AWS Big Data Blog

# Visualize over 200 years of global climate data using Amazon Athena and Amazon QuickSight

by Joe Flasher and Conor Delaney | on 13 FEB 2019 | in Amazon Athena, Amazon QuickSight, AWS Big Data | Permalink | Comments | Share

Climate Change continues to have a profound effect on our quality of life. As a result, the investigation into sustainability is growing. Researchers in both the public and private sector are planning for the future by studying recorded climate history and using climate forecast models.

To help explain these concepts, this post introduces the Global Historical Climatology Network Daily (GHCN-D). This registry is used by the global climate change research community.

This post also provides a step-by-step demonstration of how Amazon Web Services (AWS) services improve access to this data for climate change research. Data scientists and engineers previously had to access hundreds of nodes on high-performance computers to query this data. Now they can get the same data by using a few steps on AWS.

## Background

Global climate analysis is essential for researchers to assess the implications of climate change on the Earth's natural capital and ecosystem resources. This activity requires high-quality climate datasets, which can be challenging to work with because of their scale and complexity. To have confidence in their findings, researchers must be confident about the provenance of the climate datasets that they work with. For example, researchers may be trying to answer questions like: has the climate of a particular food producing area changed in a way that impacts food security? They must be able to easily query authoritative and curated datasets.

The National Centers for Environmental Information (NCEI) in the U.S. maintains a dataset of climate data that is based on observations from weather stations around the globe. It's the Global Historical Climatology Network Daily (GHCN-D) — a central repository for daily weather summaries from ground-based stations. It is comprised of millions of quality-assured observations that are updated daily.

The most common parameters recorded are daily temperatures, rainfall, and snowfall. These are useful parameters for assessing risks for drought, flooding, and extreme weather.

## The challenge

The NCEI makes the GHCN_D data available in CSV format through an FTP server, organized by year. Organizing the data by year means that a complete copy of the archive requires over 255 files (the first year in the archive is 1763). Traditionally, if a researcher wants to work on this dataset they must download it and work on it locally. For a researcher to be sure of using the latest data for their analysis, they must repeat this download every day.

For researchers, deriving insight from this data can be a challenge. They must be able to fully engage with the data, because that requires technical skill, computing resources, and subject matter expertise.

## A new efficient approach

Through AWS's collaboration with the NOAA Big Data Project, a daily snapshot of the GHCN_D dataset is now available on AWS. The data is publically accessible through an Amazon S3 bucket. For more information, see the Registry of Open Data on AWS.

Having the data available in this way offers several advantages:

- **The data is globally available to a community of users**. Users no longer must download data to work on it. Everyone can work with the same, authoritative copy.
- **Time to insight is reduced**. By taking advantage of AWS services, researchers can immediately start to perform analysis.
- **The cost of research is reduced**. Researchers can switch off resources as soon as their analysis is finished.

This blog post illustrates a workflow using Amazon S3, Amazon Athena, AWS Glue, and Amazon QuickSight that demonstrates how quickly one can derive insights from this dataset.

The workflow presented in this post follows these general steps:

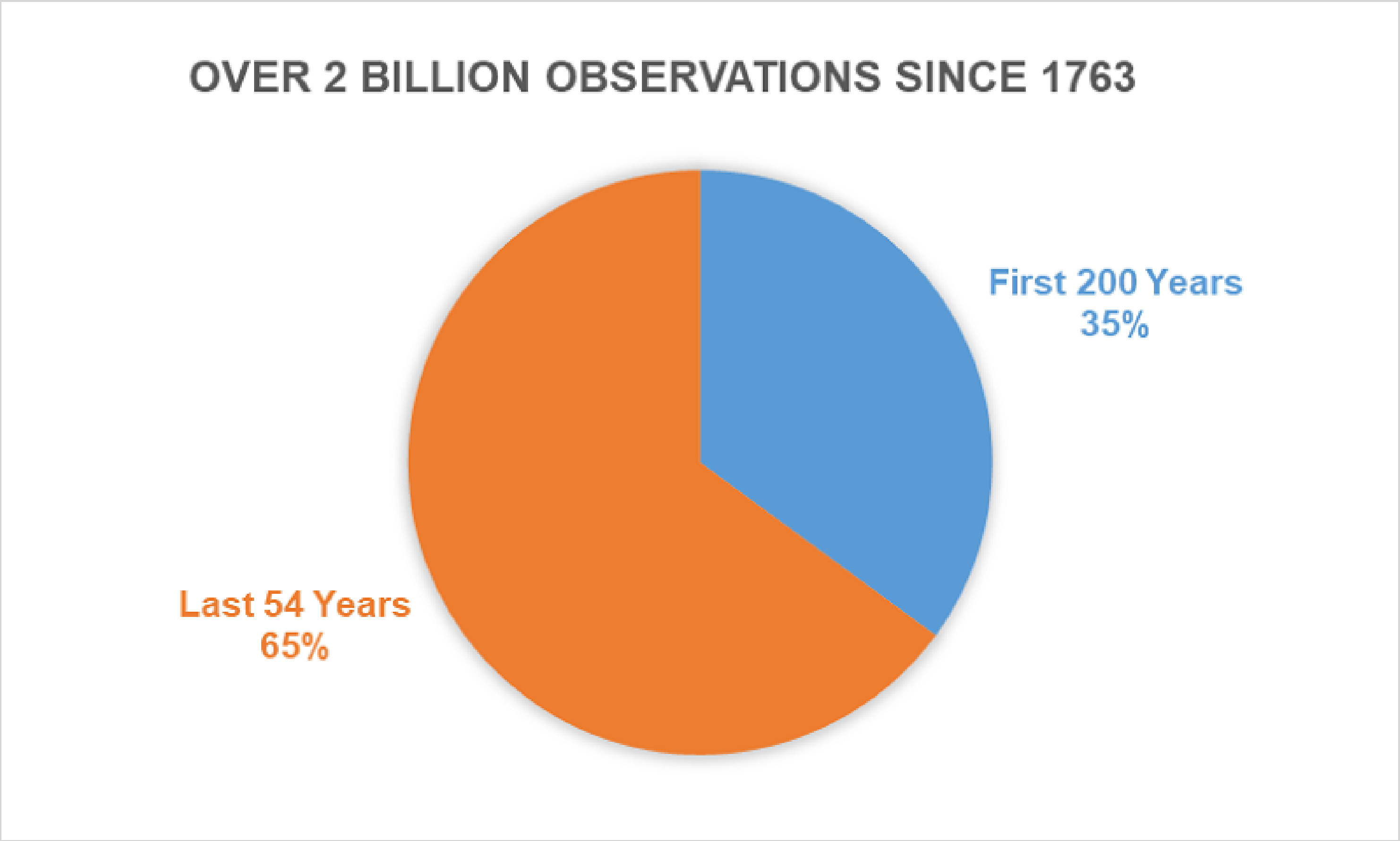
- Extract data files from the NOAA bucket and make the data available as tables.
- Use SQL to query the data contained in the tables.
- Show how to speed up analysis by creating tables from queries and storing those tables in a private Amazon S3 bucket.
- Visualize the data to gain insight.

