Foreword

Global Historical Climate Network Daily (GHCND) is an integrated database that contains daily climate summaries from weather stations across the world. It contains records for the last 260 years contributed by more than 100,000 land surfaced stations in 180 countries and territories around the world.

Processing

The data is in two parts. The first part of the data is the collection of daily climate summaries that are grouped into individual files by year and stored in a gzip compressed csv format. The second part contains metadata specific to the individual stations, countries, states, and inventory.

The data is structured as shown below.

```
|-Countries
|-Daily
|    |-1763.csv.gz
|    |-1764.csv.gz
|    |-……
|    |-2022.csv.gz
|-Inventory
|-States
|-Stations
```

Here the stations, states, countries, and inventory are the files whereas the daily is the directory which contains one file for each year and in total there are 260 years.

The schema is defined already for the daily. Every file in daily has the same schema and head as expected. The daily file had to be piped through gunzip and then the decompressed csv data was piped to head in order to see the head of daily.

The size of the total data is 15.7 GB. The daily data contributes to most of this size whereas the metadata only occupies 40.6 MB. The size of the daily individual files increases as the years pass through due to the inclusion of more stations. Also, since the daily data is compressed, the actual size of the data might be significantly higher when uncompressed. When considered the gzip compression ratio of 10.1 for csv files that contains text, the data size might increase to anywhere around 100 GB to 1 TB.

Based on the description given in the project, the schema has been defined for daily and 1000 rows have been loaded to check the data. It has been found that though the DATE and OBSERVATION TIME columns have their respected values (i.e., date, time), the data types were not accurate to parse the data automatically. They were expected to be 'DateType' and 'TimestampType' to be accurate but have been loaded as 'StringType' initially for the data loading purpose. These columns will be converted to 'DateType' and 'TimestampType' as needed to answer any other tasks.

Since the data present in countries, states, stations, inventory is in fixed width text format, a schema could not be applied while these files were read from HDFS as the columns can only be extracted by choosing the relevant substring of the text record. So, each file was loaded into separate dataframe with only one text column. Then the relevant substrings were extracted using the select statement and appropriate data types were applied from the schema based on the column names.

The stations table contains 118493 rows in total representing each row with a unique station and station ID out of which 110400 do not have WMOID. The states table contains 74 records in total where every record provides the name and two letter code of each state. The countries table consists of 219 distinct countries or territories from across the globe which are presented in 219 rows in total. The inventory table consists of 704963 rows in total and provides information about the periods of record for each station and element (i.e., TMIN, TMAX etc.).

An output directory has been created (*"hdfs:///user/ana146/ghcnd/"*) to store any outputs from the tasks performed in this project. Since the daily climate summaries data is too big to perform joins to the metadata tables (stations, staes, countries, and inventory), the metadata tables have been joined into a single table for the convenience of sort and filter the attributes at station level.

So, to allow the joining tasks, firstly the two-digit country code has been extracted from the station ID in the stations table and stored as a separate column (COUNTRYCODE) by using 'withColumn' and 'substring' functions. The column names in countries and states tables have been renamed to avoid the ambiguity. Then based on this country codes the countries table has been left joined to the stations table. The states table has been filtered for only US states then left joined to the table combined with stations and countries data.

Using the FIRSTYEAR, LASTYEAR columns in Inventory table, the first and last years that each station was active were determined. This was achieved by grouping the data by station ID and aggregating FIRSTYEAR and LASTYEAR columns by using min and max functions. Then this data has been left joined to the stations data combined with countries and states. This final data set allows us to drill through the data at individual station, state, country, and element levels. Based on this final stations data, three calculated columns have been created to count the number of elements each station collected (DISTINCTELEMENT), number of core elements each station collected (COUNTCORE), and number of non-core elements each station collected (COUNTNONCORE). It has been found that a total of 20289 stations collect all five core elements and a total of 116455 stations collected only precipitation. The final stations dataset with all the added columns has been saved to output directory (*"stations.csv.gz"*) in compressed gzip format to achieve the consistency with other files.

The stations data has been left joined to the subset of the daily data with a limit of 1000 rows to find out if there are any stations present in the daily data that are not in stations at all. The result returned to be zero, this could be due to the daily subset of 1000 rows have been taken randomly from daily data and could have missed any stations that might not be in stations data. If all the daily data is used to left join, the amount of time and resources spent will be higher and there is a possibility of left join not completing properly due to daily data contains more than 100k distinct station ID values which could skew the data. Another way to determine if there are any stations in daily that are not there in stations is by extracting the distinct values of station ID in daily and match it with the station IDs in stations table using distinct and regex functions.

Analysis

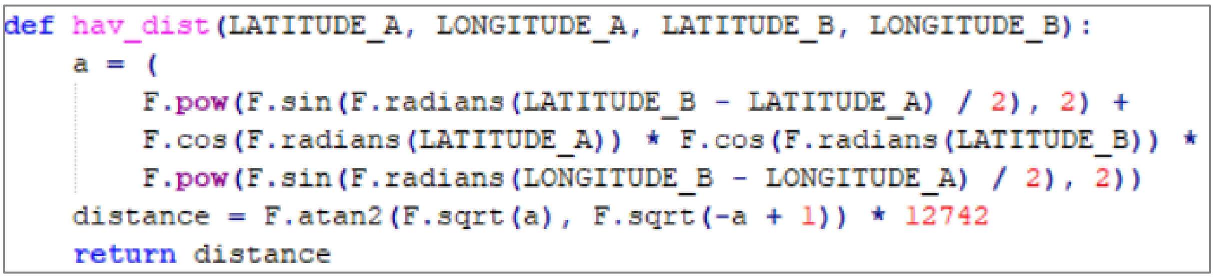
To perform analysis on the stations dataset, the resources have been increased to 4 executors, 2 cores per executor, 4 GB of executor memory, and 4 GB of master memory whenever the daily climate summaries were processed. This reduced the time taken for each task and improved the overall efficiency of Spark shell.

