Chapter 1

Week-13 Cluster Assignment

Notes

For test 4-15 You will need to recompile your **MaxTemperature** driver class and JAR to modify the following: `job.setNumReduceTasks(X)` and set `job.setJobName("Max temperature");` to include your initials: `job.setJobName("MT test 1 jrh");`

For test 4-6 You will need modify your MaxTemperature code to account for bad records use the sample code on P.174 of textbook *section 6-12* to include a counter that will output the number of bad records at the end of the job – place this new code and jar file in Week-13 -> badrecord folder (include just 1 sample)

Test 17

For MaxTemperature.*java recompile* `mt.jar` *changing* `job.setJobName` *name to add your initials For Test 1-3* A-E 1999 and small/1999 * F-N 1991 and small/1991 * O-Z 1994 and small/1994

For each bullet point in a Test you will run it using these datasets: /user/ncdc/199X/199X.txt* and /user/ncdc/small/199X/* datasets * 199X.txt * 199X.txt.bz2 * 199X.txt.gz * 199X small files (these are the induvidual 4K files not combined to a single file)

Test 1

- Without combiner, without intermediate compression, and 1 reducers
- Without combiner, without intermediate compression, and 2 reducers
- Without combiner, without intermediate compression, and 4 reducers
- Without combiner, without intermediate compression, and 8 reducers

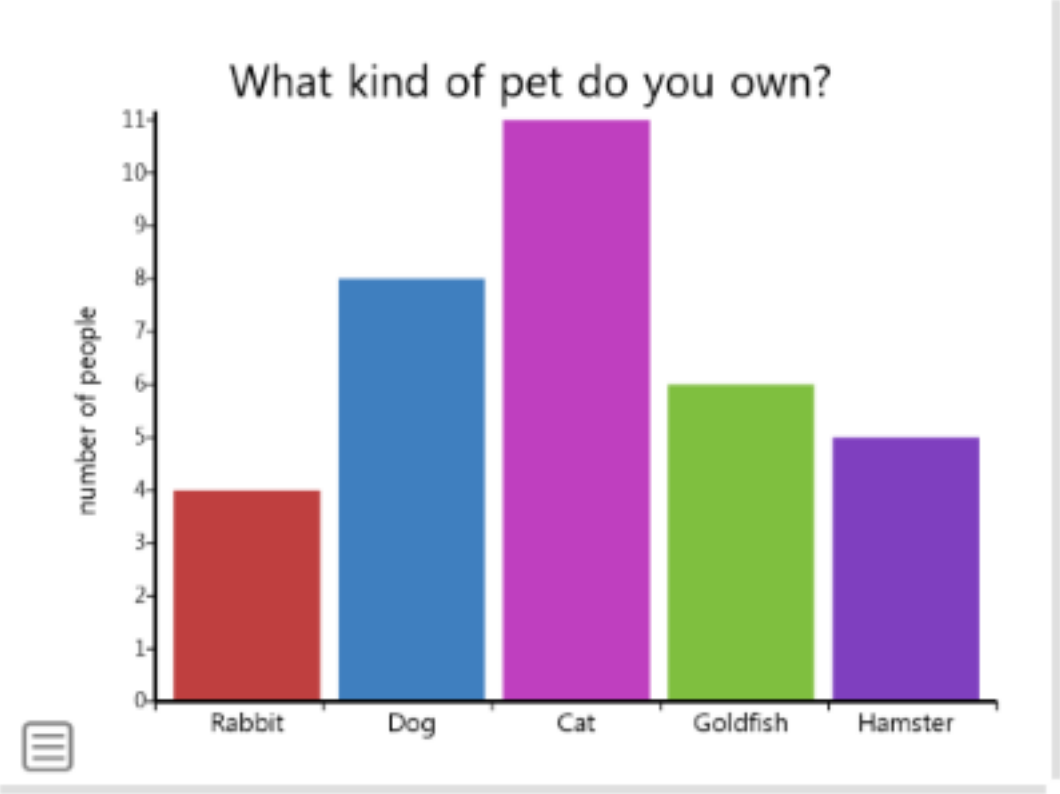

Figure 1.1: *Item 1 Results*

Answer

Your explanation of the reults of the above graph goes here. This explanation tells in detail why there are increase or decrease in execution times – and invovles information from the textbook (not the web or Google) including with page or ebook section citations.