## Overview of the GHCN_D dataset

The GHCN-D is a quality-assured dataset that contains daily weather summaries from weather stations across global land areas. It has the following properties:

- Data is integrated from approximately 30 sources that provide weather observations from various national and international networks.
- A comprehensive dataset for the US and good coverage for many parts of the world.
- There are many types of daily weather observations in this dataset, but the majority are maximum temperature, minimum temperature, precipitation, snow fall, and snow depth. These observations include:
    - Over 35,000 temperature stations.
    - Over 100,000 precipitation stations.
    - Over 50,000 snowfall or snow depth stations
- The source of each datum, the term used for a single record, is contained in the dataset. Each datum has a quality control flag associated with it.
- The dataset is updated daily. The historic sources are reprocessed weekly.

You can see in the graphic below how the data volume has grown in recent decades.



*1763 to 2018. For 1763 there are less than a thousand observations. For 2017 there are over 34 million observations.*

## Organization of the data on Amazon S3

As previously mentioned, the GHCN-D dataset is accessible through an Amazon S3 bucket. The details of the dataset are on the Registry of Open Data on AWS (RODA). The landing page for the dataset on RODA contains a link to a comprehensive readme file for the dataset. This readme contains all of the lookup tables and variable definitions.

This section shows the pertinent information required to start working with the dataset.

The data is in a text, or comma-separated values (CSV), format and is contained in the Amazon S3 bucket called noaa-ghcn-pds.

The noaa-ghcn-pds bucket contains virtual folders, and is structured like this:

- **noaa-ghcn-pds**. This is the root of the bucket with two subdirectories and a number of useful files. For the purposes of this exercise, we use only the ghcnd-stations.txt file. This file contains information about the observation stations that produced the data for the GHCN_D dataset. You must download the ghcnd-stations.txt file.
- **noaa-ghcn-pds/csv/**. This virtual folder contains all of the observations from 1763 to the present organized in .csv files, one file for every year. For this exercise, we'll collate this data into a single table.

Also for the purpose of this exercise, the data from 'ghcnd-stations.txt' and the data contained in noaa-ghcn-pds/csv/ are extracted and added to two separate tables. These tables are the basis of the analysis.

The tables are labeled as:

- **tblallyears**. This table contains all the records stored in the yearly .csv files from 1763 to present.
- **tblghcnd_stations**. This table contains information for over 106,430 weather stations.

Point of interest: the .csv file from the year 1763 contains the data for one weather station. That station was located in the center of Milan, Italy.

## The tools

To implement the general workflow in this exercise, we're using the following tools:

- Amazon Simple Storage Service (Amazon S3) to stage the data for analysis. The GHCN_D dataset is stored in a bucket on Amazon S3. We also use a private bucket to store new tables created from queries.
- Amazon Athena to query data stored on Amazon S3 using standard SQL.
- AWS Glue to extract and load data into Athena from the Amazon S3 buckets in which it is stored. AWS Glue is a fully managed extract, transform, and load (ETL) service.
- AWS Glue Data Catalog to catalog the data that we query with Athena.
- Amazon QuickSight to build visualizations, perform ad hoc analyses, and get insights from the dataset. Queries and tables from Athena can be read directly from Amazon QuickSight. Amazon QuickSight can also run queries against tables in Athena.

