 코로나



Basic Specifications

Lake Formation

- 3개의 주제(대기오염, 실업률, 주가)를 기반으로 데이터 수집
- 데이터 레이크를 생성하고 메타데이터를 저장하기

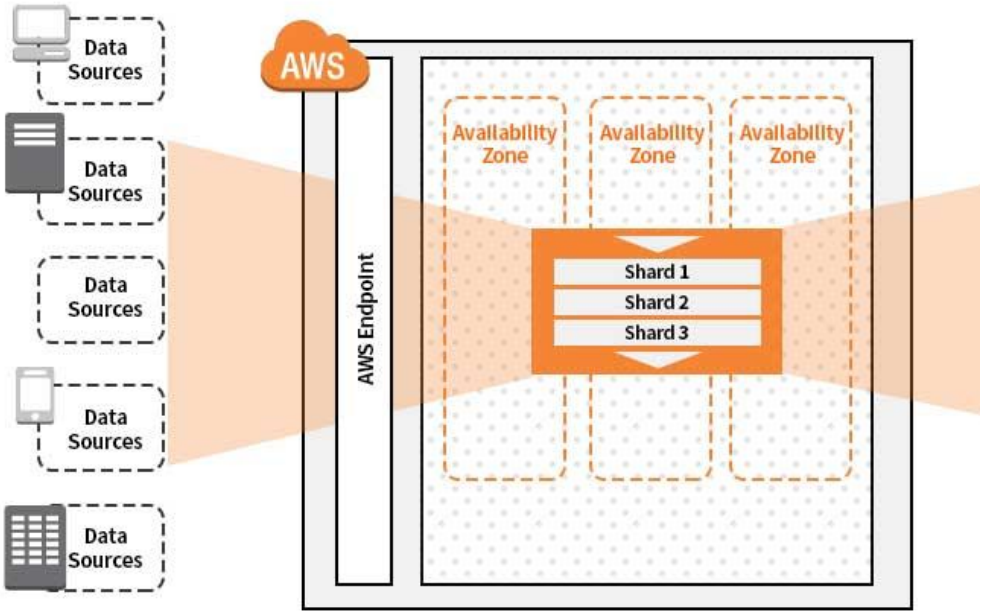
데이터 시각화

- AWS 툴을 활용해 데이터를 분석해 다른 세계적 질병과 경제적 위기를 COVID-19 데이터랑 비교하기
- AWS 서비스의 시각화 기능을 통해 Interactive 대시보드를 구축하고 Frontend 애플리케이션과 원활하게 결합해 웹사이트를 배포하기

Specifications (1)데이터 수집 및 이동

Challenge

- Data Source에 따른 효율적인 데이터 수집 및 이동 방법 확인
- Static Data vs Streaming Data의 수집 및 이동
- AWS CLI, SDK로 데이터 수집 및 이동을 자동화 시키는 방법 고안



Specifications (1)데이터 수집 및 이동

Data Source

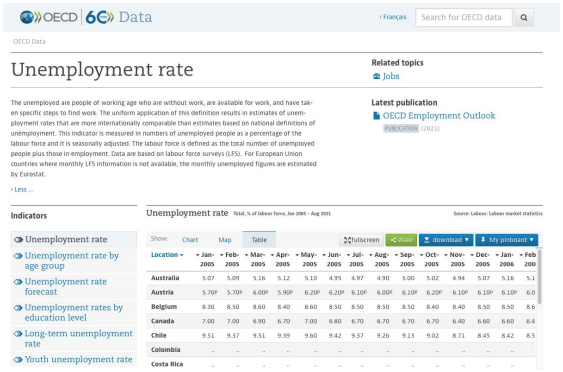
- 주식시장 지표
  - 시기 : 2000년~현재
  - NASDAQ / S&P 500 / 다우 / Nikkei225 / Shanghai / CSI 300 / FTSE 100 / KOSPI 200


나스닥 100 선물 과거 데이터

기간: 일간 | 데이터 다운로드 | 2021/09/07 - 2021/10/07

날짜	종가	오픈	고가	저가	거래량	변동 %
2021년 10월 07일	14,900.50	14,808.25	14,910.25	14,799.25	83.00K	0.96%
2021년 10월 06일	14,759.00	14,662.00	14,770.00	14,422.50	11.18K	0.71%
2021년 10월 05일	14,655.25	14,493.75	14,754.75	14,410.50	574.64K	1.33%
2021년 10월 04일	14,462.25	14,785.00	14,835.00	14,367.75	9.91K	-2.03%
2021년 10월 01일	14,761.75	14,702.25	14,818.00	14,537.00	657.09K	0.54%
2021년 09월 30일	14,682.50	14,769.75	14,883.50	14,670.25	9.14K	-0.39%
2021년 09월 29일	14,739.75	14,769.25	14,926.00	14,711.00	7.19K	-0.17%
2021년 09월 28일	14,764.75	15,172.25	15,192.25	14,740.75	7.73K	-2.83%
2021년 09월 27일	15,194.75	15,319.00	15,399.25	15,084.00	5.70K	-0.81%
2021년 09월 24일	15,318.75	15,300.25	15,334.25	15,176.75	470.51K	0.10%
2021년 09월 23일	15,303.50	15,178.25	15,347.00	15,175.25	4.53K	0.92%
2021년 09월 22일	15,163.50	14,992.75	15,227.00	14,930.50	595.84K	0.93%
2021년 09월 21일	15,024.00	14,992.00	15,163.25	14,963.25	657.00K	0.10%
2021년 09월 20일	15,009.50	15,284.00	15,338.25	14,807.50	834.32K	-3.08%
2021년 09월 17일	15,486.67	15,502.75	15,547.00	15,474.00	7.47K	-0.20%
2021년 09월 16일	15,517.75	15,516.50	15,542.00	15,369.50	105.43K	0.09%
2021년 09월 15일	15,504.00	15,395.50	15,521.75	15,315.25	156.77K	0.76%
2021년 09월 14일	15,387.00	15,452.25	15,530.75	15,351.25	237.82K	-0.32%
2021년 09월 13일	15,436.75	15,452.50	15,565.00	15,350.25	363.28K	-0.07%
2021년 09월 11일	15,447.75	15,447.80	15,447.80	15,447.75	396.89K	0.04%
2021년 09월 10일	15,441.50	15,561.00	15,668.00	15,433.50	402.82K	-0.75%
2021년 09월 09일	15,558.75	15,610.00	15,675.75	15,545.75	531.29K	-0.39%
2021년 09월 08일	15,620.00	15,671.00	15,691.00	15,523.75	559.40K	-0.35%

