

Project folder location

Google Cloud

Search (/) for resources, docs, products, and more

Search

3

?

:

A

New Project

Project name *

FA25-I535-agr-chronicdisease

Project ID: fa25-i535-agr-chronicdisease. It cannot be changed later. [Edit](#)

Organization *

iu.edu

Select an organization to attach it to a project. This selection can't be changed later.

Location *

FA25-BL-INFO-I535

Browse

Parent organization or folder

Create

Cancel

Bucket and folder creation

Google Cloud

FA25-I535-agr-chronicdisease

Search (/) for resources, docs, products, and more

Search

2

?

:

A

Cloud Storage

Overview

Buckets

Monitoring

Settings

Storage Intelligence

Insights datasets

Configuration

Marketplace

Release Notes

<

Bucket details

fa25-i535-agr-chronicdisease-lake

Location

Storage class

Public access

Protection

us-central1 (Iowa)

Standard

Not public

Soft Delete

Objects

Configuration

Permissions

Protection

Lifecycle

Observability

New

Inventory Reports

Operations

Folder browser

fa25-i535-agr-chronicdisease-lake

clean/

meta/

schema/

validation/

quarantine/

raw/

Buckets > fa25-i535-agr-chronicdisease-lake > meta

Create folder

Upload

Transfer data

Other services

Filter by name prefix only

Filter

Filter objects and folders

Show

Live objects only

Table with 7 columns: Name, Size, Type, Created, Storage class, Last modified. Rows include schema/ and validation/.

Service account creation

The screenshot shows the Google Cloud IAM & Admin console. The left sidebar contains navigation links for IAM, PAM, Security Insights, Principal Access Boundaries, Identity & Organization, Policy Troubleshooter, Policy Analyzer, Organization Policies, Service Accounts (selected), Workload Identity Federation, Workforce Identity Federation, Labels, Tags, Settings, Privacy & Security, Manage Resources, and Release Notes. The main content area is titled 'Service accounts' and shows a list of service accounts for project 'FA25-I535-agr-chronicdisease'. The list includes three accounts: 'fa25-i535-agr-chronicdisease@appspot.gserviceaccount.com', 'chronicdisease-pipeline-sa@fa25-i535-agr-chronicdisease.iam.gserviceaccount.com', and '693764657658-compute@developer.gserviceaccount.com'. All accounts are in an 'Enabled' status. A 'Recommended for you' sidebar on the right lists various actions like 'Service accounts overview', 'Create service accounts', 'List and edit service accounts', 'Disable and enable service accounts', and 'Delete and undelete service accounts'.

Email	Status	Name	Description	Key ID	Actions
fa25-i535-agr-chronicdisease@appspot.gserviceaccount.com	Enabled	App Engine default service account		No keys	
chronicdisease-pipeline-sa@fa25-i535-agr-chronicdisease.iam.gserviceaccount.com	Enabled	chronicdisease-pipeline-sa		No keys	
693764657658-compute@developer.gserviceaccount.com	Enabled	Default compute service account		No keys	

