**Project 1 : Big Data Management & Analytics - C Term 2024**

*“Are you My Friend Analytics”*

|  |  |
| --- | --- |
| **Total Points:** | **50** |
| **Given Out:** | **Week 2: Thursday, Jan 18th, 2024** |
| **Due Date:** | **Week 4: Thursday, Feb 1, 2024 (8:59PM)**  **Submit the project via CANVAS.** |
| **Teams:** | **Project is to be done in assigned teams (posted on CANVAS).** |

**Project Overview**

In this project, you will work on the Hadoop big data ecosystem that you have set up in project 0. You will load a large social media data set called “Facebook” into Hadoop HDFS. Next, you analyze the data in a scalable fashion by writing custom analytics tasks using map-reduce Java codeand execute those on your scalable Hadoop platform.

**Project Submission:**

1. **Project Code:** One single member of your team will submit a **zip file** containing the code for your MapReduce analytics via CANVAS for your team. You are welcome to submit the zip file of your IDEA project directory. However, makesure you **don’t include any large dataset** (a very small test dataset is fine). Please don’t submit a .jar file.
2. **Documentation and Analysis:** You will also submita document (pdf) called REPORT containing your documentation that describes how you accomplished each task. Discuss why your solution is good and scalable. *Keep in mind that it is not just important that it “runs”, it also should be a scalable solution.* Please include screenshots of (of a few records only) of your generated result files on HDFS (additional details will follow below).
3. **Contribution Statement:** In your project REPORT, you need to include an explicit statement labeled “Contribution Statement” that indicatesthe *relative contributions of each team member explicitly*. [[1]](#footnote-1) This requires you to discuss your expectations and how best you can work together and help each other succeed at this first project at the project kick-off and continue actively this communication throughout this project period. *By submitting, all team members confirm the division of labor as indicated in your REPORT.*
4. **Resource Usage Statement:** *You need to include an explicit “Resource Usage Statement” in your REPORT* that explicitly lists for each team member if or if not, how, and what resources you used. I encourage responsible and sensible use of *Generative AI (e.g., ChatGPT)* and other sources for this project assignment. If you do not do the work yourself, you will not understand how it works and you will learn less about this technology. In particular, you are required to report and cite all resources that you used to complete the assignment. If there is any usage of Generative AI, in particular, you are expected to explain how you validated the trustworthiness of the solution, which prompt/s you used, and what outputs the model gave you and how you adjusted and validated it (basic code/documentation/etc.).
5. **Individual Peer Feedback:** Everyonesubmits **individually** their feedback about your contributions in relation to those of your team members into the google survey LINK below. *Did you help your team? did they help you? etc.* You need to your peer assessment feedback to:

<https://forms.gle/R59VLP5PqqQqUABB8>

**by the project due date.** This team feedback is confidential; and only visible to the instructional staff.

**Project Demonstration.**

Once completed, one or at most two teams may be asked to provide a brief demonstration of their results to your classmates to review how you solved this project. *If necessary for grading, the instructor/TA will communicate with your team about your project, and also may request a demonstration of your solution.*

**Project Description**

**Documentation of your Work [10 pts]**

We expect you to submit a clear documentation of your work in a file named REPORTS following the above guidelines. This includes a description of the key components below:

* 1. Data Set Preparation and loading of test data set into Hadoop system
  2. Documentation of the logic of each of your Analytics Queries
  3. Contribution Statement by the team covering all team members
  4. Resource Usage Statement by the team covering all team members

For the data, use “datagen.zip” to produce a Facebook dataset. You then use hadoop file system commands (e.g., put) to upload your data files into Hadoop cluster to set up your data set into distributed storage.

**To get credits, submit below:**

1. Submit a screenshot of the terminal output (just a few lines).
2. HDFS Status Page (showing file blocks).

To learn about the file system commands, review the link here:

<https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html>

Note: It is good to check your files and see how the files are divided into blocks and each block is replicated. You can do that by checking the web interface of Hadoop. Check the Readme file in your virtual machine to know to do that.