- 실업률
  - 시기 : 1955~2021년 8월
  - OECD회원국


OECD Data

Unemployment rate

The unemployed are people of working age who are without work, are available for work, and have taken specific steps to find work. The uniform application of this definition results in estimates of unemployment rates that are more internationally comparable than estimates based on national definitions of unemployment. This indicator is measured in numbers of unemployed people as a percentage of the labour force and it is seasonally adjusted. The labour force is defined as the total number of unemployed people plus those in employment. Data are based on labour force surveys (LFS). For European Union countries where monthly LFS information is not available, the monthly unemployed figures are estimated by Eurostat.

Related topics: Jobs

Latest publication: OECD Employment Outlook (PUBLICATION (2021))

Indicators
- Unemployment rate
- Unemployment rate by age group
- Unemployment rate forecast
- Unemployment rates by education level
- Long-term unemployment rate
- Youth unemployment rate

Unemployment rate — Total, % of labour force, Jan 2005 - Aug 2021 — Source: Labour: Labour market statistics

Show: Chart | Map | Table — Fullscreen | share | download | My pinboard

Location	Jan-2005	Feb-2005	Mar-2005	Apr-2005	May-2005	Jun-2005	Jul-2005	Aug-2005	Sep-2005	Oct-2005	Nov-2005	Dec-2005	Jan-2006	Feb-200
Australia	5.07	5.09	5.16	5.12	5.10	4.95	4.97	4.90	5.00	5.02	4.94	5.07	5.16	5.1
Austria	5.70P	5.70P	6.00P	5.90P	6.20P	6.20P	6.10P	6.00P	6.10P	6.20P	6.10P	6.10P	6.10P	6.0
Belgium	8.30	8.50	8.60	8.40	8.60	8.50	8.50	8.50	8.50	8.40	8.40	8.50	8.50	8.6
Canada	7.00	7.00	6.90	6.70	7.00	6.80	6.70	6.70	6.70	6.70	6.40	6.60	6.60	6.4
Chile	9.51	9.37	9.51	9.39	9.60	9.42	9.37	9.26	9.13	9.02	8.71	8.45	8.42	8.5
Colombia	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Costa Rica	..	..	..	..	..	..	..	..	..	..	..	..	..	..

- 대기오염도
  - 시기: 2013년~현재 (매일 업데이트)
  - scope: 전세계 (132 나라)


Air Quality Open Data Platform

Worldwide COVID-19 dataset

With the COVID-19 spreading out all over the world, the World Air Quality Index project team saw a surge in requests for global data covering the whole world map. As a result, the WAQI project is now providing a new dedicated data-set, updated 3 times a day, and covering about 380 major cities in the world, from January 2020 until now.

The data for each major cities is based on the average (median) of several stations. The data set provides `min`, `max`, `median` and `standard deviation` for each of the air pollutant species ($PM_{2.5}$,PM_{10}, Ozone ...) as well as meteorological data (Wind, Temperature, ...). All air pollutant species are converted to the US EPA standard (i.e. no raw concentrations). All dates are UTC based. The `count` column is the number of samples used for calculating the `median` and `standard deviation`.

Download the COVID-19 Air Quality worldwide 2021 CSV dataset (Size: 86MB - generated on 2021-10-03T08:37:17+01:00)

Due to the high demand, data for prior years can also be downloaded:

2020Q1 (28 MB)	2020Q2 (31 MB)	2020Q3 (31 MB)	2020Q4 (30 MB)
2019Q1 (25 MB)	2019Q2 (27 MB)	2019Q3 (30 MB)	2019Q4 (30 MB)
2018H1 (27 MB)	2017H1 (21 MB)	2016H1 (20 MB)	2015H1 (19 MB)

Signed as hscornelia@snu.ac.kr (signout)

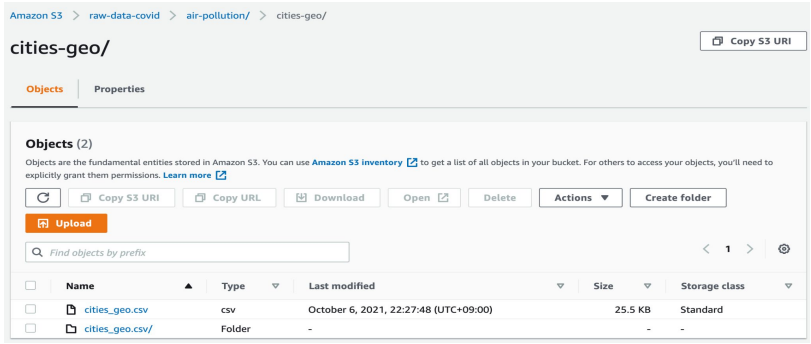
terms of service accepted on 2021-10-05T03:09:59+09:00