Downloading the dataset

The screenshot shows a Google Cloud Shell terminal window. The terminal displays the following commands and output:

```
-bash: NLH6-MH6D-601Q-05JE: command not found
agr@cloudshell:~ (fa25-i535-agr-chronicdisease) $ gcloud config get-value project
Your active configuration is: [cloudshell-6056]
fa25-i535-agr-chronicdisease
agr@cloudshell:~ (fa25-i535-agr-chronicdisease) $ cd -
agr@cloudshell:~ (fa25-i535-agr-chronicdisease) $ ls
README-cloudshell.txt
agr@cloudshell:~ (fa25-i535-agr-chronicdisease) $ mkdir -p chronicdisease_project && cd chronicdisease_project
-bash: chronicdisease_project: command not found
agr@cloudshell:~ (fa25-i535-agr-chronicdisease) $ cd -
mkdir -p cdi_project && cd cdi_project
agr@cloudshell:~/cdi_project (fa25-i535-agr-chronicdisease) $ curl -L "https://data.cdc.gov/api/views/hksd-2xuw/rows.csv?accessType=DOWNLOAD" \
-o cdi_raw_full.csv
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
           Dload  Upload   Total   Spent    Left    Speed
100 84.0M  0 84.0M    0  4237k    0 --:--:--  0:00:20 --:--:-- 4269k
agr@cloudshell:~/cdi_project (fa25-i535-agr-chronicdisease) $ ls -lh cdi_raw_full.csv
-rw-rw-r-- 1 agr agr 85M Nov 25 01:07 cdi_raw_full.csv
agr@cloudshell:~/cdi_project (fa25-i535-agr-chronicdisease) $ ls -lh cdi_raw_full.csv
-rw-rw-r-- 1 agr agr 85M Nov 25 01:07 cdi_raw_full.csv
agr@cloudshell:~/cdi_project (fa25-i535-agr-chronicdisease) $ head -5 cdi_raw_full.csv
YearStart,YearEnd,LocationAbbr,LocationDesc,DataSource,Topic,Question,Response,DataValueUnit,DataValueType,DataValue,DataValueAlt,DataValueFootnoteSymbol,DataValueFootnote,LowConfidence
Limit,HighConfidenceLimit,StratificationCategory1,Stratification1,StratificationCategory2,Stratification2,StratificationCategory3,Stratification3,Geolocation,LocationID,TopicID,Question
ID,ResponseID,DataValueUnitID,StratificationID1,StratificationID2,StratificationCategoryID2,StratificationID3,StratificationCategoryID3,StratificationID3
2020,2020,US,United States,BRPHS,Health Status,Recent activity limitation among adults,,Number,Age-adjusted Mean,2.9,2.9,,2.9,2.9,Sex,Female,,,,,59,HEA,AGEADJMEAN,SEX,SEXP,
2015,2019,AR,Arkansas,US Cancer DVT,Cancer,"Invasive cancer (all sites combined), incidence",,Number,Number,9537,9537,,,,,Sex,Male,,,,,POINT (-92.27449074299966 34.74865012400045),05,CA
N,CAN07,,NMNR,SEX,SEXM,
2015,2019,CA,California,US Cancer DVT,Cancer,"Cervical cancer mortality among all females, underlying cause",,Number,Number,486,486,,,,,Overall,Overall,,,,,POINT (-120.99999953799971 37
.63864012300047),06,CAN,CAN03,,NMNR,OVERALL,OVR,
2015,2019,CO,Colorado,US Cancer DVT,Cancer,"Invasive cancer (all sites combined), incidence",,Number,Number,2880,2880,,,,,Race/Ethnicity,Hispanic,,,,,POINT (-106.13361092099967 38.84384
0757000464),08,CAN,CAN07,,NMNR,RACE,HIS,
agr@cloudshell:~/cdi_project (fa25-i535-agr-chronicdisease) $
```

File Structure after pre processing script

The screenshot shows the Google Cloud Storage interface for the bucket 'FA25-I535-agr-chronicdisease'. The left sidebar shows the 'Buckets' section. The main area displays the bucket's contents, including folders for 'clean/' and 'quarantine/'. The 'clean/' folder contains subfolders for years 2015 through 2022, and a 'meta/' folder. The 'quarantine/' folder contains a 'cdi/' subfolder. A file named 'validation_20251125T011904Z.json' is listed under the 'clean/' folder.

The Cloud Shell terminal output shows the following commands and results:

```
Uploaded out_clean/cdi_clean_year_2021.csv -> gs://fa25-i535-agr-chronicdisease-lake/clean/cdi/year=2021/cdi_clean_year_2021.csv
Uploaded out_clean/cdi_clean_year_2022.csv -> gs://fa25-i535-agr-chronicdisease-lake/clean/cdi/year=2022/cdi_clean_year_2022.csv
Uploaded out_quarantine/cdi_quarantine_20251125T011904Z.csv -> gs://fa25-i535-agr-chronicdisease-lake/quarantine/cdi/cdi_quarantine_20251125T011904Z.csv
Uploaded validation_summary_20251125T011904Z.json -> gs://fa25-i535-agr-chronicdisease-lake/meta/validation/validation_20251125T011904Z.json
Validation summary: {'run_ts': '20251125T011904Z', 'raw_blob': 'raw/cdi_20251125.csv', 'total_rows': 309215, 'valid_rows': 209196, 'invalid_rows': 100019, 'required_fields': ['YearStart', 'YearEnd', 'LocationAbbr', 'LocationDesc', 'Topic', 'Question', 'DataValue']}
```

