Processing:

Data Structure & Metadata

There is 263 years of daily weather data stored in GHCN. It it structured similarly as figure 1, except each branch of daily is an individual year. The data spans from 1750 to the current year of 2023. There are some gaps, not every single year has a file of weather data, but the gaps are contained to the earliest years of recording. Each file contains 1 year worth of data, gzipped for maximum compression and minimal storage. The files for the oldest years are small, with the file size increasing as time goes on.

```
                                              |----- Inventory

HDFS ------------|GHCND-----------|-----Countries

                                              |------States

                                              |------ Stations                |------ 1750

                                              |------- Daily ----------------|------1833

                                                                              |-------1972

                                                                              |-------......n
```

Figure 1: General structure of GHCN data, each daily branch is a different year

The stored GHCN data uses a total of 98 G of space. Each of the files of metadata regarding station, state, country and inventory details are relatively small text files, ranging from the smallest file, states, using of 1.1 Kb (0.0000011 GB), to the largest file, inventory using 32.4 Mb (0.0324 GB). Combined, all the metadata files use approximately 0.4 G of space. The folder containing all of the daily data is the vast majority of the GHCND data, filling the other 97.6 G of space. The inventory metadata is the largest, and the states metadata is the smallest as seen in table 1.

Visual inspection of the daily data for 2023 displayed many null entries in the various measurement flags (GSN, HCN and WMO), and most observation times. Most measurement data is expected to not have an observation time, as the majority of the elements measured (core elements) are total or average for the observation date. The GHCND "readme" text file explains that the measurement flags are indications of problems with the recording, or indicators about to which day of the month the data was recorded.

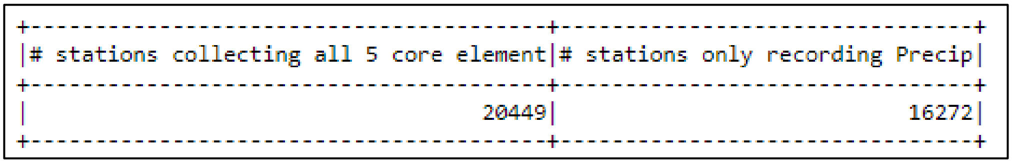
Table 1 Row count of each metadata file, and count of stations that do and do not have World Meteorological Organization (WMO) ID's

Item Counted	Row count
Inventory	737925
Countries	219
States	74
Stations	124247
Stations without WMO ID	116288
Stations with WMO ID	7959

Data Manipulation

The metadata was combined to match all data to the station it belonged to. The result was a table of unique station ID's and the corresponding location (coordinates, country ID's, etc), a complete list of all of the elements each station records, and the first and last year that station recorded unflagged data. The aggregated data was then used to find additional information about each station, such as the total number of different elements a station has recorded, and how many of those elements were "core elements" (precipitation, snowfall, snow depth, maximum temperature, and minimum temperature), and how many non-core elements were recorded at that station. Table 2 displays that 20,449 stations were observed to have collected all 5 core elements. 16,272 stations were observed to only record precipitation data. This master table has been saved to the outputs directory.

Table 2 Count of stations that collect all 5 core elements and that collect only precipitation



```
+----------------------------------------+--------------------------------+
|# stations collecting all 5 core element|# stations only recording Precip|
+----------------------------------------+--------------------------------+
|                                   20449|                           16272|
+----------------------------------------+--------------------------------+
```