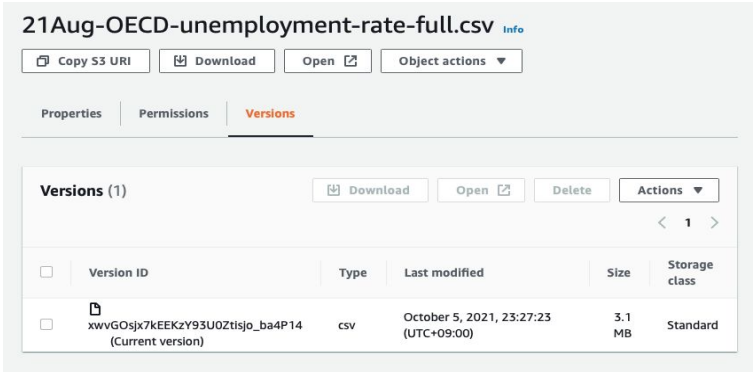
Specifications (1)데이터 수집 및 이동

Static Data Source → S3

AWS console – S3 raw bucket

- 버켓에 저장 후 S3 데이터 레이크 스토리지 위치 등록





boto3 library (python의 AWS sdk)

```python
import boto3
import requests
import json

# Retrieve data from source (api) and save to local file
def get_file_api_local_save(api, file=None):
    '''
    Example:
        response = requests.get("https://aqicn.org/data-platform/covid19/airquality-covid19-cities.json")
        if file is None:
            file = open('save_local.csv')
        data = response.text
        parsed = json.loads(data)
        file.write(parsed)
        file.close()
        return file
    '''
    # return saved file pointer
    pass


# Send file to s3 bucket
def send_file_to_s3(filePath, bucket):
    '''
        client = boto3.client(
            's3',
            aws_access_key_id = 'AK********XC',
            aws_secret_access_key = 'RONA2yOUws7QZ**********0Lv+DnHy',
            region_name = 'ap-northeast-2'
        )
        clientResponse = client.Bucket(bucket).upload_file(filepath)
    '''
    # return true if data is file in filePath is safely saved in s3Path else false
    pass
```

Specifications (2)데이터 처리

Challenge

- File format, Table Schema 등 다양한 형태의 데이터를 SQL, Python을 이용해 여러 형태로 가공
- 필요없는 Column 삭제, 중복 데이터 제거의 자동화 등 데이터 처리를 위한 Script 작성
  → vanilla python, spark (python or scala) 활용

Tools

- Python으로 데이터 가공 후 S3로 이동 → 속도 문제 때문에 작은 데이터 프로세싱에 사용
- Pyspark → 큰 데이터 프로세싱에 사용
- AWS Glue의 Crawling을 이용한 Metadata 추출

```
# Format cities_geo
```

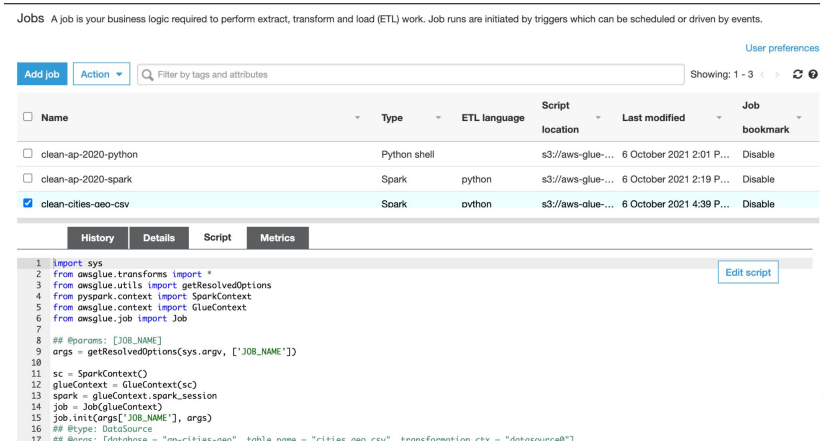
```
import requests
import json
import pandas as pd
```

```
response = requests.get("https://aqicn.org/data-platform/covid19/airquality-covid19-cities.json")
text = json.loads(response.text)
data = text['data']
```

```
df = pd.DataFrame(data)
places = df['Place']
places_list = places.values.tolist()

df = pd.DataFrame(places_list)
df = df.rename(columns={"name": "city"})
df = df.apply(lambda col: [col[0][0], col[0][1], col[1], col[2], col[3], col[4]], axis=1, result_type="expand")
df = df.rename(mapper={0: "latitude", 1: "longitude", 2:"feature", 3: "city", 4: "country", 5:"population"}, axis=
```