To implement the processes outlined in this post, you need an AWS Account. For more information about creating an AWS account, see Getting Started with AWS. You also must create a private Amazon S3 bucket located in the N. Virginia AWS Region. For more information, see Create a Bucket.

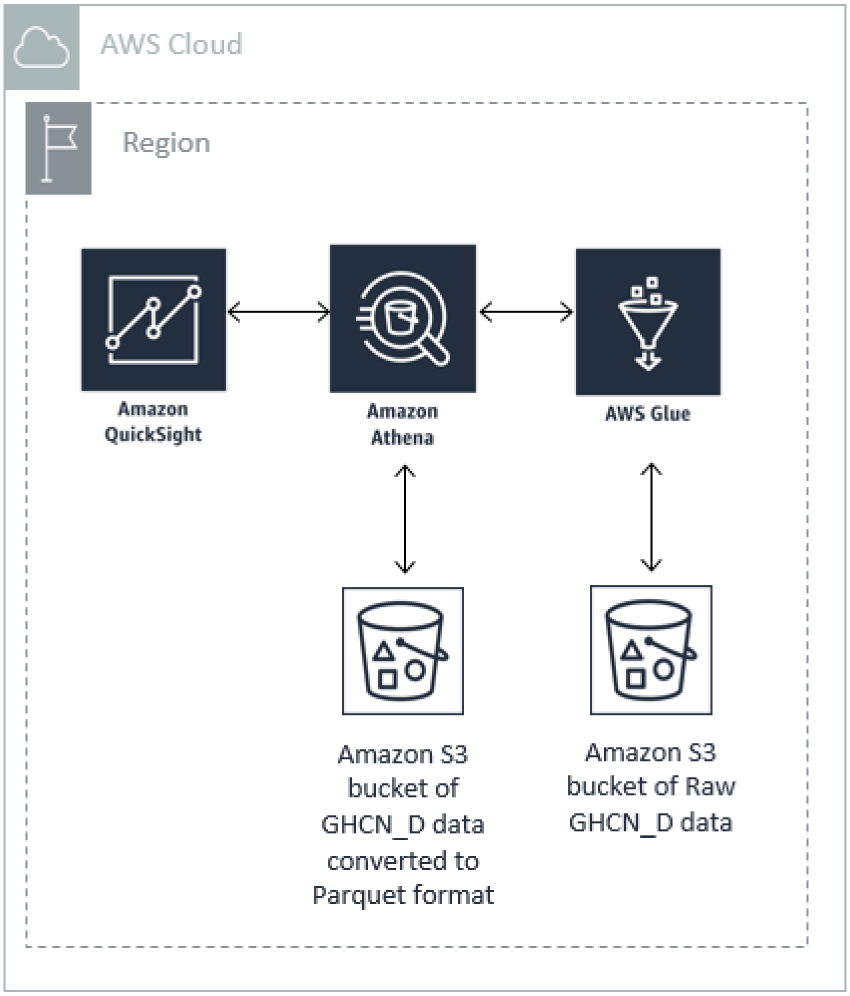
When you create the bucket, it must contain the following empty directories:

1. [your_bucket_name]/stations_raw/
2. [your_bucket_name]/ghcnblog/
3. [your_bucket_name]/ghcnblog/stations/
4. [your_bucket_name]/ghcnblog/allyears/
5. [your_bucket_name]/ghcnblog/1836usa/

The following is an overview of how the various AWS services interact in this workflow.

**Note**

The AWS services are in the same AWS Region. One of the Amazon S3 buckets is the existing one that stores the GHCN_D data. The other Amazon S3 bucket is the bucket that you use for storing tables.


AWS Cloud
Region
Amazon QuickSight
Amazon Athena
AWS Glue
Amazon S3 bucket of GHCN_D data converted to Parquet format
Amazon S3 bucket of Raw GHCN_D data

*How the AWS services work together to compose this workflow.*

In addition to using the Console for the below steps, you can also use the AWS Command Line Interface or any of the Software Developer's Kits (SDK), provided in a number of languages. Also, the queries you create in Athena can be programmatically executed on a scheduled basis or in response to triggers you chose using services like AWS Lambda.

## The workflow

Now that we have the tools and the data, we are ready to:

1. Extract the yearly .csv files and add them to a table in Amazon Athena.
2. Extract the stations text file and add it to a separate table in Amazon Athena.

### Extract the yearly .csv files and add it to a table in Amazon Athena

The complete set of daily weather observations is organized by year in one folder of the Amazon S3 bucket in .csv format. The path to the data is **s3://noaa-ghcn-pds/csv/**.

Each file is named by year beginning with 1763.csv and progressing one year at a time up to the present.

From the AWS console, click on AWS Athena. This takes you to the main dashboard for Athena. From here, click on AWS Glue Data Catalog. This brings you to AWS Glue.

In AWS Glue, choose the **Tables** section on the left side. Then, in the **Add table** drop-down menu, choose **Add table manually**. A series of forms displays for you to add the following information:

- Set up your table's properties:
  - Give the new table a name, for example, tblallyears
  - Create a database and name it ghcnblog.

