CS236 PROJECT REPORT

Team members:

Name: Abhijit Kulkarni

Student ID: 862393541

Email: akulk022@ucr.edu

Name: Suryaa Charan Shivakumar

Student ID: 862395094

Email: sshiv012@ucr.edu

Table of Contents

Execution time – 2 min 30 sec (approx..)

Including Buoys and Stations without state - 3 min 1 sec (approx.)

Join Strategy – We used a **map-side join** with the use of **distributed cache** and **in-memory hashmap** (with StationID as the key and state as the value) to perform the join (explained later). We did not use a reduce-side join because the dataset after filtering stations in USA was still large (almost 39,000 stations in a file size of 400 KB).

High level description of the flow:

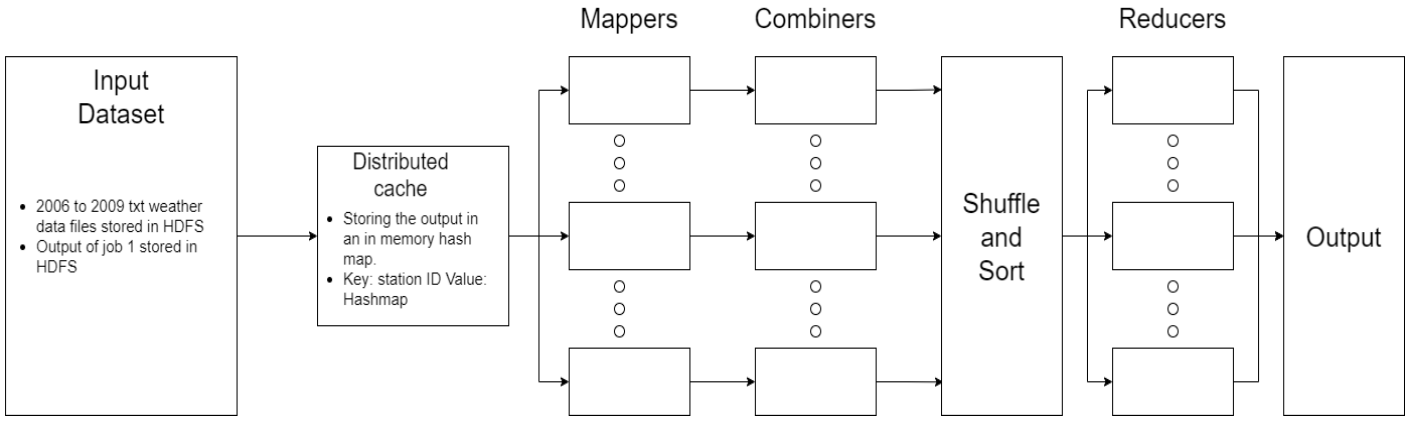
Job 1: This is a mapper only job. It filters stations based on country US, if include_approx flag is set it considers

1) Buoys - Average for all buoys in ocean and assign to buoys as the state.

2) For Locations without a state in US, we calculate closest US state with a threshold, and we assign the state.

Details of Jobs 2 and 3 given below.

Job 2:



Mappers:

a) Process: Performs a **Map-side join** with the join attribute as the station id(usafId) using the in memory HashMap.
b) Output: State and month along with temperature and precipitation values parsed from the weather data txt files. For e.g. (AK_01 8.461821978281211 6.835507314279786)

