CS455 - Introduction To Distributed Systems

[Lab Session 9]

Brandon Gildemaster
Computer Science
Colorado State University

Colorado State University

Topics Covered in Today's Lab

- Quiz 6 review (taken before spring break)
- Logistics: Microsoft Teams, remote work, online quizzes
- FAQ and HW3 tips
- Hadoop - tips, cleanup/setup method
- Term project and deliverable 0

NOTE: Feel free to bring laptops, code, and questions!

Quiz 6 review

1. Consider the case where you are using the semaphore synchronizer to implement resource pools
    a. The semaphore must be initialized to twice the number of available resources
        i. False
    b. If you perform an acquire as opposed to a release when you are done using the resource, eventually you will have liveness issues with resource pool becoming unavailable.
        i. True
    c. You can skip the acquisition phase, and access the resource directly without violating correctness requirements.
        i. False

Quiz 6 review

2. Latches wait for other threads, while barriers wait for events.
    a. False
3. On-disk data accesses that are dominated by seeks are significantly faster than streaming through them.
    a. False

Quiz 6 review

4. Consider a particular reducer RX that has received 1000 intermediate outputs from mappers; these intermediate outputs correspond to 100 unique keys. The questions below pertain to execution of the reduce function in the reducer RX.
    a. It is possible that one of the mappers may not have generated intermediate outputs (<key, value> pairs) that need to be routed to RX.
        i. True
    b. The intermediate keys are sorted but not grouped (by key) before invoking the reduce function.
        i. False
    c. The keys are grouped (by key), but not sorted before invoking the reduce function.
        i. False

Quiz 6 review

4. Consider a particular reducer RX that has received 1000 intermediate outputs from mappers; these intermediate outputs correspond to 100 unique keys. The questions below pertain to execution of the reduce function in the reducer RX.
    d. The reduce function of RX is invoked exactly 100 times
        i. True
    e. If the reducer RX fails, all mappers must be re-executed to retrieve their intermediate outputs for the newly launched replacement for reducer RX.
        i. False

Logistics

- Microsoft teams login
  - Make sure you can login, email compsci_cs455@colostate.edu if not
  - To login, use <eid>@colostate.edu
  - Office hours
- Quiz time extension
- Please post any questions about quizzes, term project, programming component, working remotely, office hours, teams, etc...

FAQ

- To access WebUI of dfs remotely without port forwarding
  - Go to https://secure.colostate.edu
  - Authenticate with one of the three options
  - Then browse http://<NAMENODE_HOST>.cs.colostate.edu:<NAMENODE_PORT>
  - Shared cluster is http://augusta.cs.colostate.edu:7477
  - Hdfs web ui: http://augusta.cs.colostate.edu:50070
  - Allows you to see how many jobs are waiting to be run, why your job failed, etc…
- Hadoop setup video reminders
  - https://infospaces.cs.colostate.edu/watch.php?id=182
  - https://infospaces.cs.colostate.edu/watch.php?id=183

HW3-PC Tips

- Do not use shared cluster for debugging your programs
  - Copy a few files to your local cluster
  - Iron out bugs, then run on the full dataset
- Shared cluster can get slow if everyone is submitting jobs simultaneously
  - Reading the entire data set every time is slow
- There may be lines in the CSV file that are not well formatted
  - Code to handle these cases
  - It is acceptable to throw away a line if it is not properly formatted, or just a certain field in that line

HW3-PC Tips

- Be mindful of the amount of data you are transferring between mappers and reducers
  - Combiner
- Consider using the cleanup method
  - Available on Mapper and Reducer classes
  - Executes after all the data has been read and processed
  - Similar in function to combiner, but is guaranteed to run
  - Drawback: can increase memory usage, shouldn't typically be a problem
- Setup method also available, can initialize variables in here

Cleanup Example - word count

No cleanup

```
public static class Map extends Mapper<Object, Text, Text, IntWritable> {

    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
        StringTokenizer itr = new StringTokenizer(value.toString());
        while (itr.hasMoreTokens()) {
            String token = itr.nextToken();
            int length = token.length();
            word.set(length+"");
            context.write(word, one);
        }
    }
}
```

- Hadoop framework first calls the setup method
- Then performs the map task
- Finally, the cleanup method is called

Cleanup

```
public static class Map extends Mapper<Object, Text, Text, IntWritable> {
    Map count = new HashMap<Integer, Integer>();

    public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
        StringTokenizer itr = new StringTokenizer(value.toString());

        while (itr.hasMoreTokens()) {
            String token = itr.nextToken();
            int length = token.length();

            if(count.containsKey(length)) {
                int sum = (int) count.get(length) + 1;
                count.put(length, sum);
            }
            else {
                count.put(length, 1);
            }
        }
    }

    public void cleanup(Context context) throws IOException, InterruptedException {
        Iterator<Map.Entry<Integer, Integer>> temp = count.entrySet().iterator();

        while(temp.hasNext()) {
            Map.Entry<Integer, Integer> entry = temp.next();
            String keyVal = entry.getKey()+"";
            Integer countVal = entry.getValue();

            context.write(new Text(keyVal), new IntWritable(countVal));
        }
    }
}
```

Term Project Introduction

- Information released on course website
- Deliverable 0 is due Wednesday April 1st by 5:00 pm MT, form a team
  - **Email team composition to compsci_cs455@colostate.edu**
  - Find a team functionality on piazza
- Work in teams of 2-3 people (no exceptions)
- Everyone is a distance student now, communication is extremely important
  - Github
  - Microsoft teams
  - Be responsive to your team members messages

Term Project Introduction

- Deliverable 1 is a project proposal and is due April 10 by 5:00 pm MT
- Basic outline: Find a dataset -> ask an interesting question about it -> answer it
- “What will happen if we try x?” is not an interesting question (unless x is something very unique)
- “What is the performance of a neural network/random forest/SVM on this dataset?” is not an interesting question
- If you’re unsure whether your idea is a good one, ask

Term Project Introduction

- Finding good datasets is the hardest part - make sure you've found one that you can access before finalizing your proposal
  - At least >1 GB
  - Datasets of >100 GB start to become difficult to manage
    - Your team will be responsible for staging your dataset for your cluster
- Search for other projects/papers that have been done with your dataset - your project
- should be unique in some respect, but you can still borrow from others

Term Project Proposal - Datasets

- UCI Machine Learning repository is a good place to look for ideas but its datasets are often too small
- Kaggle tends to have a lot more datasets but often datasets are also too small, or they are so big you cannot download them and they can only be queried
- The US government releases a lot of data on traffic/weather/crime/etc…
- Alternatively look for a paper that does something similar to what you want to do and see if they made their data available

Term Project Proposal - Datasets

- Another option is to use a streaming data source such as Twitter, a video camera or something similar
  - Spark streaming
  - Make sure you're streaming enough data re require using a cluster to compute it - ask if you're unsure

Please post questions or discussion topics on piazza