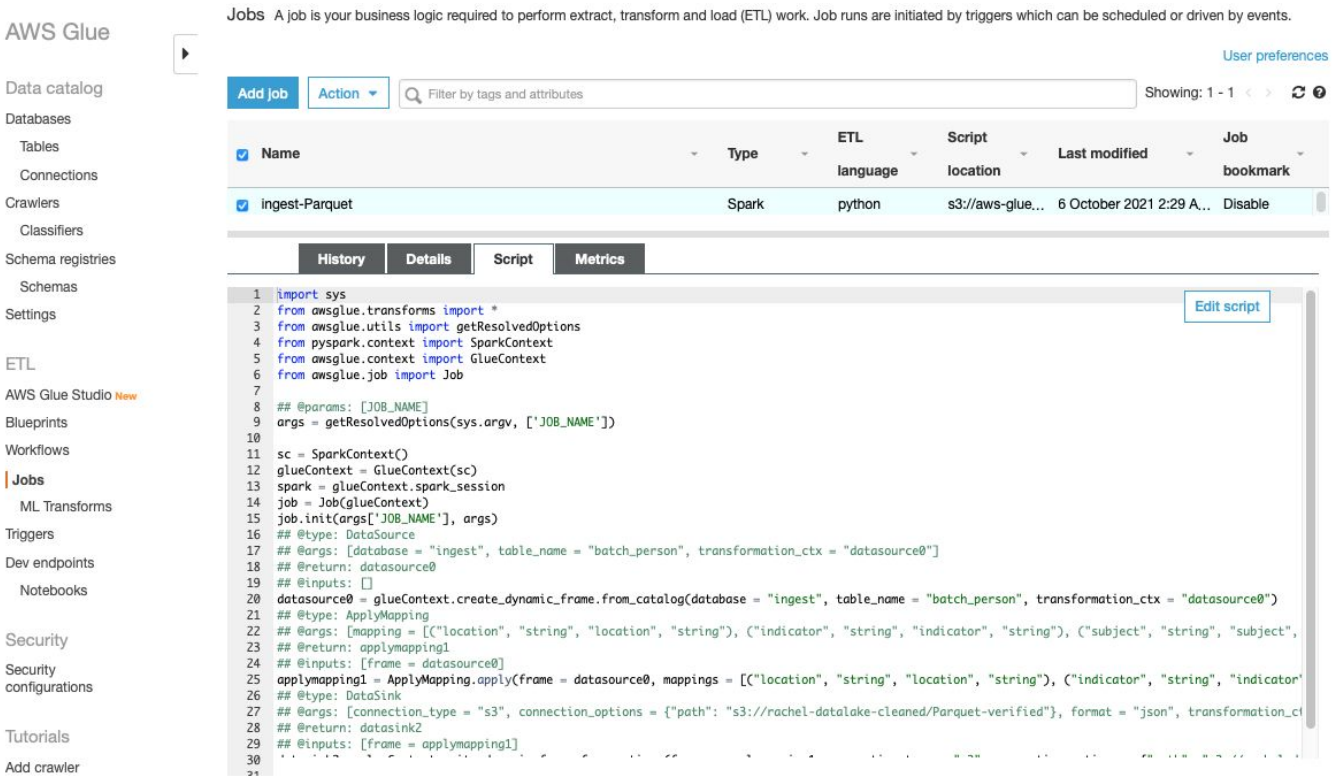
```
df.to_csv('cities_geo.csv', header=False, index=False)
```



Specifications (2)데이터 처리

Glue Crawler

- AWS Lake Formation → Crawlers → Ingest database & Crawl Data
- Crawlers를 만들어 메타 데이터 캡처 자동화 → 카탈로그 엔터티 생성



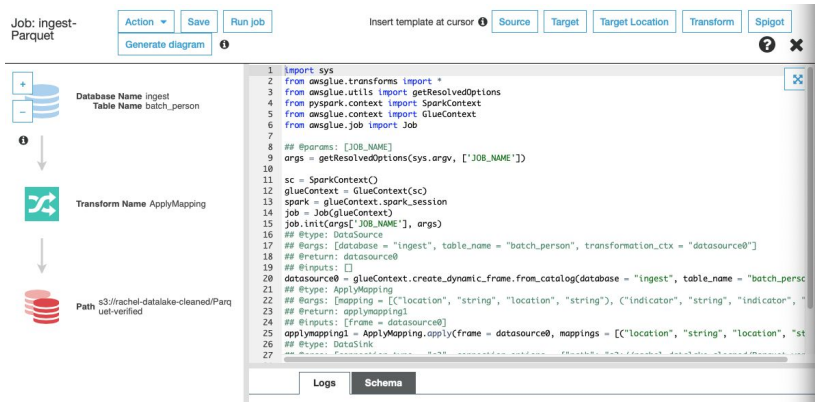
Specifications (2)데이터 처리

AWS Glue

- AWS Glue용 IAM 역할 생성

Glue ETL Job

- raw data를 파싱 → 효율적인 columnar data인 'parquet' 포맷으로 변환 → cleaned S3 bucket에 저장
- 데이터베이스 테이블 생성 확인