Sample Bigquery result

The screenshot shows the Google Cloud BigQuery interface. The left sidebar shows the 'fa25-i535-agr-chronicdisease' dataset with tables 'cdi_lake', 'cdi_clean', 'cdi_raw_external', and 'cdi_raw_external'. The main area displays the query results for the query:

```
1 SELECT
2 (SELECT COUNT(*) FROM `fa25-i535-agr-chronicdisease.cdi_lake.cdi_raw_external`) AS raw_rows,
3 (SELECT COUNT(*) FROM `fa25-i535-agr-chronicdisease.cdi_lake.cdi_clean`) AS clean_rows;
```

The query results show the following data:

Row	raw_rows	clean_rows
1	309215	30312

Chart 1 - Topic coverage

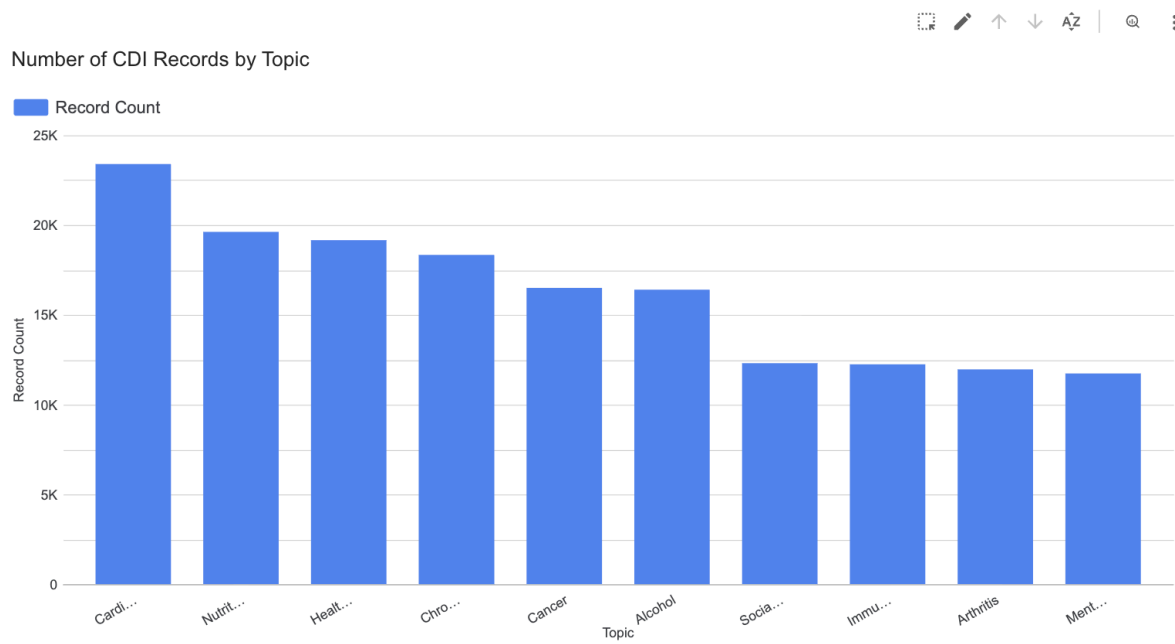


Chart 2 - Records per Year

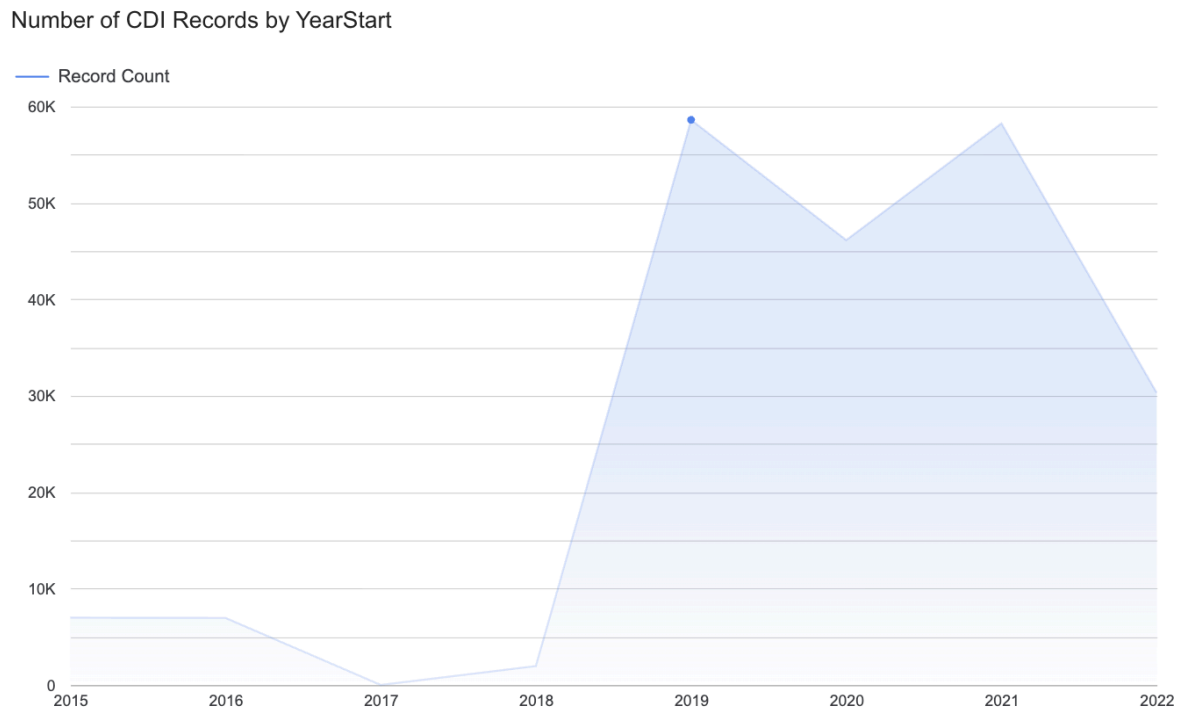


Chart 3 - Choropleth Map: Average

Average Diabetes Indicator (%) by State, 2022

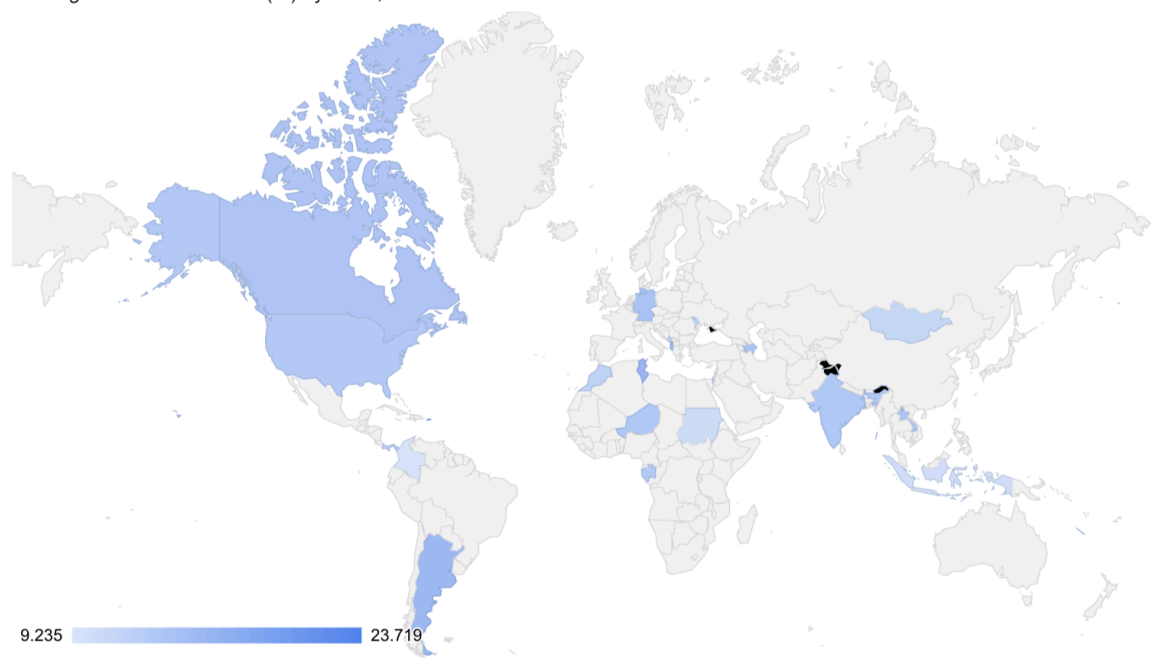


Chart 4 - Multi-Line Trend: Diabetes by Sex in One State

Average Diabetes Indicator (%) by Sex in California, Over Time

