**An Analysis on Global Climate Data**

Mohanaa Tamilselvan and Sruthi Gatta

Department of Applied Sciences – San Jose State University

DATA 228 – Big Data Technologies and Application, Fall 2021

Dr. Andrew Bond

Copyright © 2021

Mohanaa Tamilselvan and Sruthi Gatta

ALL RIGHTS RESERVED

**Abstract**

**An Analysis on Global Climate Data**

**By Mohanaa Tamilselvan and Sruthi Gatta**

Across the world, the seasons are identified as 4 or even 3 in some countries based on the local climate the area experiences. As Earth keeps warming up, the seasons shrink or elaborate, be it Summers or Winters. Over the past few years, the very definition of seasons became questionable as the climate is not how it is expected to be in a certain season. The summers are warmer and sometimes are extended to supposedly a rainy season, where there is zero precipitation. Sometimes there’s an elaborated period of snowfall which is unexpected based on the definition of our seasons. The shift in these seasons can cause extreme weather events like wildfires, snowstorms etc. To estimate the consequence of a future disaster, it is important to gather historical data and real time data that would give an inside look of how climate has been changing over the years. It is also crucial to understand the changing patterns to promptly assess human behavior when dealing with nature. This project gives a peek inside the drastic changes identified in the climate over the years across the globe. It is vital to conduct global climate analyses to assess the effects of climate change on the Earth's natural capital and ecosystem resources. A comparative analysis of the various climate indicators over the years is the primary goal of this project. NOAA captures climate measurements from land-based stations all over the world, daily maximum and minimum temperatures, temperature at the time of observation, and temperature at the time of observation. This data is from NOAA and stored, processed, and analyzed using various AWS tools such as S3, Glue, Athena and visualize the data using Tableau.

*Keywords:* NOAA, Global climate, AWS, S3, Glue, Tableau, climate change

**Table of Contents**

[**1. Introduction** 6](#_Toc90322849)

[Github URL for project 6](#_Toc90322850)

[Project Need 6](#_Toc90322851)

[Project Goal 6](#_Toc90322852)

[**2. System Requirements** 7](#_Toc90322853)

[Tools Required 7](#_Toc90322854)

[**3. System Design** 7](#_Toc90322855)

[System Architecture Design 7](#_Toc90322856)

[**4. System Implementation** 8](#_Toc90322857)

[System Workflow 8](#_Toc90322858)

[Used Tools and Technologies 16](#_Toc90322859)

[**5. System Experiment** 16](#_Toc90322860)

[Exploratory Data Analysis 16](#_Toc90322861)

[Visualization 20](#_Toc90322862)

[**6. Conclusion** 27](#_Toc90322863)

[**References** 27](#_Toc90322864)

List of Figures

Figure 1. Architecture of the Project 7

Figure 2. S3 bucket and folders 8

Figure 3. Glue Table – tblallyears 8

Figure 4. Table Schema - tblallyears 9

Figure 5. Athena query editor of new database 9

Figure 6. tblallyears data in the DB-ghcnblog 10

Figure 7. tblallyears\_qa table in S3 10

Figure 8. ghcnd-stations.csv file uploaded in S3 11

Figure 9. Glue Table - tblghcnd\_station 12

Figure 10. Table schema of tblghcnd\_stations 13

Figure 11. tblghcnd\_stations data in the DB-ghcnblog 13

Figure 12. tblghcnd\_stations\_qa table in S3 13

Figure 13. Tables on ghcnblog database on Glue and Athena 14

Figure 14. Glue ghcnblog Crawler 14

Figure 15. Weather Station Growth ……………………………………………………. 23

Figure 16. Weather Stations in the US…………………………………………………. 23

Figure 17. Snowfall Trend……………………………………………………………... 24

Figure 18. Minimum Temperature Trend ……………………………………………….24

Figure 19. Maximum Temperature Trend ……………………………………………... 25

Figure 20. Minimum Temperatures per US states ………………………………………25

Figure 21. Maximum Temperatures per US states………………………………………26

Figure 22. Precipitation trend in Mountain View, CA ………………………………… 26

## **1. Introduction**

Our Dataset is from NOAA, the Global Historical Climatology Network - Daily, which is updated Daily on the observations from various land areas. The data is of CSV format, which consists of global historical data of over 175 years. It provides the station-based observations from land-based stations all around the world, almost two-thirds of which are just for precipitation measurement. Other climatic data includes, the daily maximum and lowest temperatures, the temperature at the moment of observation, snowfall, and snow depth. Our data shows us how the volume of data has grown over the years. For 1763 there are less than a thousand observations. For 2017 there are over 34 million observations. From the period of 1763 to 2018 there are over 2 billion observations.