The database then appears in the Athena dashboard.

- Add a data store:
  - Choose the **Specified path in another account** option, and enter the following path in the text box: **s3://noaa-ghcn-pds/csv/**
- Choose a data format:
  - Select **CSV**, then select **Comma** as the delimiter.
- Define a schema:
  - Add the following columns as string variables:
    - id
    - year_date
    - element
    - data_value
    - m_flag
    - q_flag
    - s_flag
    - obs_time

For a full description of the variables and data structures, see the readme file.

- Choose **OK**, then **Finish**.

Now return to the Athena dashboard, and choose the database that you created. The table will appear in the list of tables on the left. You can now preview the data by choosing the 'Preview table' option to the right of the table.

## Use CTAS to speed up queries

As a final step, create a table using the SQL statement called CREATE TABLE AS SELECT (CTAS). Store the table in a private Amazon S3 bucket.

This step dramatically speeds up queries. The reason is because in this process we extract the data once and store the extracted data in a columnar format (Parquet) in the private Amazon S3 bucket.

To illustrate the improvement in speed, here are two examples:

- A query that counts all of the distinct IDs, meaning unique weather stations, takes around 55 seconds and scans around 88 GB of data.
- The same query on the converted data takes around 13 seconds and scans about 5 GB of data.

To create the table for this final step:

1. Open the Athena console.
2. In the dashboard, select **New query**, then enter the query as shown in the following example. Make sure to enter the information that's applicable to your particular situation, such as your bucket name.

```
/*converting data to Parquet and storing it in a private bucket*/
CREATE table ghcnblog.tblallyears_qa
WITH (
  format='PARQUET', external_location='s3://[your-bucket-name]/ghcnblog/allyearsqa
) AS SELECT * FROM ghcnblog.tblallyears
WHERE q_flag = '';
```

3. Make sure that the data format is Parquet.
4. Name your table **tblallyears_qa**.
5. Add this path to this folder in the private Amazon S3 bucket: **[your_bucket_name]/ghcnblog/allyearsqa/**. Replace your_bucket_name with your specific bucket name.

The new table appears in your database, listed on the left side of the Athena dashboard. This is the table that we work with going forward.

## Extract the stations text file and add it to a separate table in Amazon Athena

The stations text file contains information about the weather stations, such as location, nationality, and ID. This data is kept in a separate file from the yearly observations. We need to import this data to look at the geographical spread of weather observations. While dealing with this file is a bit more complicated, the steps to importing this data into Athena are similar to what we have already done.

To import this data into Athena:

1. Download the ghcnd-stations text file.
2. Open the file in a spreadsheet program and use the fixed width-delimited data import function. The fixed widths of the columns are described in the readme file in the section called **FORMAT OF "ghcnd-stations.txt" file**.
3. After you successfully import the data, save the spreadsheet as a .csv text file.
4. Copy the new .csv file to **[your_bucket_name]/stations_raw/**. Replace your_bucket_name with your specific bucket name.
5. Using this new .csv file, follow the **Add table process** steps in AWS Glue, as described earlier in this post.
    - Use the following field names:
        - id
        - latitude
        - longitude
        - elevation
        - state

        - name
        - gsn_flag
        - hcn_flag
        - wmo_id
    - Name the table **tblghcnd_stations**.
6. After the table is created, follow the CREATE TABLE AS SELECT (CTAS) steps for this table as described earlier in this post.
    - Name the new table **tblghcnd_stations_qa**.
    - Store the new table in [your_bucket_name]/ghcnblog/stations/. Replace your_bucket_name with your specific bucket name.

The two most important datasets of GHCN_D are now in Athena.

In the next section, we run queries against these tables and visualize the results using Amazon QuickSight.

## Exploratory data analysis and visualization

With our two tables created, we are now ready to query and visualize to gain insight.

### Exploratory data analysis

In the Athena query window, run the following queries to get an idea of the size of the dataset.

Query #1: the total number of observations since 1763:

```
SELECT count(*) AS Total_Number_of_Observations
FROM ghcnblog.tblallyears_qa;
```

Query #2: the number of stations since 1763:

```
SELECT count(*) AS Total_Number_of_Stations
FROM ghcnblog.tblghcnd_stations_qa;
```

### Average weather parameters for the Earth

The following figure shows a query that calculates the average maximum temperature (Celsius), average minimum temperature (Celsius), and average rainfall (mm) for the Earth since 1763.

In the query, we must convert the data_value from a String variable to a Real variable. We also must divide by 10, because the temperature and precipitation measurements are in tenths of their respective units. For more information about these details and the element codes (TMIB, TMAX and PRCP), see the readme file.

```
SELECT element,
         round(avg(CAST(data_value AS real)/10),2) AS value
FROM ghcnblog.tblallyears_qa
WHERE element IN ('TMIN', 'TMAX', 'PRCP')
GROUP BY  element;
```

It would be convenient if we could just run simple queries, such as this one, on this dataset and accept that the results are correct.

The previous query is assuming an even and equal spread of weather stations around the world since 1763. In fact, the number and spread of weather stations varied over time.