There are 118493 stations in total out which 38284 stations were active in year 2021. There are 991 stations GCOS surface network (GSN), 1218 stations in US Historical climatology network (HCN), and none in Climate Reference Network (CRN). Out of these 14 stations are both in GSN and HCN networks.

Using the stations data, the total number of countries have been calculated by grouping the country codes and counting the stations and saved this output to the countries table. The same has been done to the states table by counting the number of stations present in each state and added the output to the states table. Both files are saved to the output directory.

To find the stations in Northern hemisphere only for the territories of United States excluding the US, the stations dataset has been filtered for the territories such as Palmyra Atoll [United States], Johnston Atoll [United States], Wake Island [United States], Virgin Islands [United States], Puerto Rico [United States], American Samoa [United States], Midway Islands [United States}, and Northern Mariana Islands [United States]. There are 318 stations in total present in these territories.

A function has been created to compute the geographical distance between two stations using the longitude and latitude values.



```
def hav_dist(LATITUDE_A, LONGITUDE_A, LATITUDE_B, LONGITUDE_B):
    a = (
        F.pow(F.sin(F.radians(LATITUDE_B - LATITUDE_A) / 2), 2) +
        F.cos(F.radians(LATITUDE_A)) * F.cos(F.radians(LATITUDE_B)) *
        F.pow(F.sin(F.radians(LONGITUDE_B - LONGITUDE_A) / 2), 2))
    distance = F.atan2(F.sqrt(a), F.sqrt(-a + 1)) * 12742
    return distance
```

This function is defined using Haversine method which assumes a spherical earth ignoring ellipsoidal effects to compute the distance between two points. This function takes longitude and latitude values of the given stations as spatial point inputs and takes the earth diameter as 12742km to compute the distance between two points in kilometres.

A sample data of 5 rows from stations data is taken to test this function. Performed cross join on this subset to get two stations in each row of which the output has been tested on the above function to determine the quality of results produced by the function. This worked as expected and calculated the distance between two stations in kilometres.

To find out the pairwise distances for New Zealand stations, the stations data was filtered to get the NZ stations data only and then a cross join was performed to get the pairs of stations in one row. Using the function above the distances have been calculated and it has been found that Paraparaumu and Wellington AERO stations are the closest in New Zealand with only 52 kilometres apart. These computed distances for all of the New Zealand stations is stored to the output directory (*"NZStationPairs.csv.gz")*

To explore the climate summaries data in HDFS, commands such as hdfs dfs -ls and hdfs dfs -du are used. It has been identified that the daily data for 2022 is comparatively very less than other years with only 5971307 records. This could be due to the year 2022 is still current. Also, the file sizes have been increased over the years compared to the beginning years. This could be due to the increase in number of stations.

Only 1 block is required for the daily climate summaries 2022 data and the size of this block is 24.7MB. Similarly, 2 blocks are required for the year 2021 with the overall size of blocks being 140.12MB (1st block – 128MB, 2nd block – 12.12MB). This indicates that Spark can load and apply transformations in parallel for both the years 2022 and 2021.

When the climate summaries data for years 2022, 2021, and 2014-2022 were loaded and counted, below number of stages and tasks were executed.

Year 2022 – Total of 4 stages and 1 task for each stage
Year2021 - Total of 4 stages and 1 task for each stage
2014-2022 – Total of 4 stages and 12 tasks are executed

This shows that the number of blocks does not correspond to the number of tasks executed. For example, Year 2022 has only 1 block but still took 4 tasks to load and count the data. This is same with year 2021 but there are 2 blocks in 2021.
When the data is loaded for the years between 2014-2022 together, the number of tasks has been increased to count the data (10 tasks – 9 to shuffle write and 1 to shuffle read). This could be due to Spark writing the data individually for the 9 years between 2014-2022. However, to load the data Spark can process single or multiple datasets in the same number of stages and tasks.

Though the compressed data takes less space, in the case of parallel distributed data processing, Spark does not know how to split the compressed data for better parallelization. When loading and applying transformations to all daily data, Spark could run either 260 tasks or slightly less as it could combine the smaller multiple years into a single large partition.

There are 3000243596 rows in total in all daily data. Since it is too big to use while developing the code, a small random subset of daily has been used to develop and test the code to answer some of the tasks. Once the code was fully developed and tested, it has been applied on the whole daily dataset.

The below table shows the number of observations in daily for each of the core elements.

Element	Total Count
SNOW	341985067
SNWD	289981374
PRCP	1043785667
TMIN	444271327
TMAX	445623712

The element Precipitation (PRCP) has more observations than other core elements.

It has been found that 8808805 observations have collected TMIN but do not have the corresponding TMAX observation. A total of 27650 stations have contributed to these observations. This has been computed by filtering all the observations from daily for TMIN and TMAX then grouped by station ID and date where ‘collect_set’ function is used to find the stations with TMIN and corresponding TMAX observations. Then the stations data has been left joined with daily to filter the TMIN and TMAX observations for all the stations in New Zealand. There are 472271 observations in total collected from 1940 until 2021. This data set has been saved to the output directory. This file has been also copied to local directory and the number of observations is same when counted in Spark and in local directory. The average rainfall for each year and each country has been computed and the output has been saved to the output directory.