DATA420-25S2(C)

Assignment 1

GHCN Data Analysis using Spark

David Ewing (82171165)

2025-09-16 02:35

Author, 2025

1 Background

In this paragraph, you should give a brief overview of what you are doing in the assignment, including any useful links or references to background material and a high-level description of any difficulties that you had.

2 Processing

In this section, I explore the structure and content of the Global Historical Climatology Network Daily (GHCN-Daily) datasets stored in Azure Blob Storage. The datasets include both the daily climate observations and a set of metadata tables, all of which are necessary to conduct subsequent analysis and visualisation. The focus here is on understanding how the data is organised, verifying the schema, and preparing enriched tables for later use.

2.1 Uniform Resource Locator

Access to the Azure Blob Storage structure is via a structure URL. As an example,

wasbs://campus-data@madsstorage002.blob.core.windows.net/ghcnd/daily/2025.csv.gz

The URL structure is broken down in Table 1,

Component	Type	Description
`wasbs://`	Protocol	Like `file://` or `https://`
`campus-data`	Container	Top-level shared folder
`madsstorage002`	Storage account	Network or disk volume name
`blob.core.windows.net`	Service domain	Blob service endpoint
`/ghcnd/daily/2025.csv.gz`	Path	Blob path
`ghcnd`	Container name	`<container>`
`daily`	Directory	Optional folder(s) within container
`2025.csv.gz`	Blob name	`<filename>`

Table 1: URL structure of the GHCND Azure Blob Storage

The developer would use Hadoop commands to access the data:

- `hdfs dfs -ls wasbs://...` the ghcn files and subfolders were found, and with
- `hdfs dfs -du wasbs://...` of the size of the files were found.

.

2.2 Dataset Structure

The storage container consists of five items: one directory containing the daily climate observations and four fixed-width metadata text files covering stations, countries, states, and the inventory of elements. The overall structure is illustrated in Figure 1. Table 2 shows the sizes. The daily directory contains compressed files, with each file corresponding to one year in `csv.gz` format. In contrast, the metadata tables are small, uncompressed text files.

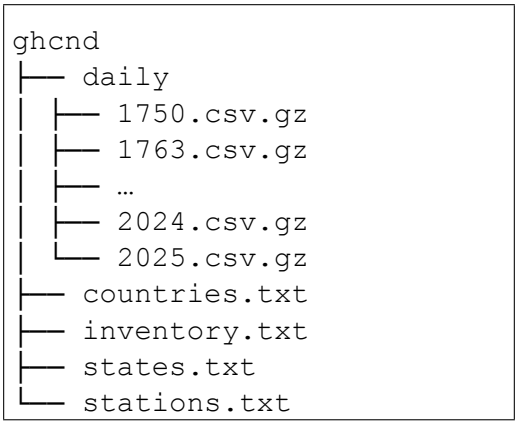

Figure 1: GHCND Azure Blob Storage Directory Structure

The daily files are stored in a compressed `−csv.gz` file format. The gzip compression algorithm, widely used for compressing text-based files such as CSVs, leverages the DEFLATE algorithm to achieve efficient data size reduction. Depending on the data's redundancy, gzip typically yields compression ratios ranging from 2:1 to 5:1 for CSV files, reducing file sizes to 20–50% of their original size (Deutsch, 1996). This efficiency makes gzip a popular choice for managing large datasets in cloud environments, such as Azure, where storage and transfer optimisation are critical. The total size of the GHCND dataset is shown in Table 2, including an estimated uncompressed size if the `..csv.gz` file format was not used.

Table 2: Summary of dataset sizes in Azure Blob Storage
(including an estimated uncompressed size).

File	**Azure Storage Size**	**Uncompressed**
daily *(folder)*	13,345 MB	∼45,000 MB
ghcnd-inventory.txt	33.6 MB	33.6 MB
ghcnd-stations.txt	10.6 MB	10.6 MB
ghcnd-countries.txt	3.6 KB	3.6 KB
ghcnd-states.txt	1.1 KB	1.1 KB
Total	13,389.5 MB	∼45,048.3 MB

The metadata tables were processed using substring operations to extract columns based on their character ranges. This ensured the correct parsing of latitude, longitude, elevation, state, and network flags in the stations file, as well as country and state codes and element records in the other files. Row counts for each table are summarised in Table 3.