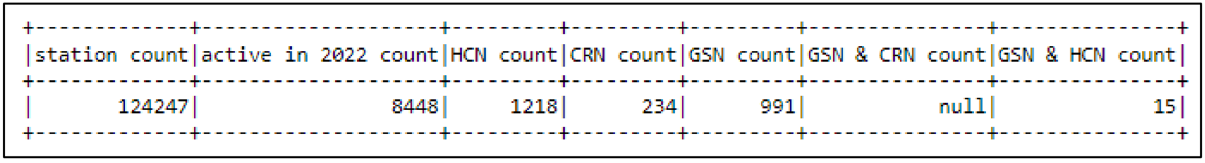
This master metadata table was broadcasted onto subset of 1,000 rows of daily data from 2023 to check for consistency. All stations in the subset of daily data were found to corresponding metadata in the master table. The master metadata table is a small enough file to be broadcasted onto each node without consuming too much of that node's local memory, which makes joining the master table onto daily data more efficient than joining the daily data to the master table would be. Additionally, each unique station has 1 row in the master table, while the daily data has multiple entries for the same station which would cause extra computation and expense than is necessary to achieve the goal of affiliating the metadata with each daily entry.

There are several options for exploring if there is any data in daily that the master table does not cover. The first and most common is a left join, as used here, followed by a filter for empty columns where the master table should have populated data. Another option could be alternative joins, such as an inner join that only keeps the data that matched successfully, and comparing if the row count of daily-joined is the same as daily not-joined. If there is a discrepancy that would indicate that not every instance of daily was matched to a station in the master list.

Station Data

As derived from the master table of station data, it was found that there are 124,247 represented in this dataset. 8,448 of those stations recorded at least one observation in 2022. The GCOS Surface Network (GSN) has 991 stations in this dataset associated with it, and the US Historical Climatology Network (HCN) has 1,218 stations in it's network, 15 of which are also apart of the GSN. The US Climate Reference Network (CRN) has 234 stations. The HCN and CRN networks are indicated in the same field and cannot co-occur in this dataset. A summary of the station information listed can be seen in table 3.

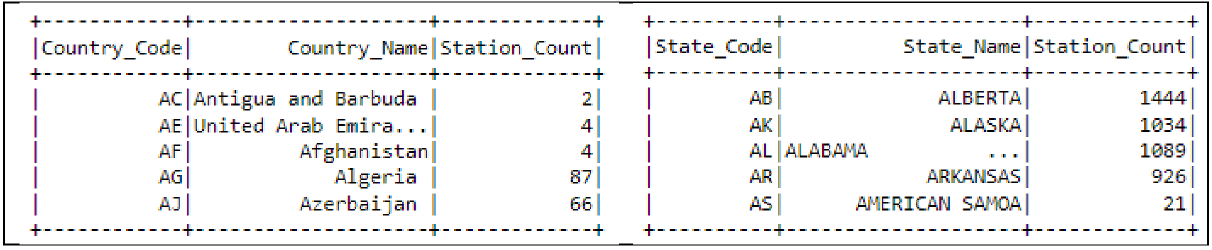
Table 3 Total count of stations, stations active in 2022, and count of various network flags



```
+-------------+--------------------+---------+---------+---------+---------------+---------------+
|station count|active in 2022 count|HCN count|CRN count|GSN count|GSN & CRN count|GSN & HCN count|
+-------------+--------------------+---------+---------+---------+---------------+---------------+
|       124247|                8448|     1218|      234|      991|           null|             15|
+-------------+--------------------+---------+---------+---------+---------------+---------------+
```

Tables with sums of how many stations are present in each state (US only) and each country, such as the examples shown in table 4, have been saved to the output directory.

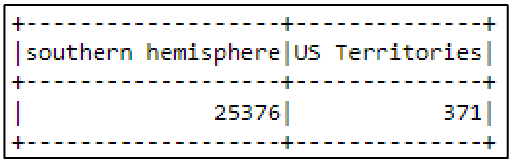
Table 4 Samples of saved data tables with count of stations per country (left) and state (right)



```
+------------+-------------------+-------------+   +----------+------------------+-------------+
|Country_Code|       Country_Name|Station_Count|   |State_Code|        State_Name|Station_Count|
+------------+-------------------+-------------+   +----------+------------------+-------------+
|          AC|Antigua and Barbuda |            2|   |        AB|           ALBERTA|         1444|
|          AE|United Arab Emira...|            4|   |        AK|            ALASKA|         1034|
|          AF|         Afghanistan|            4|   |        AL|ALABAMA        ...|         1089|
|          AG|            Algeria |           87|   |        AR|          ARKANSAS|          926|
|          AJ|         Azerbaijan |           66|   |        AS|    AMERICAN SAMOA|           21|
+------------+-------------------+-------------+   +----------+------------------+-------------+
```