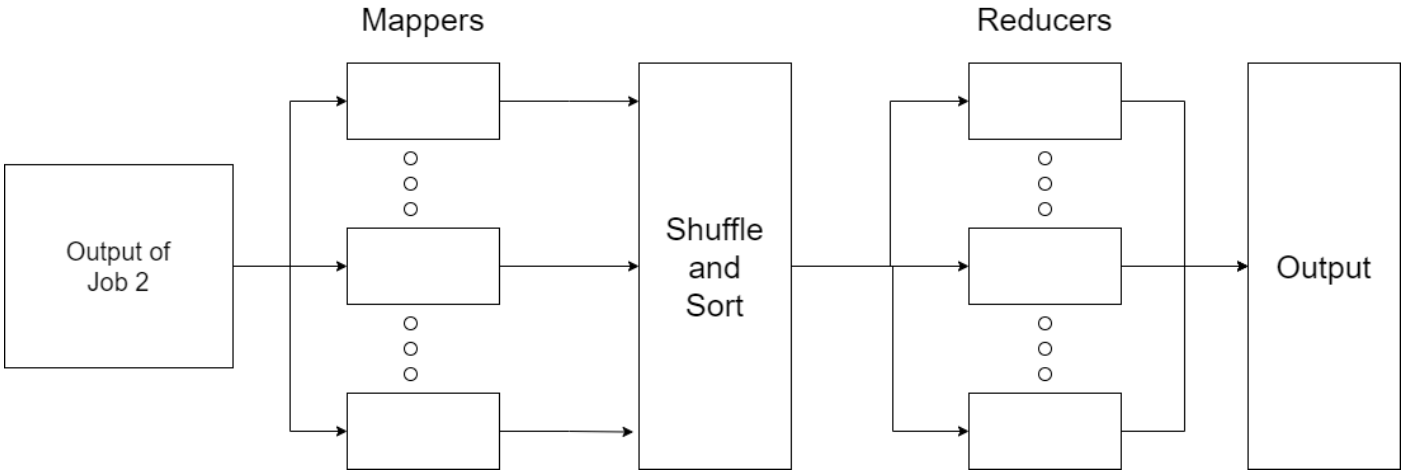
Combiners:

a) Process: Calculates the local averages and counts for temperature and precipitation readings to reduce the computation load on the reducer.
b) Output: Output goes to the reducer for this job.

Reducers:

a) Process: Calculates the weighted average by calculating the sums based on the averages and counts sent by the combiner, calculating a total sum, and then dividing it by total count.
b) Output: Reduced output with the same format as the mapper output but now with averages for the months. For e.g. (AK_01 8.461821978281211 6.835507314279786)

Job 3:



Mappers:

1. Process: Sends the state month and averages to the reducer after reformatting the key and value.
2. Output: Key: AK Value: 01_8.461821978281211_6.835507314279786

Reducers:

1. Process: Calculates the months with the maximum and minimum averages for temperature precipitation and the difference between the maximum and minimum temperatures. We sorted these records based on difference value using a sorted map.
2. Output: State Maximum Precipitation Minimum Precipitation Difference
   VI Jul 83.28 11.95 Feb 77.55 0.94 5.73

Commands and usage instructions -

- To build the maven project (attached with the project) – go to the root directory where the pom.xml is location and run the below command to build the project and find the jar in target folder generated.
  `mvn clean install`
- Use the below commands to create a directory to store the input datasets
  `hadoop fs –mkdir input_dir1`

  ```
  hadoop fs –put <PathtoWeatherStationLocations.csv> ./input_dir1
  hadoop fs –put <Pathto2006.txt> ./input_dir1
  hadoop fs –put <Pathto2007.txt> ./input_dir1
  hadoop fs –put <Pathto2008.txt> ./input_dir1
  hadoop fs –put <Pathto2009.txt> ./input_dir1
  ```
- Alternatively, we have also attached the jar to directly execute the command. Please run the below command, modify as per requirement
- Set the -include_approx flag to include stations without states and ocean buoys, if not mentioned these stations will not be involved for the computation

  ```
  hadoop jar <pathToJar> <HDFSPathtoWeatherStationLocations.csv>
  <HDFSPathToDirectoryWith2006–2009.txtFiles> <outputDirForJob1>
  <outputDirForJob2> <outputDirForJob3>| optional –include_approx
  ```

```
Eg. hadoop jar /Users/johnwesley/Desktop/GIT/MapReduce/WeatherReportProcessing/target/weather-1.0.jar /input_dir1/WeatherStationLocations.csv /input_dir1 /output_dir1 /output_dir2 /output_dir3
```

```
Eg 2. hadoop jar /Users/johnwesley/Desktop/GIT/MapReduce/WeatherReportProcessing/target/weather-1.0.jar /input_dir1/WeatherStationLocations.csv /input_dir1 /output_dir1 /output_dir2 /output_dir3 -include_approx
```

Output:

State		Maximum Precipitation			Minimum Precipitation		Difference
VI	Jul	83.28	11.95	Feb	77.55	0.94	5.73
PR	Aug	82.38	45.86	Feb	76.16	18.13	6.22
HI	Aug	78.12	4.53	Feb	70.22	5.5	7.9
FL	Aug	82.54	7.96	Jan	61.2	4.85	21.34
CA	Jul	73.38	1.85	Dec	47.86	4.69	25.52
LA	Aug	82.51	12.4	Jan	52.01	12.43	30.5
OR	Jul	67.3	3.54	Dec	36.68	7.05	30.62
WA	Jul	66.51	3.68	Dec	35.55	6.64	30.96
GA	Aug	80.82	11.33	Jan	48.31	9.78	32.51
TX	Aug	82.8	12.19	Jan	49.33	10.12	33.47
MS	Aug	81.66	6.17	Jan	48.11	7.03	33.55
AL	Aug	81.15	12.61	Jan	47.47	8.84	33.68
SC	Aug	80.85	5.97	Jan	46.69	9.74	34.16
NC	Aug	78.93	19.34	Jan	44.3	15.85	34.63
VA	Aug	77.08	15.81	Feb	39.29	14.2	37.79
NM	Jul	74.95	10.82	Jan	35.27	4.63	39.68
TN	Aug	80.13	3.39	Jan	40.18	3.36	39.95
WV	Aug	72.73	22.51	Feb	32.68	23.38	40.05
DE	Jul	76.53	2.91	Feb	36.39	1.31	40.14
AR	Aug	80.89	7.27	Jan	40.68	5.95	40.21
MD	Jul	76.59	6.46	Feb	36.25	7.05	40.34
AZ	Jul	84.71	6.91	Dec	44.28	6.62	40.43
RI	Jul	72.83	2.87	Feb	31.99	4.0	40.84
KY	Aug	78.08	7.34	Feb	36.46	10.71	41.62
NJ	Jul	75.34	4.0	Feb	33.39	3.89	41.95
MA	Jul	71.46	3.99	Feb	29.32	4.12	42.14
CT	Jul	72.86	1.2	Feb	30.63	2.52	42.23
OK	Aug	82.27	17.62	Jan	39.98	10.61	42.29
PA	Jul	72.57	5.4	Feb	29.19	9.13	43.38
NY	Jul	71.36	3.62	Feb	27.0	4.9	44.36
OH	Aug	73.04	5.35	Feb	28.04	8.46	45.0
MO	Aug	78.26	2.16	Jan	33.15	1.92	45.11
ME	Jul	66.2	1.84	Jan	21.02	0.52	45.18
DC	Jul	77.44	0.0	Jan	32.23	0.0	45.21
NH	Jul	67.89	5.25	Feb	22.37	5.02	45.52
IN	Aug	73.89	7.54	Feb	28.35	12.92	45.54
CO	Jul	68.57	15.77	Jan	23.01	10.71	45.56
AK	Jul	54.23	6.03	Jan	8.46	6.84	45.77
IL	Aug	74.67	24.09	Feb	28.64	27.93	46.03
MI	Jul	68.58	20.18	Feb	22.02	32.35	46.56
KS	Jul	79.21	2.68	Dec	32.57	1.95	46.64
VT	Jul	68.28	12.0	Jan	21.1	10.0	47.18
ID	Jul	72.19	1.7	Jan	24.22	4.74	47.97
NV	Jul	83.31	1.96	Dec	34.11	1.72	49.2
WY	Jul	71.5	2.38	Dec	20.93	6.19	50.57
NE	Jul	76.36	18.04	Dec	25.62	19.76	50.74
MT	Jul	73.34	1.97	Dec	21.95	1.4	51.39
IA	Jul	74.45	24.61	Jan	23.0	28.09	51.45
WI	Jul	70.05	22.24	Feb	18.49	27.04	51.56
UT	Jul	79.18	1.8	Jan	25.02	2.76	54.16
SD	Jul	75.43	3.66	Feb	20.32	4.5	55.11
MN	Jul	71.12	29.16	Feb	13.83	38.52	57.29
ND	Jul	72.37	2.09	Feb	11.91	1.95	60.46