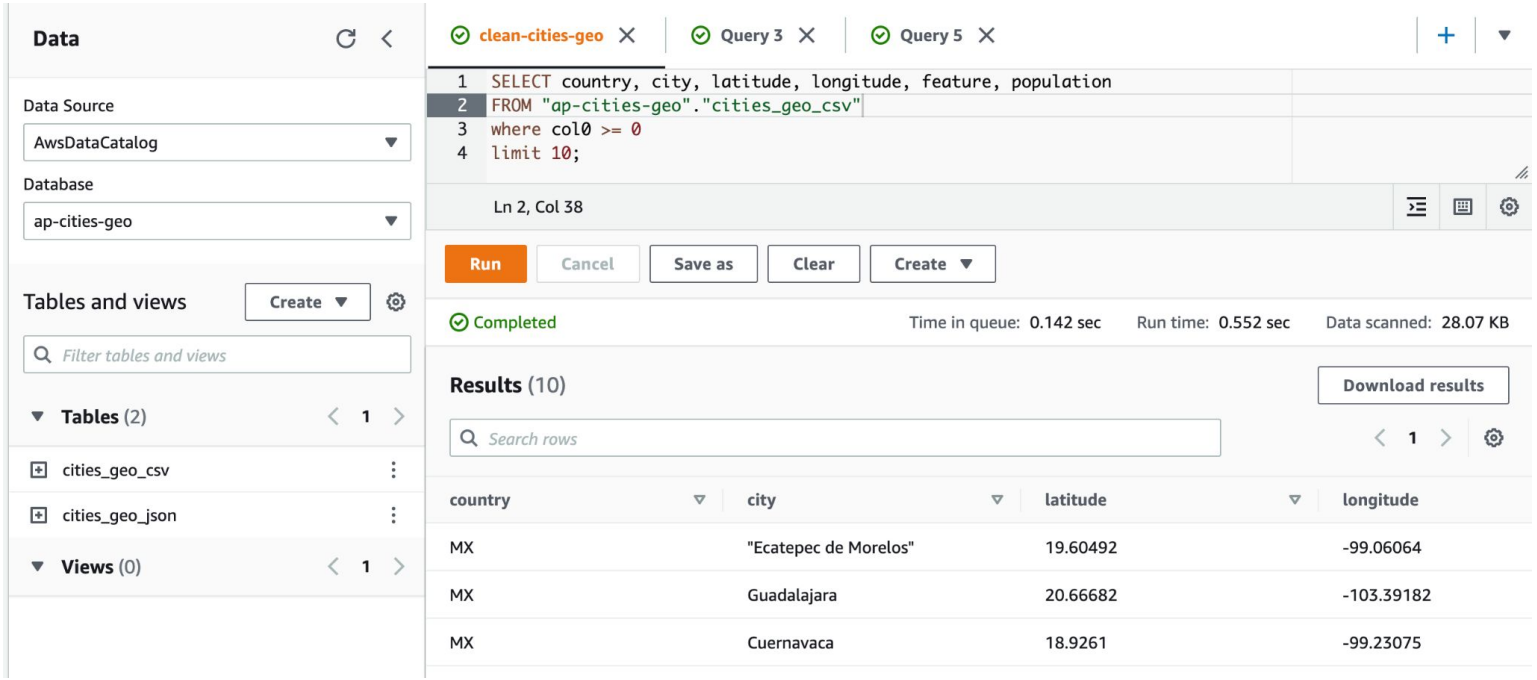
Schema

	Column name	Data type	Partition key	Comment
1	location	string		
2	indicator	string		
3	subject	string		
4	measure	string		
5	frequency	string		
6	time	string		
7	value	double		
8	flag codes	string		

Specifications (3)데이터 분석

Challenge

- AWS EMR (Hadoop, Spark 등)을 이용해서 대규모 데이터 처리→ 서로 다른 결과 도출
- AI 툴인 AWS Sagemaker를 활용한 데이터 상관관계 및 미래 예측
- SQL 사용 → Amazon Athena를 이용해 S3에 저장된 데이터를 query해 원하는 종류와 기준의 데이터만 추출
- 자주 사용되는 작업을 효과적으로 처리할 수 있는 쿼리 생성 & 저장