The master table can be filtered for stations located at a latitude below 0° to find the count of stations located in the southern hemisphere. Similarly, locating stations in United States territories exclusively can be found using the fact that territories are denoted using square brackets. There are 25,376 stations located in the southern hemisphere, and 371 stations in exclusively US territories (Table 5).

Table 5 Count of stations in the southern hemisphere and US territories

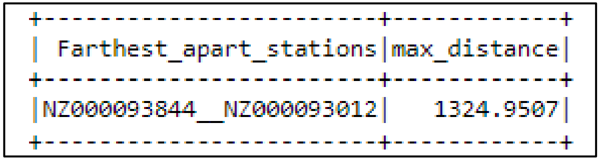


```
+-------------------+--------------+
|southern hemisphere|US Territories|
+-------------------+--------------+
|              25376|           371|
+-------------------+--------------+
```

Pairwise distances

The haversine formula can be used to find the distance between two geographic points while taking into account the curvature of the earth. Using a python friendly version of this by Michael0x2a, as the haversine module of python was not installing properly, a table with the pairwise distances between all stations in New Zealand was generated and saved to the output directory. The two stations farthest from each other in New Zealand are NZ000093844 & NZ000093012, being 1,324.95 km apart as shown in table 6. It might be better in the future to utilize station names rather than station ID's for similar comparisons.

Table 6 Station ID's of the two stations farthest from each other in New Zealand and the distance away from each other they are in km



```
+-----------------------+------------+
| Farthest_apart_stations|max_distance|
+-----------------------+------------+
|NZ000093844__NZ000093012|   1324.9507|
+-----------------------+------------+
```

HDFS Transformations, Blocking and Partitions

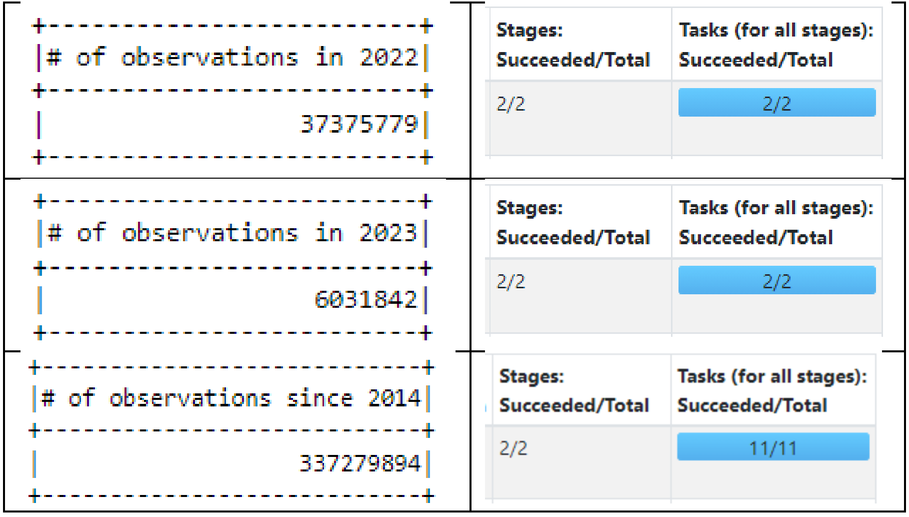
HDFS has a default, fixed block size of 128 mb, or 134217728 B. Blocks are HDFS's built-in function to distribute data uniformly. The daily file for 2023 has an average block size of 27521531 B, so it will fit into 1 HDFS block. The daily file for 2022, however, has a block size of 166075423 B, which is large than 1 HDFS block. HDFS splits the 2022 file up into two blocks averaging 83037711 B.