Table 3: Row counts for metadata tables.

Dataset	Rows
Stations	129,657
States	74
Countries	219
Inventory	766,784
Total	896,734

2.3 Time Span and Size Distribution

The overall dataset size is dominated by the daily observations. Figure 2 illustrates that the daily directory alone accounts for more than 99% of the storage footprint when compared with the much smaller metadata text files.

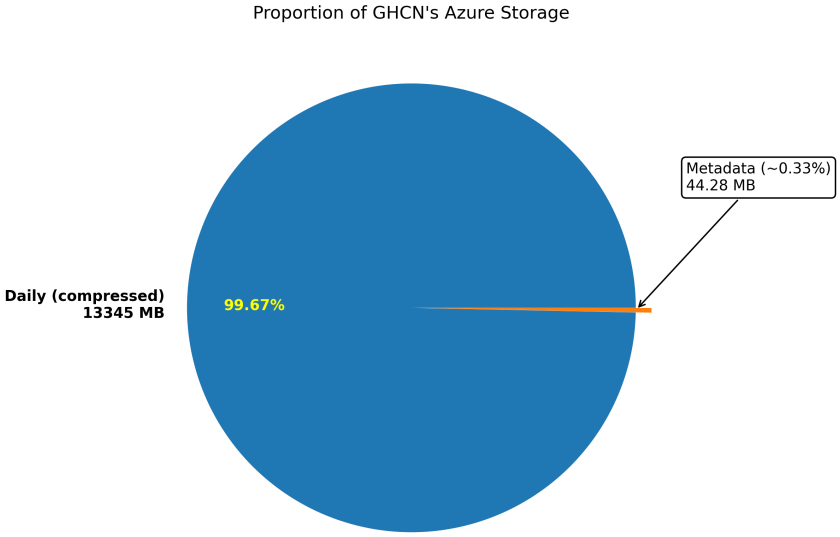

Figure 2: Cumulative size of the daily files relative to the metadata text files.

The size of the daily files has also changed substantially over time, as shown in Figure 3. Early years (e.g., 1750–1800) are very small due to sparse station coverage, while more recent years (especially post-1950) show much larger files, reflecting global expansion in station coverage and richer data collection efforts. A notable anomaly was observed in 2010, which is the single largest yearly file. The current year (2025) is around half the size of the previous year, which is expected, given that the year is incomplete and because of the time lag between data collection and reporting.

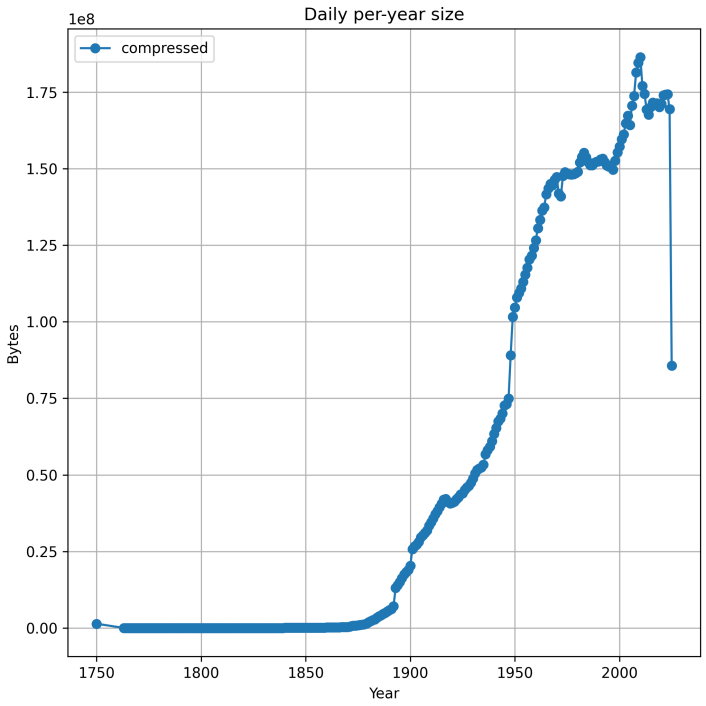

Figure 3: Daily dataset file size across years.