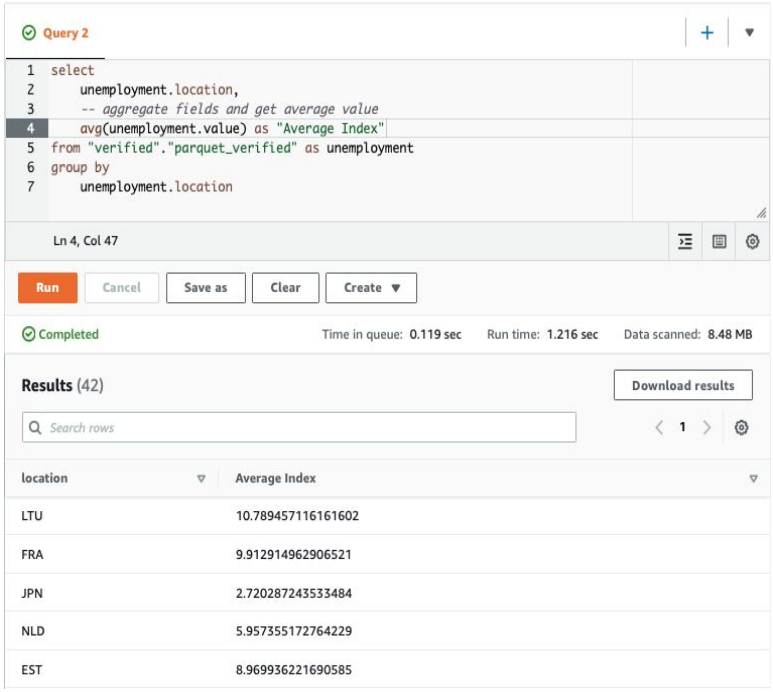


Specifications (3)데이터 분석

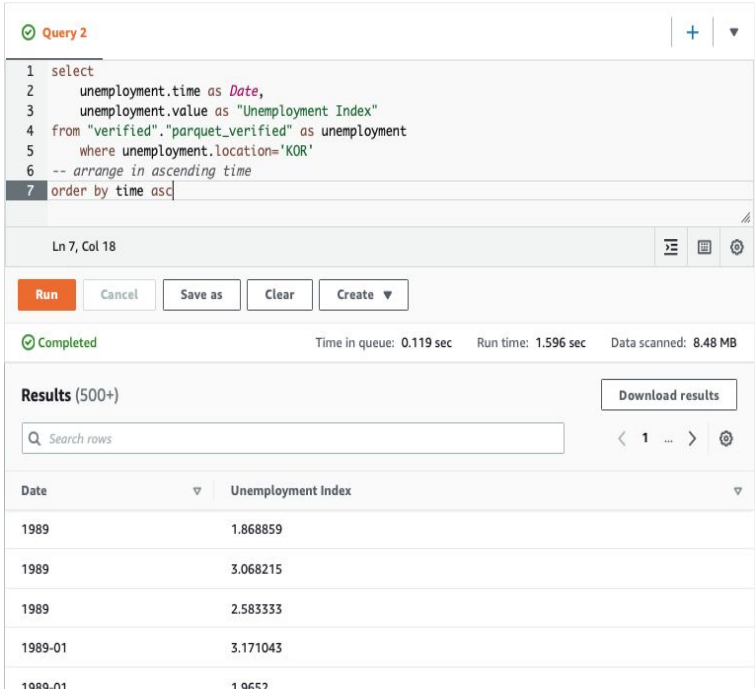
Amazon Athena SQL Query

- SQL문을 사용하여 쿼리 진행
- 쿼리 결과를 S3 curated 버켓 AthenaResults/ 폴더에서 저장

SQL 함수를 이용해 특정 테이블 열을 합계



SQL ORDER BY 데이터 테이블 조회 정렬



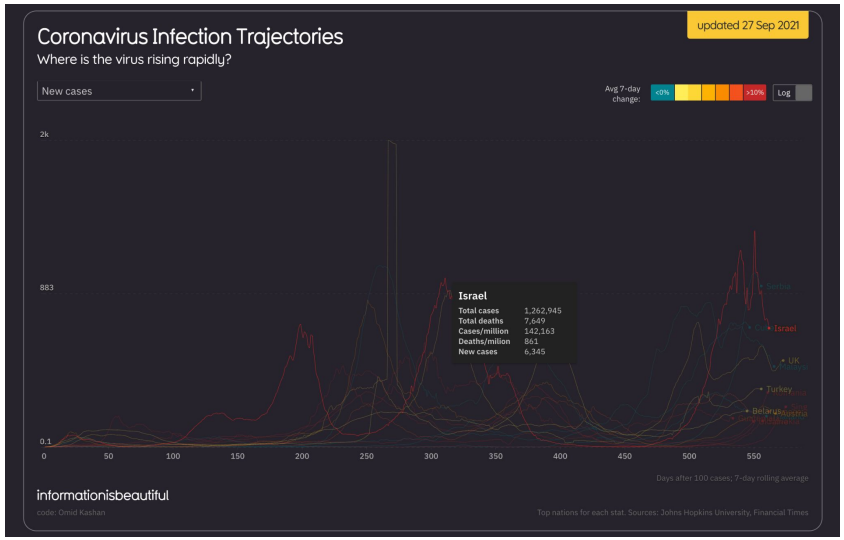
Specifications (4)데이터 시각화

Challenges

- Amazon QuickSight를 이용해 최종 결과를 도출, QuickSight 내에서 여러 형태로 데이터를 볼 수 있도록 customize
- Interactive하게 여러 데이터의 관계를 보여주기 위해 python libraries (folium, bokeh) 활용

결과물 : Interactive Website

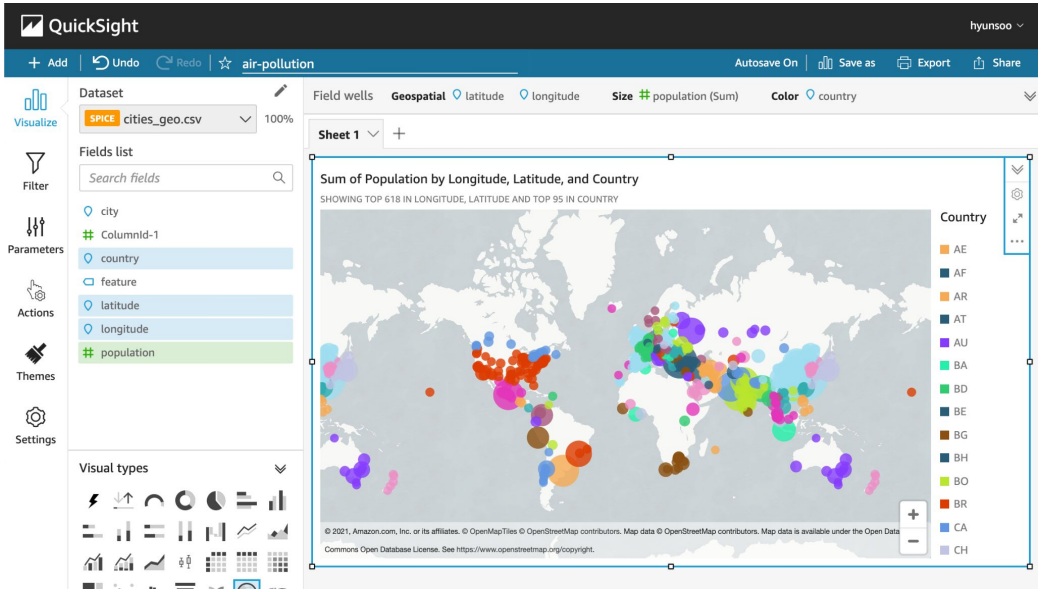
→ 여러가지 그래프가 보여지고, 그래프를 클릭하면 세부 정보를 더 볼 수 있는 형태



Specifications (4)데이터 시각화

1. Heat Map

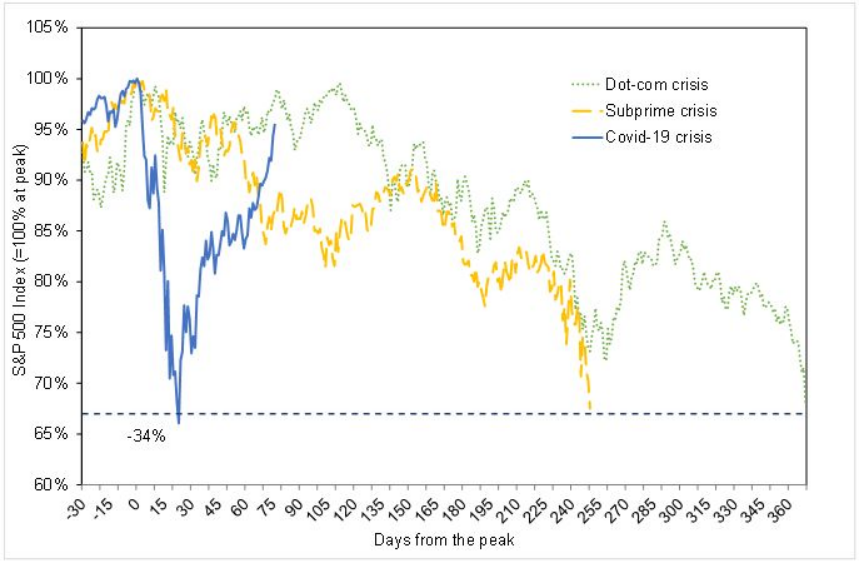
- Interactive Feature
  - heatmap에 air quality 변화를 시간에 따라 변화 시켜서 코로나 전후 비교
  - Geo heatmap, slider로 시간 조정하면 시간에 맞는 air quality를 세계 지도에 표시
- Prediction
  - 코로나가 없었을 경우의 air quality 추정