## Visualizing the growth in numbers of weather stations over time

The following query visualizes the number of weather stations for each year since 1763, by using Amazon QuickSight.

**Note:** You must be signed up for Amazon QuickSight to complete these steps. During the sign-up process, you are prompted to manage your Amazon QuickSight data source permissions. At this time, use step 3 in the following procedure to grant access to the Amazon S3 buckets and to Athena.

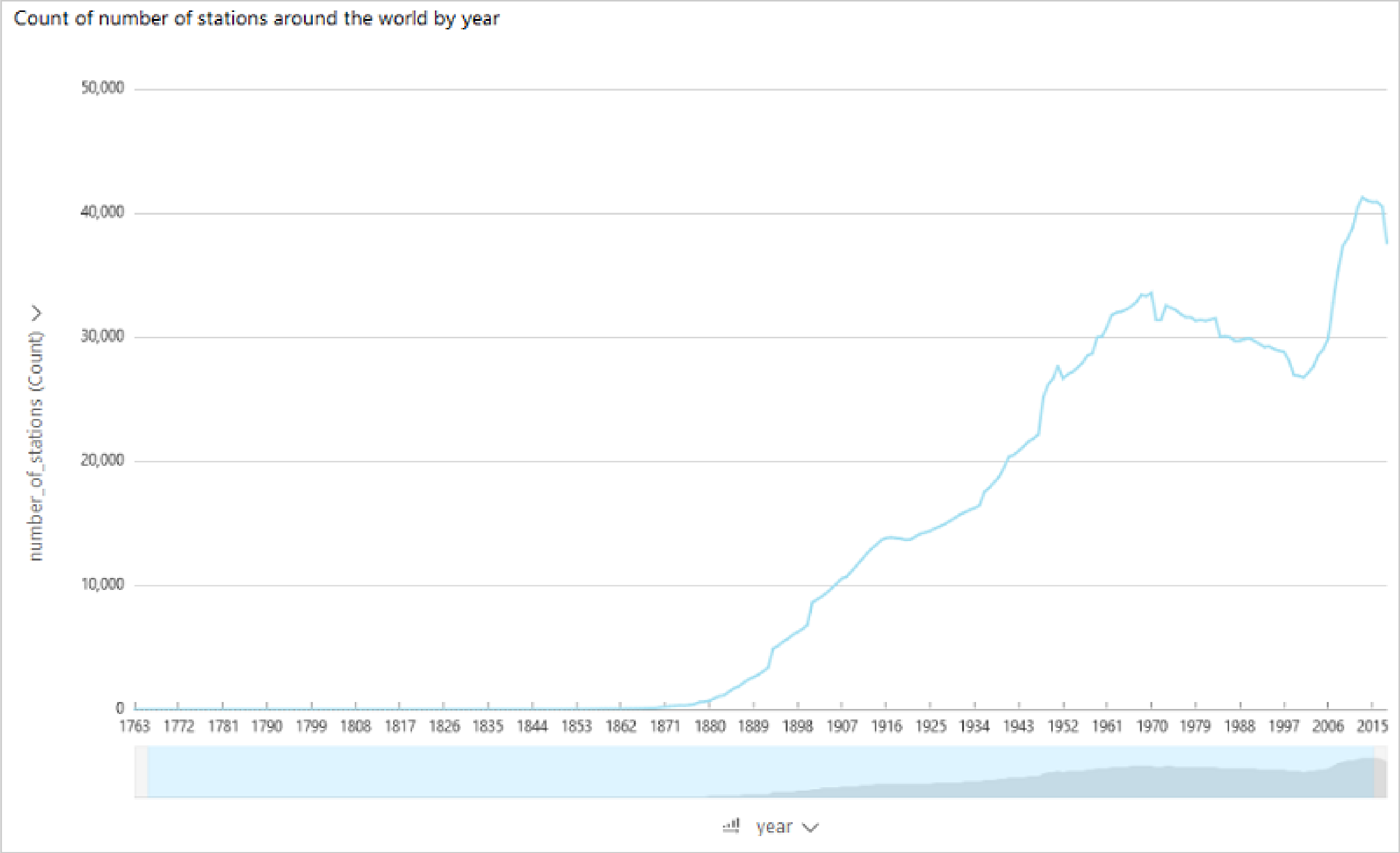
The steps are as follows:

1. Open the Amazon QuickSight console.
2. On the far right of the dashboard, choose **Manage QuickSight**.
3. Choose **Account Setting**, then **Manage QuickSight permissions**. Give Amazon QuickSight permission to access Athena, and to read the Amazon S3 bucket that contains the new tables.
4. Return to Amazon QuickSight by choosing the logo on the top left side of the screen.
5. From the Amazon QuickSight dashboard, choose **New analysis**, then **New data set**.
6. From the **Create a Data Set** tiles, choose **Athena**. Name the data source, for example ghcnblog, then choose **Create data source**.
7. Choose the option to add a custom SQL, then add the SQL, as shown in the following example:

```
SELECT DISTINCT id AS numberofstations,
substr(year_date,1,4) as year
FROM ghcnblog.tblallyears_qa
GROUP BY substr(year_date,1,4), id
ORDER BY substr(year_date,1,4)
```

8. Choose **Confirm query**.
9. Choose **Directly query your data**.
10. Choose **Visualize**.

11. To make the graph, choose the line chart graph. Add **year** to the X-axis and **number_of_stations** to the Value field wells. The options appear to the left of the visualization dashboard.



