CS 6240: Assignment 3

Goals: (1) Gain deeper understanding of grouping and aggregation in Spark. (2) Implement joins in Spark.

This homework is to be completed individually (i.e., no teams). You must create all deliverables yourself from scratch: it is not allowed to copy someone else's code or text, even if you modify it. (If you use publicly available code/text, you need to indicate what was copied and cite the source in your report!)

Please submit your solution as a *single PDF file* on Gradescope (see link in Canvas) by the due date and time shown there. During the submission process, you need to tell Gradescope on which page the solution to each question is located. Not doing this will result in point deductions. In general, treat this like a professional report. There will also be point deductions if the submission is not neat, e.g., it is poorly formatted. (We want our TAs to spend their time helping you learn, not fixing messy reports or searching for solutions.)

For late submissions you will lose one point per hour after the deadline. This HW is worth 100 points and accounts for 15% of your overall homework score. To encourage early work, you will receive a 10-point bonus if you submit your solution on or before the early submission deadline stated on Canvas. (Notice that your total score cannot exceed 100 points, but the extra points would compensate for any deductions.)

To enable the graders to run your solution, make sure your project includes a standard **Makefile** with the same top-level targets (e.g., *local* and *aws*) as the one presented in class. As with all software projects, you must include a **README** file briefly describing all the steps necessary to build and execute both the standalone and the AWS Elastic MapReduce (EMR) versions of your program. This description should include the build commands and fully describe the execution steps. This README will also be graded, and you will be able to reuse it on all this semester's assignments with little modification (assuming you keep your project layout the same).

You have about 2 weeks to work on this assignment. Section headers include recommended timings to help you schedule your work. The earlier you work on this, the better.

Important Programming Reminder

As you are working on your code, **commit and push changes frequently.** The commit history should show a natural progression of your code as you add features and fix bugs. Committing large, complete chunks of code may result in significant point loss. (You may include existing code for standard tasks like adding files to the file cache or creating a buffered file reader, but then the corresponding commit comment must indicate the source.) If you are not sure, better commit too often than not often enough. Here are some rules of thumb:

- Work for about 1 hour on your code, then commit your changes and include a comment explaining concisely what changed. Early on, when you produce a lot of initial code quickly, you may need to commit more frequently. Later, when you maybe spend 50 minutes on tracking down a single bug that results in a 1-line change in your code, you may commit less frequently. Use common sense: you want your commit history to tell a sufficiently detailed story about how you worked on your code, but you do not need to commit every single line you modify.
- To commit and periodically push your changes, use your git client, e.g., your IDE. There are explicit commit and push operations, and the git client will ask you for a comment summarizing the changes you made. You must execute those commit/push commands—simply saving your code from your IDE will not automatically commit it. The commit comment is a short explanation, e.g., "added reduce function for word count" or "fixed memory error in RDD join."

Repeat this work-commit-push cycle until you are done. This way we see your progress in terms of the intermediate commits. If you simply commit a complete program in a single step, it looks like you just got the solution from somebody else, and this may trigger an academic integrity investigation.

Combining in Spark (First 2/3 of Week 1)

The first part of this assignment compares different implementations of the Twitter-follower counting problem in Spark. We only work with the edges.csv data. Like in the first assignment, your programs should output the number of followers for each user who has at least one follower and whose ID is divisible by 100, returning output formatted with each user and follower count in a different line:

(userID1, number_of_followers_this_user_has)
(userID2, number_of_followers_this_user_has)

This time we only focus on Spark, but you need to implement different programs:

RDD-G: This program only uses RDD and pair RDD, but not DataSet or DataFrame. The grouping and aggregation step must be implemented using groupByKey, followed by the corresponding aggregate function.

RDD-R: This program only uses RDD and pair RDD, but not DataSet or DataFrame. The grouping and aggregation step must be implemented using reduceByKey.

RDD-F: This program only uses RDD and pair RDD, but not DataSet or DataFrame. The grouping and aggregation step must be implemented using foldByKey.

RDD-A: This program only uses RDD and pair RDD, but not DataSet or DataFrame. The grouping and aggregation step must be implemented using aggregateByKey.

DSET: The grouping and aggregation step must be implemented using DataSet, with groupBy on the appropriate column, followed by the corresponding aggregate function.

Joins in Spark (Last 1/3 of Week 1, All of Week2)

Solve the Twitter-follower *triangle*-counting problem from Homework 2, but this time in Spark Scala. Write four different versions of the program:

1. RS-R implements the equivalent of Reduce-side join (hash+shuffle) using RDD and pair RDD only.
2. RS-D implements the equivalent of Reduce-side join (hash+shuffle) using DataSet or DataFrame only.
3. Rep-R implements the equivalent of Replicated join (partition+broadcast) using RDD and pair RDD only.
4. Rep-D implements the equivalent of Replicated join (partition+broadcast) using DataSet or DataFrame only.

Hints and requirements:

- For the DataSet/DataFrame programs, it is allowed to load the data initially as an RDD, but then the actual join must be applied to DataSets/DataFrames.
- Do not use SparkSession.sql(someSQLquery) to implement the join. Instead, use functions such as join, joinWith, map, filter, flatMap etc.
- For functions such as join and joinWith, find out if they implement hash+shuffle and partition+broadcast. If they do, use them accordingly.
    - DataSet/DataFrame: You may need to explore optimizer hints and broadcast-size-threshold settings as discussed in class. If despite your best efforts the optimizer keeps choosing only one, but never the other join strategy, then report all settings you tried and explain which strategy was chosen by the optimizer.
    - (Pair) RDD: Last time we checked, the join was implemented using hash+shuffle, but not partition+broadcast. Verify if this still holds true. If so, you need to implement partition+broadcast in user code as shown in class.
        - When using broadcast(), check that the data structure you broadcast is properly representing the edge-RDD information. For instance, for a hashmap or similar, can the same key appear with different values? If not, then you need to find a workaround, e.g., encode (k1, v1), (k1, v2), (k1, v3) as (k1, (v1, v2, v3)).
- Like for the MapReduce program, you may use the MAX-filter idea to reduce data size by eliminating all input records for users with large IDs, if you believe you cannot resolve excessive memory/storage consumption or running time in any other way.
- It is perfectly acceptable to search for example programs in textbooks and on the Web to get inspiration and resolve syntax issues, as long as you cite them in report and source code. E.g., look at the Spark textbook join chapter we reference in the module.