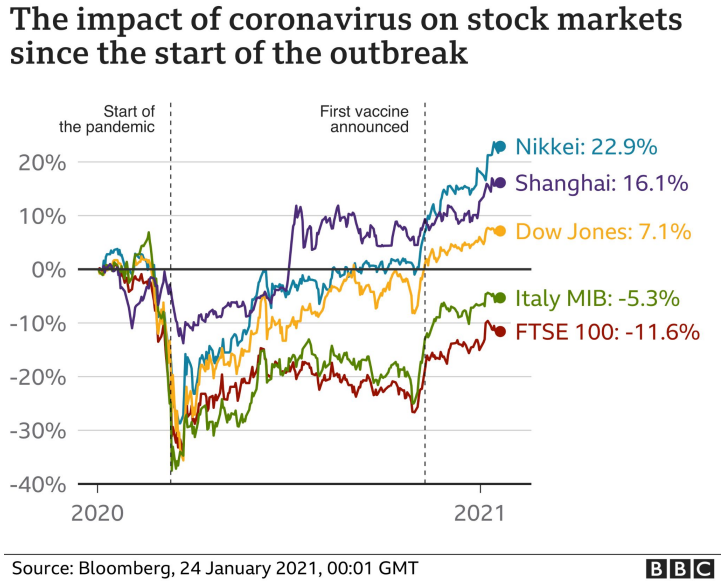


Specifications (4)데이터 시각화

2. Line Graph

- Interactive Feature
  - 주가 지수 그래프를 클릭하면 sector 별 변동량 그래프 노출
  - 그래프 위 hover 시 major event 관련 정보 및 그래프 상세보기 가능
  - 선택한 국가별 지수 그래프 비교
- Data Relation
  - major event의 지속일과 주가 지수 관계

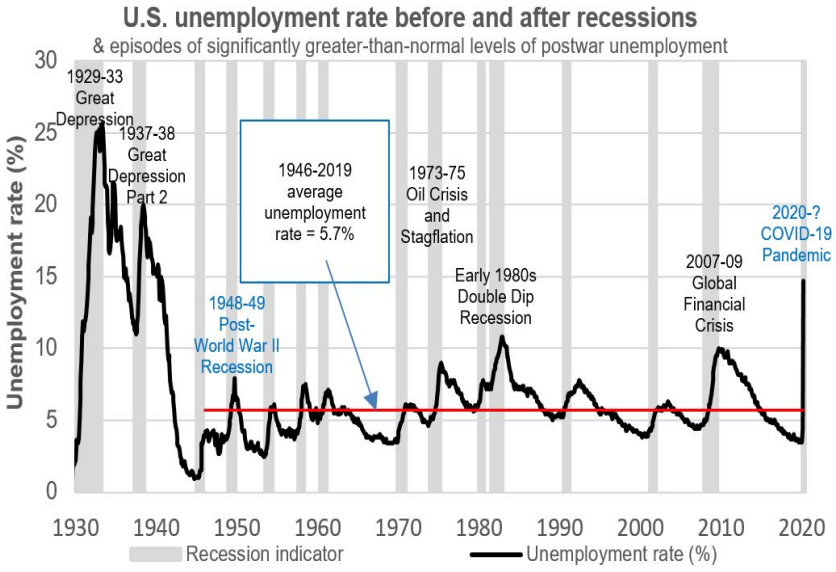


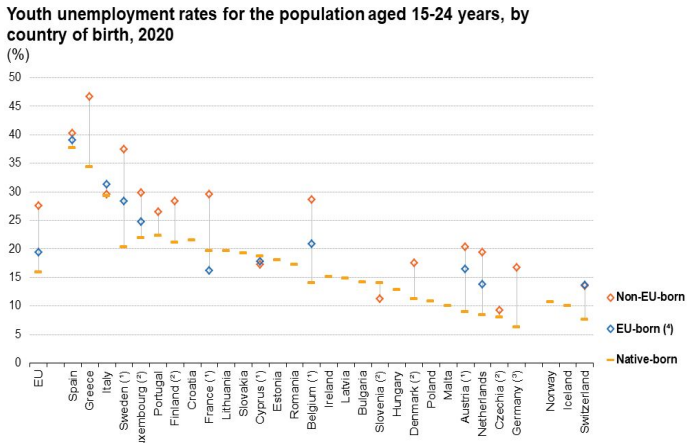


Specifications (4)데이터 시각화

3. Line + Stacked Bar Chart

- Interactive Feature
  - 관심국가 highlight, 외에 나머지 국가 hover로 확인 가능
  - 분산, 가장 높은 수치의 데이터, ATH 확인
  - 국가 선택 및 hover 시 detailed index number 확인 가능
- Prediction
  - 2021 Q4 ~ 2022 Q4 실업률 예측





Development Environment

Data Gathering

- Crawling (python)
- API

Data Processing and Analysis

- AWS services via CLI and console
- boto 3 library (python)
- Pyspark
- AWS Glue Crawling / Data Catalog
- AWS EMR (Hadoop, Spark 등)
- AWS Athena + SQL
- AWS Sagemaker
- vanilla python

Data Visualization

- AWS Quicksight
- python libraries (folium, bokeh)