### Github URL for project

<https://github.com/mohi-selvan/global-climate-data>

### Project Need

* Climate Analysis is essential for researchers to assess the effect of climate change on Earth
* There’s a need for high quality datasets, that can hold the essential parameters
* Climatic analysis on global data is often described to be complex and challenging to work with

### Project Goal

* Obtain and use the large dataset, which consists of millions of rows of climate data from various stations and observations.
* Explore data modeling and data processing to perform data analysis and visualization.
* Investigate and apply cloud services available on AWS to clean the data, store the data to perform Analysis, and draw fascinating insights with Tableau from the dataset

## **2. System Requirements**

### Tools Required

* Amazon Simple Storage Service (Amazon S3) to stage the data for analysis. The NOAA Global Climate data is known as the ‘**GHCN\_D**’ dataset, is stored in a bucket on Amazon S3. We use our private bucket **‘my22bucket’** 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 Data Catalog to catalog the data that we query with Athena.
* Tableau to build visualizations, perform ad hoc analyses, and get insights from the dataset. Queries and tables from Athena can be read into Tableau.

## **3. System Design**

### System Architecture Design

The architecture of our project involves the tools described above and Figure 1 represents the architecture of the project.

**Figure 1**

*Architecture of the Project*

Chart, diagram, funnel chart

Description automatically generated

## **4. System Implementation**

### System Workflow

To store the csv files and queries performed on Athena, we have created an S3 bucket **‘my228bucket’**, which has the following folders created as shown in the Figure 2 below.

* ghcnblog
* queries
* stations\_raw

**Figure 2**

*S3 bucket and folders*

Graphical user interface, text, application

Description automatically generated

The daily weather observation data is stored in files in a yearly format, starting from the year 1763 to the present year, which is named as ‘1763.csv’. This data can be accessed from the Amazon S3 bucket **s3://noaa-ghcn-pds/csv/**. From the AWS Glue console, we created a table manually, by the name ‘tblallyears’, and grouped it under the database ‘ghcnblog’. Added the data store with the path ‘s3://noaa-ghcn-pds/csv/’. Followed by the Serde parameters, which is ‘CSV’ file format and ‘Comma’ for the seperatorChar.

**Figure 3**

*Glue Table - tblallyears*

Graphical user interface, text, application

Description automatically generated

Figure 4 below shows the table schema for tblallyears, the following are the columns which are string variables.

* Id
* year\_date
* element
* data\_value
* m\_flag
* q\_flag
* s\_flag
* obs\_time

**Figure 4**

*Table Schema - tblallyears*

Graphical user interface

Description automatically generated with medium confidence

We can find the new database ‘ghcnblog’ in our AWS Athena dashboard as shown in Figure 5.

**Figure 5**

*Athena query editor of new database*

Graphical user interface, text, application

Description automatically generated

By clicking on the table ‘tblallyears’ and click on ‘Preview Table’ we are able to view a few rows of data loaded into the table as shown below in Figure 6.

**Figure 6**

*tblallyears data in the DB-ghcnblog*

Graphical user interface, text, application

Description automatically generated

To speed up the querying process we have created a table ‘tblallyears\_qa’ in our S3 bucket from Athena. The following query, from Figure 7 was used to create the table in S3 bucket, which stores data in a columnar, Parquet format, and this new table is also added to the database in where we work on Athena.

**Figure 7**

*tblallyears\_qa table in S3*

Graphical user interface, application, Word

Description automatically generated

Similarly, we are uploading the ghcnd-stations.csv file, as shown in the Figure 8 below, into our S3 bucket in the following location, **‘my228bucket/stations\_raw’**. The same process as followed for the **tblallyears** has been followed for this file as well as shown in the following figures 9-12.

**Figure 8**

*ghcnd-stations.csv file uploaded in S3*

Graphical user interface, application, table

Description automatically generated

**Figure 9**

*Glue Table - tblghcnd\_stations*

*Graphical user interface, text, application

Description automatically generated*

The below Figure 10 shows the table schema for tblghcnd\_stations, where these are the columns which are of string variables.

* Id
* Latitude
* Longitude
* Elevation
* State
* Name
* gsn\_flag
* hcn\_flag
* wmo\_id

**Figure 10**

*Table schema of tblghcnd\_stations*

***Graphical user interface

Description automatically generated with medium confidence***

**Figure 11**

*tblghcnd\_stations data in the DB-ghcnblog*

**A screenshot of a computer

Description automatically generated with medium confidence**

**Figure 12**

*tblghcnd\_stations\_qa table in S3*

*Graphical user interface, text, application, Word

Description automatically generated*

The image here in Figure 13, shows the tables created on AWS Glue and Athena under the database **ghcnblog**. Which are being used for our data analysis and visualizations via Tableau.

**Figure 13**

*Tables on* ***ghcnblog*** *database on Glue and Athena*

Graphical user interface, text, application, email

Description automatically generated

We have created a crawler on AWS Glue which will crawl the csv files being added to the S3 bucket to populate tables on Glue. The Crawler has been scheduled to run on every Tuesday at 8.15PM, as shown in the Figure 14.