Given that the 2023 data only requires 1 block, this block can be loaded onto all available nodes and run in parallel, however it is unnecessary. All of the transformations can be performed locally, eliminating the need for any shuffling. The 2022 data Spark can run transformations on block 1 and block 2 in parallel before shuffling and mapping the results back together.

The 2022 daily data has 37,375,779 rows of observations, and the 2023 daily data has only 6,031,842 observations. Despite the large difference between the size of both of these files, and the n umber of blocks for both being different, both counts were completed in 2 stages and 2 tasks total – although the execution time was faster for the 2023 data. Having more blocks did not affect the number of tasks that had to be performed for the 2022 daily data.

When 10 years of data, from 2014 to 2023, is counted, the task is still completed in only two stages. However, the map and reduce is now completed in 11 tasks as seen in table 7. Stage 1 is comprised of 10 tasks, one task per year, or rather- per file- of data read and counted.

Table 7 comparison of stages and tasks for different file sizes, and different number of files evaluated



```
+-------------------------+
|# of observations in 2022|
+-------------------------+
|                 37375779|
+-------------------------+
```

Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
2/2	2/2

```
+-------------------------+
|# of observations in 2023|
+-------------------------+
|                  6031842|
+-------------------------+
```

Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
2/2	2/2

```
+--------------------------+
|# of observations since 2014|
+--------------------------+
|                 337279894|
+--------------------------+
```

Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
2/2	11/11

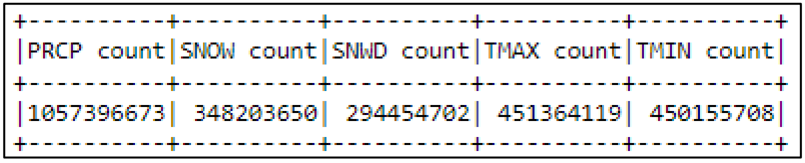
Each yearly file of daily data is compressed with .gzip. gzip is a compression format that is not splittable, the whole file needs to be opened at once to access the contents. This means that each node in the cluster is forced to operate on 1 and only 1 file. This can mean that the node possibly is idle while other nodes are still processing, it also means that the node has to uncompress the entire file and can't delegate any of the file to other nodes if the size of the file is too large when uncompressed. Gzip is a very efficient way to compress files, but compressing with gzip means that the files should be split, in a logical manner, prior to compression. If data being evaluated has a component of time, that is a reliable metric to chunk data on.

In the current file format, to load and transform the whole folder of daily data would require at leats 264 tasks – one task for each of the 263 files and at least one to map, if not more. As many tasks as there are cores in use can run in parallel, so if 16 cores were in use to open and transform all of daily, 16 tasks would run in parallel at a time and it would take appriximately 16 "rounds" of completed tasks across all nodes. However, this does not account for the size of each file, causing some tasks will exceute faster or slower than others on a node. The most efficient composition of the data would be for all sizes to be roughly the same size. Perhaps larger files could be split into 6 month chunks, and smaller files could be combined into one gzip file.

Daily Temperature

In the entire folder of daily data, there are 3,064,620,240 rows of observation data. Of these observations, precipitation was recorded the most frequently with 1,057,396,673 observations. Precipitation had the most observations, followed by maximum temperature, minimum temperature, snowfall, and finally snow depth in descending order. The exact count of observations for each core element can be seen in table 8.