Test 2

- With combiner, with intermediate compression, and 1 reducers
- With combiner, with intermediate compression, and 2 reducers
- With combiner, with intermediate compression, and 4 reducers
- With combiner, with intermediate compression, and 8 reducers

Chart goes here

Answer

Your explanation of the reults of the above graph goes here. This explanation tells in detail why there are increase or decrease in execution times – and invovles information from the textbook (not the web or Google) including with page or ebook section citations.

Test 3

- With combiner, without intermediate compression, and 1 reducers
- With combiner, without intermediate compression, and 2 reducers
- With combiner, without intermediate compression, and 4 reducers
- With combiner, without intermediate compression, and 8 reducers

Chart goes here

Answer

Your explanation of the reults of the above graph goes here. This explanation tells in detail why there are increase or decrease in execution times – and invovles information from the textbook (not the web or Google) including with page or ebook section citations.

Test 4

- Repeat test 1 using the datasets located in: `/user/ncdc/80/` (Lastname A-K) and `/user/ncdc/90/` (Lastname L-Z)
- For test 4-6 You will need modify your MaxTemperature code to account for bad records use the sample code on P.174 of textbook *section 6-12* to include a counter that will output the number of bad records at the end of the job – place this new code and jar file in Week-13 -> badrecord folder

---

Test 5

- Repeat test 2 using the datasets located in: `/user/ncdc/80/` (Lastname A-K) and `/user/ncdc/90/` (Lastname L-Z)
- For test 4-6 You will need modify your MaxTemperature code to account for bad records use the sample code on P.174 of textbook *section 6-12* to include a counter that will output the number of bad records at the end of the job – place this new code and jar file in Week-13 -> badrecord folder

---

Test 6

- Repeat test 3 using the datasets located in: `/user/ncdc/80/` (Lastname A-K) and `/user/ncdc/90/` (Lastname L-Z)
- For test 4-6 You will need modify your MaxTemperature code to account for bad records use the sample code on P.174 of textbook *section 6-12* to include a counter that will output the number of bad records at the end of the job – place this new code and jar file in Week-13 -> badrecord folder

---

Test 7

- Repeat test 1 using the datasets located in: `/user/ncdc/80-90/` (ALL students)

Test 8

- Repeat test 2 using the datasets located in: `/user/ncdc/80-90/` (ALL students)

Test 9

- Repeat test 3 using the datasets located in: `/user/ncdc/80-90/` (ALL students)

Test 10

- Repeat test 1 using the dataset and the `/user/ncdc/256/80-90/` datasets (ALL students)
- This increases the default Block Size to 256 mb

Test 11

- Repeat test 2 using the dataset and the `/user/ncdc/256/80-90/` datasets (ALL students)
- This increases the default Block Size to 256 mb

Test 12

- Repeat test 3 using the dataset and the `/user/ncdc/256/80-90/` datasets (ALL students)
- This increases the default Block Size to 256 mb

Test 13

- Repeat test 1 using the dataset and the `/user/ncdc/512/80-90/` datasets (ALL students)
- This increases the default Block Size to 512 mb

Test 14

- Repeat test 2 using the dataset and the `/user/ncdc/512/80-90/` datasets (ALL students)
- This increases the default Block Size to 512 mb

Test 15

- Repeat test 3 using the dataset and the `/user/ncdc/512/80-90/` datasets (ALL students)
- This increases the default Block Size to 512 mb

Test 16

- Using the datasets contained in `/user/logs/large-logs` write and compile a MapReduce Program to do the following: For each month in the year find the page (html or php) name that has the highest number of HTTP 200 responses that is not the index.* page as well as the number (count) of those results graphed together. Use the following years and files:
  - Lastname A-E 2012 + 2013
  - Lastname F-N 2014 + 2015
  - Lastname O-Z 2016 + 2012

  - `web-server-logs.txt`
  - `web-server-logs.bz2`
  - `web-server-logs.gz`
- Graph the retrieved data, based on results of previous 15 tests, you decide the single configuration that is optimum for achieving the results and explain why you chose that configuration for this test. Include your code inside your Github account in a folder

named test-16. Make sure all coded needed to compile and run the job is present. Add a ReadMe.md with any instructions or assumptions in that folder.

Conclusion (test 17)

Write your general conclusion and recommendation for the optimal performance characteristics relating to **intermediate compression**, **compression**, **combiners**, **block size**, and **number of reducers** based on this work load and dataset results (reference your test results above).