**Figure 14**

*Glue* ***ghcnblog*** *Crawler*

Graphical user interface

Description automatically generated

### Used Tools and Technologies

Amazon S3, AWS Glue, Amazon Athena has been used here to create ethe platform where we are able to extract the required data and load them into the database. The data has been transformed to Parquet format while loading into the S3 bucket. We can use the Glue Jobs which can convert the data and store them on an automated basis as these jobs can be scheduled just like the Crawlers which are scheduled.

## **5. System Experiment**

Upon implementing the system where we could load and transform the data, we have used Amazon Athena for further analysis and Tableau for Visualizations and creating the dashboard from our project.

### Exploratory Data Analysis

As part of our exploratory data analysis and to perform visualization which will allow us to gain insights, we have created additional tables as shown here.

***Precipitation Table***

***Text

Description automatically generated***

***Chart

Description automatically generated with low confidence***

***Minimum Temperature***

Graphical user interface, text, application

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

***Maximum Temperature***

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

***Growing Degree Days Table***

Graphical user interface, text, application, email

Description automatically generated

A picture containing table

Description automatically generated

***Table for Every 10th Year***

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 10th year from 1831 to 2021.

Graphical user interface, text, application, Word

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

As part of EDA, we queried a few tables to get the numbers on the stations, observations and the average max & min Celsius, and average mm of rainfall, average mm of snowfall and snow depth, on Earth since the time of the dataset, since 1763. In the querying the String values are converted to Real variables to do the necessary calculations for querying the weather elements.

***Total Number of Observations***

Graphical user interface, text, application, Word

Description automatically generated

Graphical user interface, text, application, website

Description automatically generated

***Total Number of Stations***

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated with medium confidence

***Average value of the various elements***

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated with medium confidence

### Visualization

To perform analysis, we have connected Tableau to Amazon Athena Server. We performed queries to obtain tables necessary for our project analysis. One such table required for our analysis to understand the trend of temperatures, snowfall, precipitation for every 10 years. We queried the data from our S3 bucket that hold our data as parquet files as shown in the figure *Decennial Climate data*. Figure *Mountain View, California weather station data* shows the query to create a table that holds the data recorded by the Mountain View, California weather station over the years.

As part of exploratory data analysis, we studied the weather station growth in the world from the year 1851 to 2021. Figure 15 the trend of growing number of weather stations from the year 1851 till date. We could notice that at the start of the 20th century, there have been new weather stations set up. Figure 16 shows the number weather stations set up in the US per a period of 30 years. The circle represents the number of weather stations per state in the US, showing that Colorado currently has the highest number of weather stations followed by California.

Now as part of our climate analysis, we have considered the snowfall in mm from the year 1851 and showing the trend of snowfall over the world at a period of every ten years. As seen in figure 17 there is very little variation in the snowfall recorded in every decade, we could notice a drop in the year 2021 representing warmer temperatures recorded. The figures 18 and 19 shows the trend average minimum and maximum temperatures recorded in the world from the years 1851 till date. Figures 20 and 21 shows the average temperatures in the US per each state for the year 2016. As the graph, an interactive map with year filters show that Alaska state has the least minimum and maximum temperature recorded for the year 2016 while Florida recorded the highest minimum and maximum temperatures.

For our final analysis, we considered the Mountain View, California weather station to understand the precipitation recorded over the years. As shown in the figure 22, the graph shows that Mountain View has been dry over many years and the highest rainfall recorded is 3.365 millimeters. With these analyses we could plot and understand the rainfall, snowfall, minimum and maximum temperature indicators of the climate over the world from past few centuries.

***Decennial Climate data***

Timeline

Description automatically generated with medium confidence

***Mountain View, California weather station data***

Table

Description automatically generated with low confidence

**Figure 15**

***Weather Station Growth***

Chart, line chart

Description automatically generated

**Figure 16**

***Weather Stations in the US***

Map

Description automatically generated

**Figure 17**

***Snowfall Trend***

Chart, line chart

Description automatically generated

**Figure 18**

***Minimum Temperature Trend***

Chart

Description automatically generated

**Figure 19**

***Maximum Temperature Trend***

Chart, bar chart

Description automatically generated

**Figure 20**

***Minimum Temperatures per US states***

Map

Description automatically generated

**Figure 21**

***Maximum Temperatures per US states***

Map

Description automatically generated

**Figure 22**

***Precipitation trend in Mountain View, CA***

Chart, bar chart

Description automatically generated

## **6. Conclusion**

To conclude, we have used AWS big data tools to retrieve, store and transform hundreds of GBs of data. We performed analysis on the global climate data to understand the patterns of parameters like snowfall, precipitation, maximum and minimum temperatures. From our analyses we have identified that snowfall has reduced over periods of time indicating raise in temperatures. The maximum and minimum temperatures have rose over the years, showing the climate change scenario in the world. Apart from analyzing the data, with our project, we have created a platform where researchers, scientists and user could plugin their application to Athena data lake and perform their analyses.

## **References**

<https://registry.opendata.aws/noaa-ghcn/>

<https://github.com/awslabs/open-data-docs/tree/main/docs/noaa/noaa-ghcn>