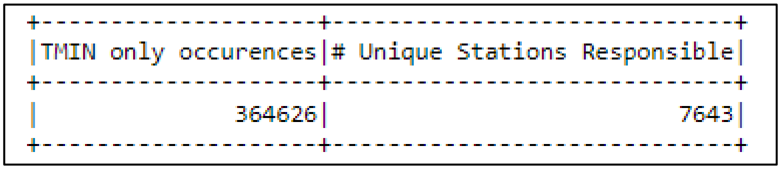
Table 8 count of total number of observations for each core element



```
+----------+----------+----------+----------+----------+
|PRCP count|SNOW count|SNWD count|TMAX count|TMIN count|
+----------+----------+----------+----------+----------+
|1057396673| 348203650| 294454702| 451364119| 450155708|
+----------+----------+----------+----------+----------+
```

Table 9 shows that 7,643 stations were found to have at least once occurance of recording the minimum temperature but not the maximum temperature for a given day, which occurred 364,626 times. If averaged across all stations that contributed to this discrepancy, each station would have had to record minimum temperature without also recording maximum temperature around 47 times. It is more likely, however, that a number of stations consistently did not record maximum temperature despite recording minimum temperature, and the other stations had less consisitent discrepancies.

Table 9 count of occurrences of only minimum temperature being recorded and how many stations contributed



```
+-------------------+----------------------------+
|TMIN only occurences|# Unique Stations Responsible|
+-------------------+----------------------------+
|             364626|                        7643|
+-------------------+----------------------------+
```

A table listing all of the observations of minimum and maximum temperature throughout new zealand was saved to the output directory. This table was condensed to include only the station ID and name, which temperature (min or max) was recorded, the value of those temperatures, and the date of the observation to be space efficient. Additional station metadata, such as location coordinate of the station, is not readily availible on this table.

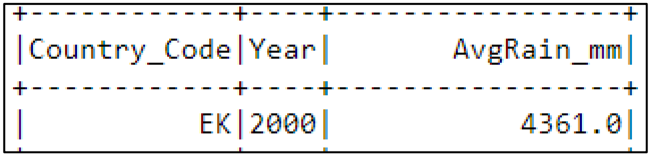
In the GHCN data, weather observation recordings for New Zealand began 83 years ago in 1940. There may be additional weather data observed prior to 1940 for New Zealand availiable from other data sources than the GHCN. Since the first observation of New Zealand weather available in the GHCN data, there have been 478,712 observations recorded, averaging around 5,767 recordings per year if the recordings were evenly distributed across all years, which is unlikely. Given new weather recording technologies since 1940, it is likely that more recent years contain much more data than older years, as evidenced by storage size of each year of data increasing drastically over time. The daily observation file for 1940 across all stations is 58.1 M, while the file for all stations for the last full year of observations (2022) is almost three times as large, at 158.4 M. To confirm the number of observations recorded in New zealand since 1940 the output table was transferred to the local directory and unzipped and counted using the “zcat” and “| wc -l” bash commands, which positively confirmed the row count.

Temperature maximums and minimuns were isolated for each station across New Zealand and plotted across the 83 years of activity can be seen in appendix A files in the supplimentary material . Due to the scale of the time aspect, finer details of each year are lost. Three stations, Auckland Aerodome, Wellingtom Aerodome, and Kaikoura, were found to have begun recording data after the year 2000. Three stations were found to have not recorded any data since before 2010. Four stations had extended periods of no recorded observations, with two of those four having had two extended periods of no recorded observations. Minimum and maximum temperatures were averaged across all recordings on each date throughout New Zealand to plot the average minimum and maximum temperatures for new zealand as a whole. There appears to have been a trend from higher temperatures to lower temperatures in the first 10 years of recorded data, however it is worth noting that only 2 stations recorded data between 1940 and 1950, and both stations were located on remote islands, not on New Zealand proper.

Average Rainfall

The average precipitation for each country across all stations within that country was calculated for each year of data and plotted and can be seen in appendix B files in the supplementary material. The country that recorded the highest amount of rainfall in any single year was Equatorial Guinea, with a rainfall of 4,361 mm in the year 2000 (table 10). According to the UNDP website, Equatorial Guinea experiences heavy rainfall due to frequent and heavy monsoons, so Equatorial Guinea having the highest rainfall is not out of the realm of reasonable doubt (Equatorial Guinea, n.d.). The quantity of rainfall is the aspect that raises concern. As seen on the world plot fig, the majority of the average rainfall is below about 300 mm, so the recorded temperature of 4,361 mm is extremely outlying and likely indicates the possibility of a measurement error.