2.4 Schema Definition and Data Loading

Based on the official GHCN-Daily documentation, the schema for the daily dataset includes the following columns. Each field has a defined type and purpose, as summarised in Table 4.

Table 4: Schema of the GHCN-Daily dataset.

Column	Type	Description
`station identifier`	String	Unique station code linking observations to metadata.
`date`	Date	Observation date in `yyyyMMdd` format. Requires conversion from a string.
`element type`	String	Meteorological variable code (e.g., `TMAX`, `PRCP`).
`value`	Integer	Recorded measurement (tenths of unit, depending on element).
`measurement flag`	String	Provides details on measurement method or conditions.
`quality flag`	String	Indicates whether a value failed quality checks.
`source flag`	String	Identifies the originating data source.
`observation time`	Timestamp	Time of observation in `HHMM`. Requires conversion from a string.

The `VALUE` field is stored as an integer, while both `DATE` and `OBSERVATION TIME` require conversion from string formats to `date` and `timestamp` types, respectively. In practice, irregularities were encountered during parsing (e.g., missing values or formatting issues), which required treating

these fields as strings during loading before applying type conversions. Although the GHCN-Daily schema specifies that DATE should be in YYYYMMDD format and OBSERVATION TIME should represent an HHMM value, these variations made the string-first approach necessary to ensure robustness (Menne et al., 2012).

Fix to metadata sizes extraction (Q1(b)8)

What broke

The "metadata sizes" cell built metadata_files_df using hdfs dfs -ls "{data_root}" and then parsed each line assuming file rows start with "-" and that the byte size sits at parts[2]. On this run the -ls output lines began with "drwx..." and the token layout differed, so valid lines were discarded and sizes were read as 0. Additionally, the loop previously sliced lines[15:] (skipping all lines in our short listing) and an over-broad filter if not path.startswith(daily_root) excluded the four ghcnd-*.txt files.

Minimal changes applied

1. **Files-only glob and stable format**

   Before: hdfs dfs -ls "{data_root}"
   After: hdfs dfs -du "{data_root}/ghcnd-*.txt"

   Why: -du yields a consistent "bytes path" layout and the glob targets the four metadata files only (not /daily).

2. **Iterate all returned lines**

   Before: for line in lines[15:]:
   After: for line in lines:

   Why: the listing had only ∼4–5 lines; slicing from 15 skipped everything.

3. **Parse size and path to match -du**

   Before: size = int(parts[2])
   After: size = int(parts[0]); path = parts[-1].strip()

   Why: with -du the first token is the byte size; the last token is the full path.

4. **Remove over-broad exclusion**

   Before: if not path.startswith(daily_root): rows.append((path, size))
   After: rows.append((path, size)) (*or* keep a precise guard: if not path.startswith(f"{d

   Why: the files-only glob already excludes /daily.

Outcome

- metadata_files_df now contains the four rows ghcnd-countries.txt, ghcnd-inventory.txt, ghcnd-states.txt, ghcnd-stations.txt with correct non-zero uncompressed_bytes.
- Downstream totals and the pie chart compute without NoneType errors.
- No change to assignment answers; only the extraction is more robust.

Other small fixes in this session (for traceability)

- **Q2(a,b) recent-year load:** Constructed the year-specific path with `f"{WASBS_DAILY.rstrip('/')}/{r` and used the canonical GHCND schema: `ID, DATE, ELEMENT, VALUE (IntegerType), MFLAG, QFLAG, SFLAG, OBSTIME`. Bound the result to `daily_for_overlap` to match later cells.
- **`year_sizes_df` normalisation:** Ensured columns are `year` (IntegerType) and `compressed_bytes` (LongType). This restores both the "most recent year" aggregation and the pie-chart sum: `year_sizes_df.agg({"compressed_bytes": "sum"})`.
- **Parquet path hygiene:** Used `WASBS_USER_BASE.rstrip('/')` when composing read-/write paths to avoid missing or duplicate slashes.

