Lab 5

Big Data 2017
February 27, 2017

Homework Review

- Homework 1 was due today
- Review tasks, talk about issues

Homework Review: Task 1

- Challenges: reading multiple input files, determining which file a line came from
- One way: using environment variables set by Hadoop Streaming
- Alternative: use the fact that the two CSV files have different numbers of columns
- Can select the needed columns (from parking-violations) in the mapper, or just pass the whole lines through and do this in the reducer.

Homework Review: Task 2, 3, 4

- Very similar to wordcount - should have been easy
- Task 3 required computing an average

Homework Review: Task 5,6

Task 5:

- Used a key with 2 parts in it (license and reg state)
- Wanted the maximum so needed to use only one reducer
    - In the reducer, needed to keep track of the maximum element seen so far, only print this at the end

Task 6:

- Similar to task 5, but wanted top 20
- Reducer needs to maintain some data structure of max 20 seen so far (or just store everything and sort at the end)

Homework Review: Task 5,6

In practice, using one reducer won’t scale well.

A better idea is to not restrict the number of reducers, let each reducer output its max (or top 20), and use a quick (non-mapreduce) script to take the max among reducers at the end.

Homework Review: Task 7

- Also similar to wordcount, but need to parse whether day is weekday or weekend
- For this task, we had you hard-code the dates of the month that were weekends and gave you the total number of weekdays/weekend days to compute the averages
- How would you do this if you weren’t given these values, or didn’t know them for your dataset?

Homework Review: Task 7

- Could do this as a 2-stage mapreduce job
  - Stage 1: count number of weekend/weekend days in dataset
  - Stage 2: same as assignment, but use results of stage 1 in computing averages

- Note you can use the weekday() function in python’s datetime package to determine what day of the week a given date is

Resources

Big Data FAQ Google doc (see link in announcements)

NYU Classes Forum

Reread previous labs/lab slides

HPC Wiki (see tutorial link in Lab 1!)

Apache Hadoop Streaming API, python API

Google, stackoverflow, etc.

Homework Review: Common Issues

Activity: get into groups of 2-4 people and talk about issues you ran into while doing this assignment

Deliverable: (due Wednesday, March 1, 6pm)

- Write a short report on your group discussion. Include:
  - Names of group members
  - For each group member:
    - A problem/bug/error they had to resolve (be specific)
    - How they resolved it (reading API, yarn logs, stackoverflow, office hours, etc.)

Note: each group member should submit a copy of your document (every group member submitting the same document is ok)