*The number of global weather stations used by GHCN_D over time.*

The resulting graph shows that the number and spread of stations around the world has varied over time.

## A look at the development of observation in the US

1836 is the year of the first US observation station in the data set. To get an insight into the development of observations the US, we extracted a subset of US data from the main data source (tblallyears_qa). This dataset features annual data every 30th year from 1836 to 2016.

This query generates a large dataset. To improve performance, save the query as a table stored in an Amazon S3 bucket using the previously described procedure.

The query to do this in one step is shown in the following figure.

```
CREATE TABLE ghcnblog.tbl1836every30thyear
WITH (
  format='PARQUET',
  external_location='s3://[your-bucket-name]/ghcnblog/1836every30years/'
) AS
SELECT TA.id as id, substr(TA.year_date,1,4) as year, TS.state, CAST(TS.longitude as re
FROM "ghcnblog".tblallyears_qa as TA, "ghcnblog".tblghcnd_stations_qa as TS
WHERE substr(TA.year_date,1,4) IN ('1836', '1866', '1896', '1926', '1956', '1986', '20
AND substr(TA.id,1,2) = 'US'
AND state <> 'PI'
AND TRIM(TA.id) = TRIM(TS.id)
GROUP BY TA.id, substr(TA.year_date,1,4), state, longitude, latitude, element, data_va
```

The table appears in the Amazon Athena dashboard as **tbl1836every30thyear** and it forms the basis for our analysis.

In the Amazon QuickSight console, use the follow SQL to generate a new dataset.

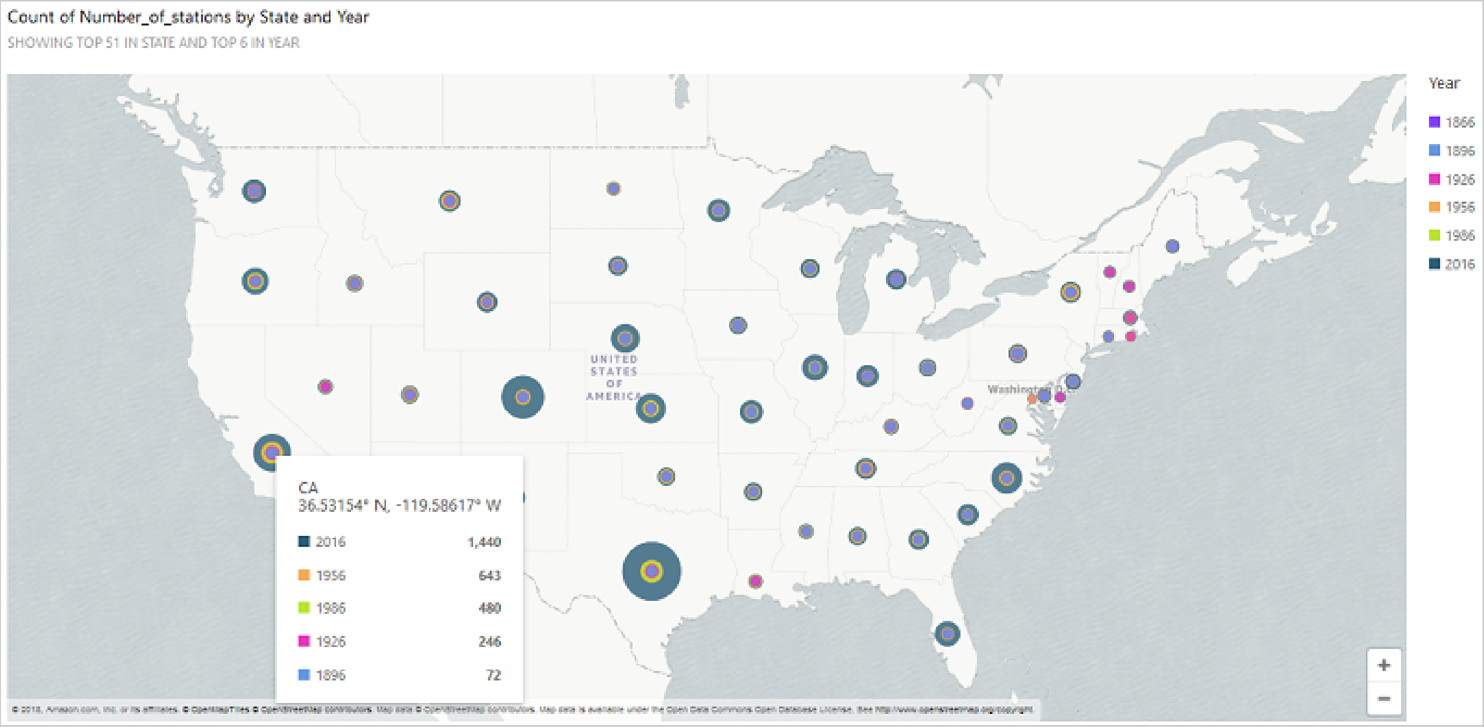
```
SELECT DISTINCT(id) AS number_of_stations, year, state
FROM ghcnblog.tbl1836every30thyear
GROUP BY year, id, state
ORDER BY year
```