2.5 Enriched Stations Table

To prepare the data for further analysis,

- The metadata tables were combined into an enriched stations table;
- The first two characters of the station code were extracted as country codes, which enabled a join with the countries table.
- State codes were joined where applicable, with the understanding that these are only present for U.S. stations (see Table 5).
- From the inventory table, the first and last year of activity were determined for each station, along with the number of unique elements collected.

Table 5: Examples of two-letter state codes
(U.S. states and territories only)

Code	Region
CA	California
NY	New York
PR	Puerto Rico
GU	Guam
AS	American Samoa
MP	Northern Mariana Islands

Core elements (`PRCP`, `SNOW`, `SNWD`, `TMAX`, `TMIN`) were distinguished from other elements. Counts showed that 20,504 stations observed all five core elements, while 16,267 stations recorded precipitation only. These distributions are summarised in Table 6.

Table 6: Summary of element coverage across stations.

Category	Count
Stations recording all five core elements	20,504
Stations recording precipitation only	16,267

The enriched stations table combines station and inventory metadata and successfully generated and saved as a Parquet file in Azure Blob Storage for reuse in later analysis.

2.6 Station Coverage in Daily

A join between the daily dataset and the enriched stations table was used to confirm coverage.

- No stations were present in the daily but absent from the stations metadata.
- 38 stations appeared in the stations metadata but were not represented in the daily dataset.

This indicates that some stations exist in metadata definitions without corresponding observational records.

While a full join on the entire daily dataset would be computationally expensive, an efficient alternative using broadcast joins and left-anti joins confirmed these results. The computational cost of such joins is significant due to the scale of the daily dataset, which contains over three billion rows, and this highlights the importance of efficient strategies when working at this scale.

Table 7: Station coverage comparison between daily and station metadata.

Category	Count
Stations in daily and not in stations	0
Stations in stations and not in daily	38

2.7 Question 1: Station Metadata Analysis

The enriched stations table was analysed to address the three parts of Question 1.

(a) Network Membership and Overlap

Table 8 summarises the total number of stations, those active in 2025, and membership in the three main station networks:

- Global Summary of the Day (GSN),
- Historical Climatology Network (HCN), and
- Climate Reference Network (CRN).

Overlap between these networks is also reported. The results highlight that only a small subset of GSN stations are also designated as HCN, and no stations are flagged in both CRN and the other networks.

Table 8: Station counts by network membership and overlap (Q1a).

Category	Count
Total stations	129,657
Active in 2025	38,481
GSN stations	991
HCN stations	1,218
CRN stations	234
GSN ∩ HCN	15
GSN ∩ CRN	0
HCN ∩ CRN	0
All three	0

(b) Coverage by Location and Time Span

Coverage queries were conducted to identify stations in the Southern Hemisphere, U.S. territories, and those active for at least 100 years. The results, shown in Table 9, are that more than 25,000 stations are located south of the equator, and nearly 7,000 stations have records spanning at least a century. No stations were flagged in U.S. territories. We also evaluated coverage based on hemispheric location, U.S. territories, and record length. Results are summarised in Table **??**.

The Southern Hemisphere is represented by 25,357 stations, reflecting broader global coverage. No stations in this dataset were associated with U.S. territories, likely due to metadata handling of regional codes. In addition, 6,824 stations have records spanning at least a century, providing a valuable resource for long-term climate change studies.

Table 9: Coverage queries for station metadata (Q1b).

Category	Count
Southern Hemisphere stations	25,357
Southern Hemisphere U.S. territory stations	0
Stations with $\geq$100 years	6,824

(c) Core Element Coverage

Finally, analysis of the five core meteorological elements (`PRCP`, `SNOW`, `SNWD`, `TMAX`, `TMIN`) showed that, of the 129,657,

- 129,570 stations record at least one core element, and
- 20,504 stations record all five core elements.