We strongly encourage collaboration via Piazza for all Spark syntax issues, especially questions like "which data structure should I use to broadcast the edges list," "what is the syntax to use map() to look up values associated with a specific key in that data structure" and so on.

Report

Write a brief report about your findings, answering the following questions:

1. [10 points] Show the pseudo-code for all 5 Twitter-follower-count programs (RDD-G, RDD-R, RDD-F, RDD-A, DSET) in Spark Scala. Since many Scala functions are similar to pseudo-code, you may copy-and-paste well-designed (good variable naming!) and well-commented Scala code fragments here.
2. [10 points] Show the link(s) to the source code for these programs in your Github Classroom repository.
3. [10 points] Run these programs *locally* (not on AWS) and determine if Spark performs aggregation before data is shuffled, i.e., the equivalent of MapReduce's Combiner. Look for evidence that supports your answer, e.g., by using toDebugString() and explain(), looking at the log files, and consulting the Spark documentation. For each of the 5 programs, (1) state clearly if it aggregates before shuffling (like a Combiner) and (2) present evidence to support your answer.
4. [30 points] Show the pseudo-code for the 4 triangle-counting programs, including MAX-filter functionality. Since many Scala functions are similar to pseudo-code, you may copy-and-paste well-designed (good variable naming!) and well-commented Scala code fragments here.
    a. [10 points] RS-R
    b. [5 points] RS-D
    c. [10 points] Rep-R
    d. [5 points] Rep-D
5. [16 points] Show the link(s) to the source code for these programs in your Github Classroom repository.
6. [8 points] Run each triangle-counting program on EMR using 1 master and 4 worker nodes, using the same machine type and MAX setting as for the corresponding MapReduce program (RS-R and RS-D like RS-join in HW 2; Rep-R and Rep-D like Rep-join in HW 2). If a program runs faster than 15 min for the same settings as HW 2, simply report the faster time. If a program runs more than twice as long as the corresponding MapReduce version, you may choose a lower MAX setting. If the Spark version crashes with out-of-memory error for the same settings where the MapReduce program worked fine, explore Spark parameters to control memory size (container size, heap size), but do not use larger machine instances. If you still cannot resolve the memory issue, report the settings you tried and fix the problem by using a lower MAX. For each of the 4 programs (RS-R, RS-D, Rep-R, Rep-D), report machine type, MAX value, running time, and triangle count. For comparison, also report machine type, MAX, running time, and triangle count for the corresponding join versions from HW 2. The machine type must be the same. If you changed MAX, briefly explain why (see discussion above).
7. [4 points] Show the link to the stderr file and the file that contains the triangle count for the run that produced the above running-time numbers for each of the 4 programs (RS-R, RS-D, Rep-R, Rep-D).
8. [8 points] Run the triangle-counting programs on the larger cluster with 1 master and 8 workers, following the same instructions (same configuration as for HW 2; change MAX only if necessary).

For each of the 4 programs (RS-R, RS-D, Rep-R, Rep-D), report machine type, MAX, running time, and triangle count. For comparison, also report machine type, MAX, running time, and triangle count for the corresponding join versions from HW 2. The machine type must be the same. If you changed MAX, briefly explain why (see discussion above).

9. [4 points] Show the link to the stderr file and the file that contains the triangle count for the run that produced the above running-time numbers for each of the 4 programs (RS-R, RS-D, Rep-R, Rep-D).

Important Notes

Choose carefully where you host the log and output files. You can copy them to the corresponding Github Classroom repo, but Github has a size limit for version-managed files and things can get tricky. If you leave the files on S3, make sure the TAs and instructor have access to them and keep in mind that Amazon charges per GB of data on S3. So, possibly the easiest solution would be to use Google Drive, Dropbox or similar. Again, make sure the TAs and instructor (but nobody else) can access the corresponding directories and files.

Check that the log file is not truncated—there might be multiple pieces for large log files!

The **submission time** of your homework is the *latest timestamp of any of the deliverables included*. For the PDF it is the time reported by Gradescope; for the files on Github it is the time the files were pushed to Github, according to Github. If you want to keep things simple, do the following:

1. Push/copy all requested files to github and make sure everything is there. (Optional: Create a version number for this snapshot of your repository.)
2. Submit the report on Gradescope. Open the submitted file to verify everything is okay.
3. Do not push any more changes to the files for this HW on Github.

If you *cannot get your program to run on AWS*, then you can instead include the log files and output from execution on your local machine for partial credit.

Please ensure that your code is properly documented. There should be comments concisely explaining the role/purpose of a class. Similarly, if you use carefully selected keys or custom Partitioners, make sure you explain their purpose (what data will be co-located in a Reduce call; does input to a Reduce function have a certain order that is exploited by the function, etc.). But do not over-comment! For example, a line like “SUM += val” does not need a comment. As a rule of thumb, you want to add a brief comment for a block of code performing some non-trivial step of the computation. You also need to add a brief comment about the role of any major data structure you introduce.