Final output

State		Maximum Precipitation			Minimum Precipitation		Difference
VI	Jul	83.28	11.95	Feb	77.55	0.94	5.73
PR	Aug	82.38	45.86	Feb	76.16	18.13	6.22
HI	Aug	78.24	3.73	Feb	70.81	4.82	7.43
FL	Aug	82.64	6.71	Feb	61.83	4.99	20.81
BUOY	Aug	69.12	0.61	Feb	46.27	1.76	22.85
CA	Jul	71.93	1.68	Dec	48.64	4.15	23.29
OR	Jul	65.52	3.03	Dec	38.2	6.06	27.32
WA	Jul	65.19	3.2	Dec	36.52	5.81	28.67
LA	Aug	82.81	9.99	Jan	53.78	9.96	29.03
GA	Aug	80.88	10.81	Jan	48.62	9.41	32.26
AL	Aug	81.4	11.61	Jan	48.76	8.41	32.64
MS	Aug	81.77	5.93	Jan	48.57	6.59	33.2
TX	Aug	82.81	12.07	Jan	49.48	9.99	33.33
SC	Aug	80.94	5.34	Jan	47.02	8.81	33.92
NC	Aug	78.94	19.25	Jan	44.57	15.64	34.37
RI	Jul	71.11	1.98	Feb	34.64	2.63	36.47
VA	Aug	77.08	15.81	Feb	39.29	14.2	37.79
DE	Jul	75.85	2.0	Feb	37.25	0.92	38.6
NM	Jul	74.95	10.82	Jan	35.27	4.63	39.68
TN	Aug	80.13	3.39	Jan	40.18	3.36	39.95
WV	Aug	72.73	22.51	Feb	32.68	23.38	40.05
AR	Aug	80.89	7.27	Jan	40.68	5.95	40.21
MD	Jul	76.62	6.2	Feb	36.26	6.75	40.36
AZ	Jul	84.71	6.91	Dec	44.28	6.62	40.43
NJ	Jul	75.1	3.59	Feb	33.65	3.46	41.45
KY	Aug	78.08	7.34	Feb	36.46	10.71	41.62
MA	Jul	71.29	3.84	Feb	29.41	3.96	41.88
CT	Jul	72.98	0.95	Feb	31.08	2.0	41.9
OK	Aug	82.27	17.62	Jan	39.98	10.61	42.29
PA	Jul	72.46	5.12	Feb	29.01	8.66	43.45
ME	Jul	65.69	1.7	Jan	21.83	0.47	43.86
NH	Jul	67.47	4.69	Feb	23.2	4.47	44.27
NY	Jul	71.25	3.38	Feb	26.74	4.62	44.51
AK	Jul	54.22	5.86	Jan	9.33	6.6	44.89
OH	Aug	72.98	5.14	Feb	27.96	8.3	45.02
MO	Aug	78.26	2.16	Jan	33.15	1.92	45.11
DC	Jul	77.44	0.0	Jan	32.23	0.0	45.21
IN	Aug	73.89	7.54	Feb	28.35	12.92	45.54
CO	Jul	68.57	15.77	Jan	23.01	10.71	45.56
IL	Aug	74.65	23.68	Feb	28.6	27.47	46.05
MI	Jul	67.93	18.01	Feb	21.86	30.29	46.07
KS	Jul	79.21	2.68	Dec	32.57	1.95	46.64
VT	Jul	68.28	12.0	Jan	21.1	10.0	47.18
ID	Jul	72.19	1.7	Jan	24.22	4.74	47.97
NV	Jul	83.31	1.96	Dec	34.11	1.72	49.2
WY	Jul	71.5	2.38	Dec	20.93	6.19	50.57
NE	Jul	76.36	18.04	Dec	25.62	19.76	50.74
MT	Jul	73.34	1.97	Dec	21.95	1.4	51.39
WI	Jul	69.97	21.6	Feb	18.55	26.59	51.42
IA	Jul	74.45	24.61	Jan	23.0	28.09	51.45
UT	Jul	79.18	1.8	Jan	25.02	2.76	54.16
SD	Jul	75.43	3.66	Feb	20.32	4.5	55.11
MN	Jul	70.87	28.44	Feb	13.89	37.8	56.98
ND	Jul	72.37	2.09	Feb	11.91	1.95	60.46

Final output with inclusion of buoys and stations without state assigned to nearest state (-include_approx flag set)

Our Efforts towards gaining extra credit:

1. A good use of combiners
    - **We used a combiner to calculate local averages before passing the data to the reducer to reduce the computational load for it.**
    - **Since the calculation of averages is not commutative by nature, we used a weighted average method as mentioned earlier.**
2. A clever way to achieve faster execution time
    - **We used a distributed cache and an in memory HashMap to perform the map-side join. Other approaches that we attempted involved additional jobs (increased execution time by roughly 40 seconds because of job initialization) or performing a reduce-side join (increased execution time by roughly 1 minute because of the size of the dataset).**
    - **Use of TreeMap to sort the states with the difference in extreme temperatures as the key in the final reducer. We made use of the heuristic that there only about 50 states/regions and this might not require another job to sort them. Hence, in the cleanup function of the final reducer we wrote the output from the sorted treeMap.**
3. Enriching the data, e.g., including the average precipitation for the two months
    - **For the months for which maximum and minimum temperatures were calculated. We used a similar approach used for temperatures, for the precipitation values to enrich the data.**
4. Bonus: Include the stations with "CTRY" as "US" that don't have a state tag, finding a way to estimate the state using a spatial distance with the known stations. There are some stations that are Ocean Buoys so you may want to have a maximum distance to be required in order to be included in a state, or you could create a separate "state" representing the "pacific" and "atlantic" ocean (Checked by using coordinates).
    - **If include_approx flag is set we include**
        1) **Ocean buoys that belong to the US by classifying their state to BUOY. We include them in the computation for our primary solution.**
        2) **For all the stations which are in the US but are not assigned to any state, we calculate the nearest state by calculating the distance from the centroids of the state. We filter out noise by using the radius of the biggest state in the US (Alaska) as the threshold. Then we assign the station to the state with the least distance as per our earlier calculations as long as it is below the mentioned threshold.**

Contributions:

1) Contributions by Suryaa Charan:

- Creating a maven project and configuring it.
- Coming up with the high-level approach for the desired solution through brainstorming sessions and abstraction of jobs, mappers, combiners, and reducers.
- Implementation of distributed cache and map-side join.
- Inclusion of buoys and stations with unassigned states within US.
- Developing parsers and filtering out data.
- Code optimizations to achieve faster execution times.

2) Contributions by Abhijit:

- Refactoring the structure and logic of the code for mappers, combiners, reducers.
- Coming up with the high-level approach for the desired solution through brainstorming sessions.
- Implementation of the logic for calculating the maximum and minimum average temperatures and precipitations for every month.
- Code optimizations for better memory management.
- Documenting all the efforts and keeping track of the progress of the project.

commit bd53a480dd3ef1674edda80f3ec7a70124b66c58

Author: suryaacharan <johnwesley@MacBook-Pro-2.local>

Date: Tue Nov 29 17:33:21 2022 -0800

Code Refactoring, Added optional flag; Added functionality to approximate states if not mentioned; Added functionality to include ocean buoys classifed as a state

commit 0bf5b67cbacac07ddb2cc58538da290838ec2b08

Merge: 4ebc59e eeb49c2

Author: suryaacharan <97846602+suryaacharan@users.noreply.github.com>

Date: Sat Nov 26 18:29:14 2022 -0800

Merge pull request #1 from suryaacharan/feature-restructuringAddingLogicForMonthsPrecipitation

Feature restructuring adding logic for months precipitation

commit eeb49c2394c76cfd88f9d002a3cfce311236895a

Author: akulk022 <116849480+akulk022@users.noreply.github.com>

Date: Sat Nov 26 17:18:39 2022 -0800

Delete Test.java

commit 2c496472335c269395a86e8d43e8bc334e4b20c8

Author: akulk022 <116849480+akulk022@users.noreply.github.com>

Date: Sat Nov 26 17:17:30 2022 -0800

Delete bin directory

commit 1fc26fca1a230777cf0b84634b57ec4d6d1f356b

Merge: 489fa21 4ebc59e

Author: akulk022 <116849480+akulk022@users.noreply.github.com>

Date: Thu Nov 24 19:33:40 2022 -0800

Merge branch 'dev' into feature-restructuringAddingLogicForMonthsPrecipitation

commit 489fa21d1fe09517e5fb64883cc5382b1f6f302d

Author: Abhijit Kulkarni <akulk022@ucr.edu>

Date: Thu Nov 24 19:27:08 2022 -0800

mappers and reducers refactoring

commit 60058ce4d56584df8dec1c40ba4842c694473cb4

Author: Abhijit Kulkarni <akulk022@ucr.edu>

Date: Thu Nov 24 19:26:31 2022 -0800

The Git history for our repository of this project.