# DATA420-25S2(C)

## Assignment 1

# GHCN Data Analysis using Spark

David Ewing (82171165)

2025-03-28 07:51

Hello ├── test

**testentry**

# 1 Analysis of the GHCN-Daily Dataset

## 1.1 Q1(a), Data Structure

The dataset is stored in an Azure Blob Storage container named `ghcnd`, structured using virtual directories. As shown in Figure 2, each file path follows the format:

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

Within the `ghcnd` container: :

- Metadata files in the root are `stations.txt`, `states.txt`, `countries.txt`, and `inventory.txt`,
- The main subdirectory is `daily` which holds 264 compressed `.csv.gz` file,
- Each compressed files holds the daily records for one year,
- The records span 1750 to 2025 with 12 years data missing.

The layout is illustrated in Figure 1, showing a tree-like directory structure with yearly files nested under `daily`.

```
ghcnd
├── daily
│  ├── 1750.csv.gz
│  ├── 1751.csv.gz
│  ├── …
│  ├── 2024.csv.gz
│  └── 2025.csv.gz
├── countries.txt
├── inventory.txt
├── states.txt
└── stations.txt
```

Figure 1: Azure Blob Storage GHCND directory structure

## Directory Structure of `ghcnd/`

The `daily/` folder contains yearly data files in compressed CSV format:

The `daily/` folder contains yearly data files in compressed CSV format:

| Component | Type | Description |
|---|---|---|
| `wasbs://` | Protocol | Azure Blob Storage over SSL (used with Hadoop-compatible connectors) |
| `campus-data` | Container | Like a top-level directory within the blob storage account |
| `madsstorage002` | Storage account | The Azure storage account name |
| `blob.core.windows.net` | Service domain | Azure's blob service endpoint domain |
| `/ghcnd/daily/2024.csv.gz` | Path | Virtual directory and file structure within the container |
| `ghcnd` | Container name | `<container>` |
| `daily` | Optional folder(s) within container | `<directory>` |
| `2024.csv.gz` | Blob name (file) | `<filename>` |

Figure 2: Structure of a ‘wasbs://‘ Azure Blob Storage URI used for GHCN-Daily datasets

## 2 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.

## 3 Processing

In this paragraph you should describe the structure and content of the data and provide references for discussion in the sections below. You should describe the steps that you took to load, join, and check the data and discuss anything that you discovered along the way, but you should not include outputs other than answers to the questions that you have been asked.

## 4 Analysis

In this paragraph you should answer the questions that you have been asked, give a high level summary of what you have done, and discuss any insights that you had. You should not list answers you have not explained. You should talk about any tasks that you were unable to complete and explain why.

## 5 Visualizations

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

## 6 Conclusions

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

## 7 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.