1. Choose **Confirm query**.
2. Choose **Directly query your data**.
3. Choose **Visualize**.

This brings you back to the visualization dashboard. From this dashboard, chose the **Points on a map** visualization, and set up the fields as follows:

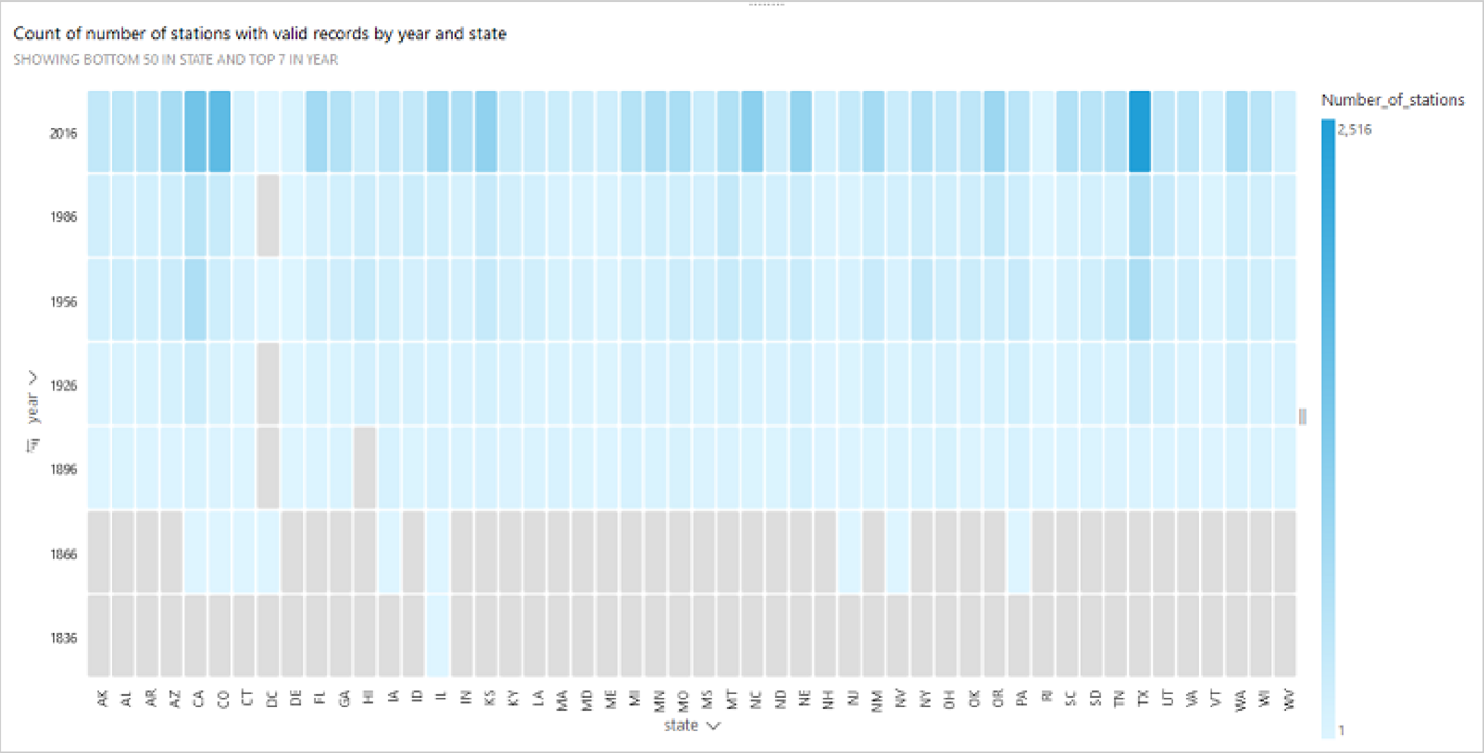
- **Geospatial**: state
- **Size**: number_of_stations, aggregate by count.
- **Color**: year

The results should be the following map of the US showing the growth of weather stations used by GHCN_D from 1836 to 2016 at 30-year increments. In 1836, there was one station. By 2016, there were thousands.


Count of Number_of_stations by State and Year
SHOWING TOP 51 IN STATE AND TOP 6 IN YEAR
Year
1866
1896
1926
1956
1986
2016
UNITED STATES OF AMERICA
Washington D.C.
CA
36.53154° N, -119.58617° W
2016 1,440
1956 643
1986 480
1926 246
1896 72
+
−

*The growth of the number of observations stations in the US.*

Interestingly, some states had more stations in 1956 than they did in 1986. This is also illustrated in the following figure. The data for the figure was derived from the previous dataset.

