AMS 598: Big Data Analysis (Fall 2024)

Project #2
Due Oct 17th, 2024

---

1. You need to submit (1) a report in PDF and (2) your .ipynb code file, both to Brightspace.
2. Your PDF report should include results and analysis of the programming part. For the programming part, your PDF report should at least include the results you obtained. You should also analyze your results as needed.
3. Please put all your files (PDF report and code files) into a compressed file named "Proj#_FirstName_LastName.zip"
4. Unlimited number of submissions are allowed on Brightspace and the latest one will be timed and graded.
5. All students are highly encouraged to typeset their reports using Word or LaTeX. In case you decide to hand-write, please make sure your answers are clearly readable in scanned PDF.
6. No starting code is given for this project.
7. Please read and follow submission instructions. No exception will be made to accommodate incorrectly submitted files/reports.
8. Please start your submission to Brightspace at least 15-30 minutes before the deadline, as there might be latency. We do NOT accept E-mail submissions.

---

1. Implement Grouping in Relational Algebra using MapReduce

Overview: In this assignment, you will perform the **grouping** operation in Relational Algebra using the **MapReduce** programming paradigm. You will group a dataset based on the `customer_id` and calculate the total amount spent by each customer.

Problem Illustration: You are given a **Sales** dataset with the following schema:

```
Sales(customer_id, product_id, amount)
```

The dataset represents transactions, where:

- `customer_id`: ID of the customer.
- `product_id`: ID of the product bought.
- `amount`: The amount spent by the customer in this transaction.

You will use MapReduce to:

1. **Group by** `customer_id`.
2. Calculate the total `amount` spent by each customer.

Data Example: Consider the following sample data stored in a text file `sales.txt`:

```
customer_id,product_id,amount
1,101,50
2,102,30
1,103,70
3,101,20
2,101,20
1,101,40
```

The final output should be:

```
1    160
2    50
3    20
```

Data Location: /gpfs/projects/AMS598/projects/proj2.

Tasks:

1. Use the MapReduce concept and SeaWulf to tackle the problem.
2. Explore different numbers of Mappers and Reducers.
3. Explore and implement at least one optimization strategy.
4. Write a report about the analysis.

2. Natural Join of Three Tables (R1, R2, R3) Using MapReduce

Overview: In this assignment, you will implement a natural join operation using the MapReduce programming model to combine three large datasets (tables). Each table shares a common attribute, and your task is to write a MapReduce program that performs the join across all three tables.

Problem Illustration: You are given three tables:
1. Table R1 with schema R1(A, B, C).
2. Table R2 with schema R2(A, D, E).
3. Table R3 with schema R3(A, F, G).

The three tables share the common attribute A, and your task is to perform a natural join on attribute A, resulting in a table with schema (A, B, C, D, E, F, G). Intuitively, the goal is to output rows where R1.A = R2.A = R3.A, combining the corresponding rows from all three tables.

Data Location: /gpfs/projects/AMS598/projects/proj2.

Tasks:

1. Use the MapReduce concept and SeaWulf to tackle the problem.
2. Explore different numbers of Mappers and Reducers.
3. Explore and implement at least one optimization strategy.
4. Write a report about the analysis.