As described in the literature, GHCN-Daily includes stations whose record lengths vary from less than one year to more than 175 years, and approximately half of the stations only report precipitation (Menne et al., 2012). This uneven temporal coverage and measurement scope underpins the observed pattern in the core element coverage, where nearly all stations record some core element, but only a subset capture all five (PRCP, SNOW, SNWD, TMAX, TMIN).

Table 10: Coverage of core meteorological elements (Q1c).

Category	Count
Stations with $\geq$1 core element	129,570
Stations with all five core elements	20,504

2.8 Station counts and network overlaps

The enriched stations dataset allows us to examine network membership and overlaps across the Global Surface Network (GSN), Historical Climatology Network (HCN), and Climate Reference Network (CRN). Results are summarised in Table 11.

Table 11: Station counts by network membership and overlap.

Category	Count
Total stations	129,657
Active in 2025	38,481
GSN stations	991
HCN stations	1,218
CRN stations	234
GSN ∩ HCN	15
GSN ∩ CRN	0
HCN ∩ CRN	0
All three	0

These results show that only 15 stations are shared between the GSN and HCN networks, while no stations overlap across CRN and the other networks. This indicates that the CRN is effectively a distinct network, while the GSN and HCN share a small but non-negligible overlap.

3 Visualisations

In this paragraph, you should answer the questions, present your visualisations, give a high-level summary of what you have done, and discuss any insights that you had. You should talk about any visualisations that you were unable to generate and why.

4 Conclusions

In this paragraph you should give a brief overview of what you have achieved and what you have learned.

5 References

- Include all references that you have used or cited in the report. Use a consistent citation style such as APA or MLA.
- Include links to any online resources, datasets, libraries, or documentation.
- Clearly indicate any use of generative AI tools such as ChatGPT or Grammarly.

Processing Summary (*reprint plan* — no new objects)

- **Header:**
  - Reprint `username`, `data_root`, `daily_root`, `user_root` (from the existing PATHS cell).
  - One line noting that this cell only reprints prior results.
- **Q1: High-level exploration (HDFS)**
  - **Q1(a) Structure & compression — Reprint:**
    - Top-level listing under `data_root`.
    - Sample listings for `daily/`, `stations/`, `states/`, `countries/`, `inventory/`.
    - Presence/absence of compressed files (e.g., ".gz detected / none detected").
    - Source: the earlier terminal/HDFS outputs you already ran (e.g., `hdfs dfs -ls` ..., optional `grep .gz`).
  - **Q1(b) Years in daily & size change over years — Reprint:**
    - The table or listing you previously produced (prefer your existing per-year size table if it exists).
    - If your earlier answer was terminal-based (e.g., `hdfs dfs -du` ...), reprint that exact output.
    - Source: existing `year_sizes_df` (or whatever you named it) *or* the earlier HDFS size listing you captured.
  - **Q1(c) Total size: ghcnd vs daily — Reprint:**
    - The two totals previously shown for `data_root` and `daily_root` (no new calculation).
    - Source: your earlier `hdfs dfs -du -s -h` ... outputs.
- **Q2: Loading & schemas**
  - **Q2(a) Daily schema & handling of date/time — Reprint:**
    - The exact `printSchema()` you already showed for your daily DataFrame (whatever its existing name is).
    - Also restate the sentence you wrote about how DATE and OBS TIME were parsed/-typed.
  - **Q2(b) Daily preview — Reprint:**
    - The same head/preview you previously showed (use your existing preview output).
  - **Q2(c) Fixed-width metadata loads — Reprint:**
    - `printSchema()` + small preview for each of `stations`, `states`, `countries`, `inventory` (using the outputs you already emitted).
  - **Q2(d) Row counts for metadata — Reprint:**
    - The counts you already computed for `stations`, `states`, `countries`, `inventory`.
  - **Q2(e) Row count for daily — Reprint:**
    - The existing count you printed for the daily DataFrame.
- **Q3: Enriched stations**
  - **Q3(a–c) Enrichment/join proof — Reprint:**