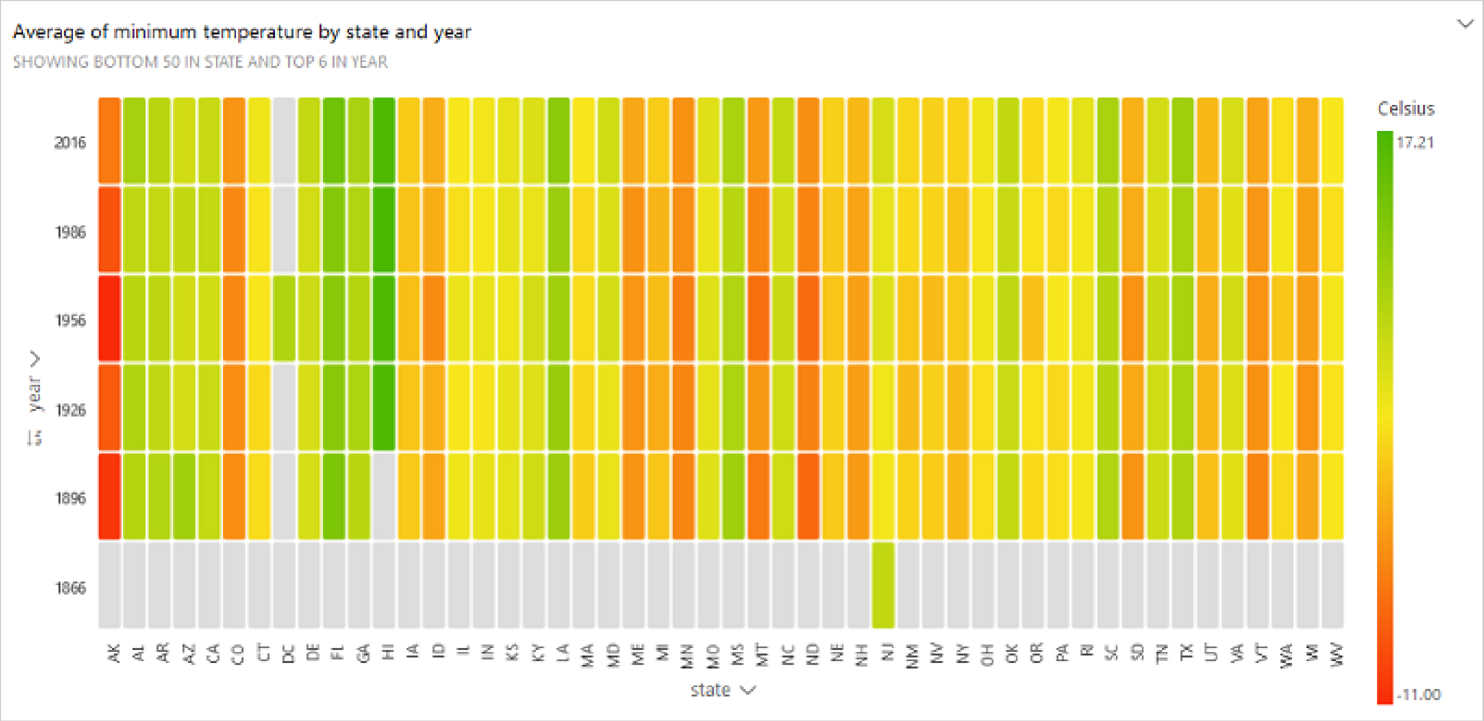


*This heat map illustrates the number of stations per state over time. This is a 30th year snapshot.*

## A look at more data

We have now a data lake of GHN_D data. By using the tools that we have assembled, we are free to experiment with the data. It is now possible to construct queries and visualization on this huge dataset to gain insights.

The following figure shows the heat map that we created. It shows the average minimum temperature in US states over time. As before, we are looking at 30-year snapshots; that is to say, every 30th year we take a yearly average.



*This heat map illustrates the minimum temperate for each state over time. A yearly average every 30th year starting at 1836.*

## Summary

Our headlines are full of Climate Change and Sustainability stories, and research and analysis has become more crucial than ever.

We showed researchers, analysts, and scientists how AWS services have reduced the level of technical skills required to fully use the GHCN_D dataset.

This GHCN-D is available on AWS. The details can be found on the Registry of Open Data on AWS. This data is available to researchers studying climate change and weather impacts.

This blog post demonstrated a typical workflow that a researcher could use to engage with and analyze this important data by using Amazon Athena, AWS Glue, and Amazon S3, and how they can visualize insights by using Amazon QuickSight.

By making this data available, Amazon has removed the heavy lifting that was traditionally required to work with the GHCN_D, thus expanding the opportunity for new research and new discoveries.

## About the Authors

**Joe Flasher is the Open Geospatial Data Lead at Amazon Web Services**, helping organizations most effectively make data available for analysis in the cloud. Joe has been working with geospatial data and open source projects for the past decade, both as a contributor and maintainer. He has been a member of the Landsat Advisory Group, and has worked on projects, ranging from building GIS software to making the space shuttle fly. Joe’s background is in astrophysics, but he kindly requests you don’t ask him any questions about constellations.

**Conor Delaney, PhD. is an environmental data scientist.**

TAGS: Amazon Athena, Amazon Quicksight