Table 10 country code and rainfall total of the largest recorded average rainfall in a year



```
+------------+----+-----------------+
|Country_Code|Year|       AvgRain_mm|
+------------+----+-----------------+
|          EK|2000|           4361.0|
```

This is consistent with general trend of higher rainfall in equatorial areas, however it might be expected that locations such as the Amazon Rainforest, which are known for heavy rainfall, to have the highest rainfall in a year. It is worth noting that since these are average measurements across all stations in a country, the number produced cannot take into account localized events if there are many other stations that counterbalance that reading, or vice versa. For example, Equatorial Guinea has only two stations, so the average rainfall in a country with very few stations, such as Equatorial Guinea, is more likely to not be representative.

The average rainfall per year has been plotted on a world map, as well as local continental maps. Each map has a PDF in the supplementary materials that is high resolution and areas of interest can be zoomed in to. Several countries are plotted with no rainfall data, although there are stations located in them. It is possible that the stations in those countries do not record rainfall data, or that the rainfall data is not shared to GHCN.

References (Report):

Equatorial Guinea. (n.d.). UNDP Climate Change Adaptation. Retrieved April 28, 2023, from https://www.adaptation-undp.org/explore/africa/equatorial-guinea#:~:text=The%20main%20wet%20season%20lasts,moist%20air%20from%20the%20ocean.

References (Code):

Analysis Q4e:

GeeksforGeeks. (2021, June 10). How to add a column from another DataFrame in Pandas? [Blog post]. GeeksforGeeks. Retrieved April 28, 2023, from https://www.geeksforgeeks.org/how-to-add-column-from-another-dataframe-in-pandas/

Geopandas. (2022, January 31). Missing and Empty Geometries — GeoPandas 0.9.0 documentation. Retrieved from https://geopandas.org/en/stable/docs/user_guide/missing_empty.html

Geopandas. (2022, January 31). Creating a GeoDataFrame from a DataFrame with coordinates — GeoPandas 0.9.0 documentation. Retrieved from https://geopandas.org/en/stable/gallery/create_geopandas_from_pandas.html

Matplotlib. (n.d.). Legend guide — Matplotlib documentation. Retrieved from https://matplotlib.org/stable/tutorials/intermediate/legend_guide.html

Shapely deprecation warning message when plotting GeoPandas GeoDataFrame. (2021, June 23). GIS Stack Exchange. Retrieved from https://gis.stackexchange.com/questions/420046/shapely-deprecation-warning-message-when-plotting-geopandas-geodataframe

Overlapping legend: how to put GeoPandas legend next to the map. (2020, June 11). GIS Stack Exchange. Retrieved from https://gis.stackexchange.com/questions/371979/overlapping-legend-how-to-put-geopandas-legend-next-to-the-map

Gómez, R. (2017, May 23). Geopandas choropleth. Ramiro.org. Retrieved from https://ramiro.org/notebook/geopandas-choropleth/

Country Codes. (n.d.). Worlddata.info. Retrieved from https://www.worlddata.info/countrycodes.php

nkmk. (2018, Oct 28). Python: textwrap - Wrap, fill and shorten texts [Blog post]. Note.nkmk.me. Retrieved from https://note.nkmk.me/en/python-textwrap-wrap-fill-shorten/

Analysis Q2:

Michael0x2a. [Michael0x2a]. (2013, Oct 17). Getting distance between two points based on latitude/longitude [Answer to the question]. In Stack Overflow. https://stackoverflow.com/questions/19412462/getting-distance-between-two-points-based-on-latitude-longitude

ChatGPT/BingChat Use:

- Analysis Q3b (file retrieval name issues)
- Analysis Q4d (read rows in directory, ”zcat … *.gz” specifically)
- Analysis Q4d (date range for x axis index for shared axis)
- Asked a few times if an aggregation was the most efficient way of doing it
- All citations