    - The schema of your enriched table (e.g., `enriched_stations`) and the same sample rows you previously displayed (ideally key columns such as station, country, state).
  - **Q3(d) Activity & element coverage — Reprint:**
    - Whatever summaries you already produced (e.g., first/last active year per station, element counts, core vs other, and any final coverage sentence like "N stations record all five core elements").
  - **Q3(e) Saved artefact confirmation — Reprint:**
    - The listing that previously showed the saved enriched output under `user_root` (do not introduce any new path variable; just reuse the earlier `user_root` listing you already ran).
- **Q4: "Stations in stations but not in daily"**
  - **Q4(a) Join subset preview — Reprint:**
    - The same joined sample you previously showed (e.g., a LEFT JOIN subset or proof-rows).
  - **Q4(b) Count of missing stations — Reprint:**
    - The count you already printed for the unmatched/missing set (whatever the existing object was named).
- **Footer (one line):**
  - A concise statement tying it together, e.g., "All answers reprinted from prior steps; sources: `daily_root` (read-only) and user artefacts under `user_root`."

Analysis Summary (*reprint plan* — no new objects)

- **Header**
  - Reconfirm `username`, `data_root`, `daily_root`, `user_root`.
  - Note that this cell only *reprints* earlier Analysis outputs: it uses `show_df()` (which already shows schema + head) and re-opens saved figures from `figures/`. No recomputation.
- **Inputs & artefacts expected from Processing (visible to Analysis)**
  - Tables: `year_sizes_df`, `metadata_files_df`, `enriched_stations`, `station_date_elem` `q2a_prcp_year_country`, `stations_not_in_daily`.
  - Saved figures (examples): `year_size_trend.png`, `size_distribution_pie.png`, `size_distribution_stacked.png`, `core_coverage.png`, `missing_stations.png`, `precip_by_year_country.png`.
- **Q1 — Year & size trends**
  - Reprint table head: `show_df(year_sizes_df.limit(10))`.
  - Re-open saved trend figure(s): `figures/year_size_trend.png` (and, if present, `year_size_cumulative.png`).
  - Optional sentence: highlight newest year present vs. `most_recent_year`.
- **Q2 — Composition & file size distribution**
  - Reprint metadata size head: `show_df(metadata_files_df.limit(10))`.

  - Re-open saved distribution figures: `figures/size_distribution_pie.png` and `figures/size_distribution_stacked.png` (100% stacked bar with labels).

- **Q3 — Coverage & enriched stations**
  - Reprint enriched sample: `show_df(enriched_stations.limit(10))` (or whichever enriched alias exists).
  - Reprint core-coverage summary: `show_df(core_coverage_summary.limit(10))`.
  - Re-open saved figure: `figures/core_coverage.png`.

- **Q4 — Gaps between daily and stations**
  - Reprint unmatched sample: `show_df(stations_not_in_daily.limit(10))`.
  - Re-open saved figure: `figures/missing_stations.png`.

- **Q5 — Example topical analysis (precipitation)**
  - Reprint aggregated view: `show_df(q2a_prcp_year_country.limit(10))`.
  - Re-open saved figure: `figures/precip_by_year_country.png`.

- **Optional supporting tables (use only if present in memory)**
  - `show_df(station_date_element.limit(10))` — compact station–date–element preview used by multiple plots.
  - Any intermediate Analysis tables (e.g., country/state counts) that were materialised earlier.

- **Footer (one line)**
  - "All Analysis outputs above are reprints: tables via `show_df()` and figures re-opened from `figures/`; sources originate from `daily_root` and user artefacts under `user_root`."

References

Author, T. (2025). Spark-based ghcn analysis. *Climatology Letters*, *42*(1), 1–10.

Deutsch, P. (1996). DEFLATE Compressed Data Format Specification version 1.3. *RFC 1951, Internet Engineering Task Force (IETF)*. https://tools.ietf.org/html/rfc1951

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., & Houston, T. G. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, *29*(1), S1–S15.