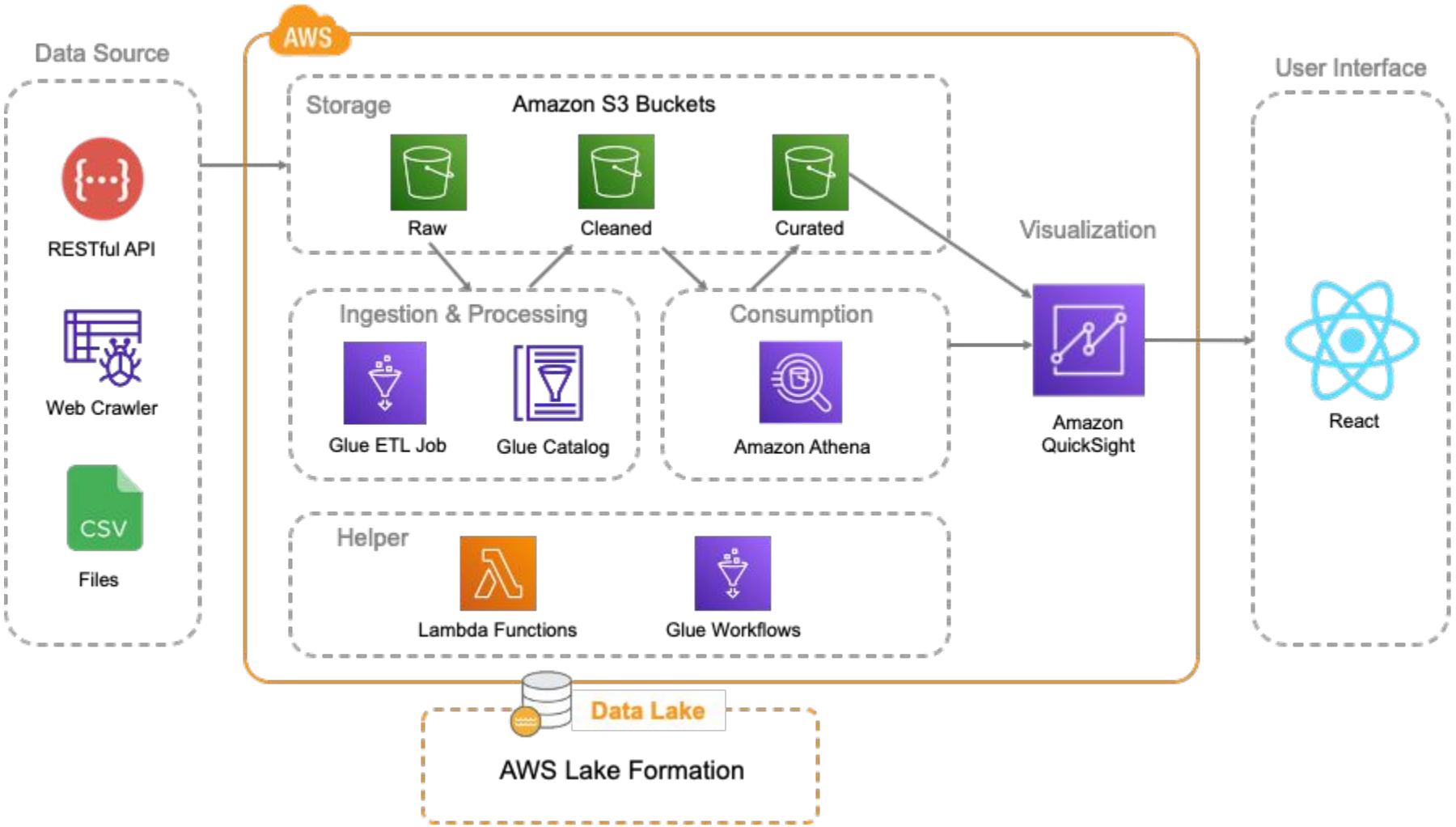
Web Framework

- React (javascript)

and more...

aws

Architecture



Current Status

AWS 사용자 계정 관리

- AWS 계정 루트 사용자 생성하고 역할을 만들어 IAM 사용자에게 권한 부여

AWS 데이터 레이크

- COVID-19 데이터 레이크용 Amazon S3 버킷 생성:
  - Raw 데이터, cleaned 데이터, curated 데이터
- AWS Lake Formation 구조 잡기:
  - COVID-19 데이터 레이크
  - 데이터 레이크 루트 위치로 S3경로를 등록
  - 데이터베이스를 만들어 메타데이터 저장

간단한 데이터 수집 및 프로세싱

- python, AWS console 활용해 데이터 S3에 업로드
- Glue 활용한 간단한 데이터 프로세싱

Division and Assignment of Work

주제별 데이터 전처리와 분석

1. 대기오염: 김현수
2. 실업률: 레이첼 간
3. 주가: 나혜주

Further Plans

향후 계획

- 웹크롤링을 통해서 대오염도, 주가 및 실업률에 관한 데이터를 수집하여 데이터베이스에 저장
- Amazon Athena로 저장된 데이터 분석
- AWS QuickSight로 데이터를 시각화하여 COVID-19 확산과 각 주제별 데이터 간의 상관관계 나타내기
- React 웹 애플리케이션 개발
- AWS QuickSight 임베딩 SDK를 사용하여 React 앱에서 대시보드 구현

Schedule

내용	9월		10월				11월				12월		
	3주	4주	1주	2주	3주	4주	1주	2주	3주	4주	1주	2주	3주
프로젝트 소개 및 자료조사													
주제확정 및 Infrastructure 구축													
프로젝트 스펙 발표													
주제별 데이터 수집 및 분석													
프로젝트 중간 발표													
데이터 시각화													
React 앱 개발 및 대시보드 구현													
Optimization													
프로젝트 최종 발표													

감사합니다