**Dataset for Project 1: The FaceBook Big Data Application**

We provide you with a python data generator “datagen.zip” that creates data sets for your Facebook-like application. It generates a “Facebook dataset” that contains three separate data files: ***MyPage,***  ***Friends,***  and ***AccessLogs.***

Each line in the **MyPage** file represents one person, and includes the attributes describing the person as listed below.

Each line in the ***Friends*** dataset file describes which person p1 has indicated that they are friends with another person p2 (this is a one-directional relationship from p1 to p2) and the timestamp when this friend relationship was declared.

Each line in the ***AccessLog*** data file indicates which person p1 has accessed the Facebook page that belongs to a second person p2, including the time stamp of the access.

The datasets have the following attributes. The attributes within each line are comma separated (CSV format). Find a snapshot of a very small sample data set on CANVAS under projects.

The ***MyPage*** dataset has the following attributes for each Facebook page:

|  |  |
| --- | --- |
| ID: | unique sequential number (integer) indicating the owner of the page |
| Name: | characters of length between 10 and 20, |
| Nationality: | characters of length between 10 and 20, |
| CountryCode: | number (integer) between 1 and 50, |
| Hobby: | sequence of characters of length between 10 and 20. |

The ***Friends*** dataset has the following attributes for each friend relationship:

|  |  |
| --- | --- |
| FriendRel: | unique sequential number (integer) taken from value in the range from 1 to 20,000,000 |
| PersonID: | Person-ID of a person p1 who has a Facebook page, |
| MyFriend: | References ID of a person p2 that also has a Facebook page and that the person p1 listed in attribute PersonID says they are friends with. This relation is not mutually necessarily, i.e., it just indicates that p1 declares that they are friends with p2. Also, person p1 needs to be distinct from person p2. |
| DateofFriendship: | random number (integer; or some other sequential data type to use as date) between 1 and 1,000,000 to indicate when the friendship started |
| Desc: | text of characters of length between 20 and 50 explaining the type of friendship: college-friend, family, etc. |

The ***AccessLog*** dataset has the following attributes for each Facebook access:

|  |  |
| --- | --- |
| AccessId: | unique sequential number (integer) from 1 to 10,000,000 |
| ByWho: | References the Id of the person who has accessed the Facebook page |
| WhatPage: | References the Id of the page that was accessed |
| TypeOfAccess: | text of characters of length between 20 and 50 explaining if just viewed, left a note, added a friendship, etc. |
| AccessTime: | random number between 1 and 1,000,000 (or epoch time) |

**Note:** A column name does not include a comma. The column names are not stored in the file. Only the values are listed. Each value should be separated by a comma. From the order of the columns; you will know what column a value represents.

**Analytics Tasks using MapReduce Jobs [40 Points]**

You will write Java programs to realize the following tasks on your data to analyze your data. Before writing your code and/or queries, review the “WordCount” example. It is part of your project 0 package that you worked with in week 1 of this course. It has also been discussed in the class.

**General Guidelines:**

* Learn how Hadoop reads and writes integers, floats, and text fields. Check IntWritable, FloatWritable, and Text classes to know which one to use.
* You should determine whether a query is a map-only job, a map-reduce job, or several map-reduce jobs. If a task can be done with a simpler solution, then you should describe this simpler solution to get full credit.
* Develop a solution with and without a map-reduce combiner, when possible. If not possible for your query, please state so explicitly in your description. Work with as many features of Hadoop as possible (e.g., such as to control how many mappers are used) to get to know Hadoop. Explain explicitly any ideas you have tried out.
* Report the performance for the execution each of your tasks below. In particular, compare the relative performance of different solutions.
* Document each of your tasks in the report.

**Point distribution:** You will get up to 5 pts per task (plus, possibly one extra bonus point). Deliverables for each task include:

* A (simple) working solution with its explanation [4 pts],
* a description of an advanced/optimized solution (such as combiner) that is more optimized and thus efficient than your first solution, or a clear description that your above solution is already scalable and a better solution is not possible [1 pt],
* in the case that you have identified an effective optimized solution that is more advanced than your first solution below, then you can get extra credit by providing its working implementation [Bonus of 1pt].

**TIP:** You want to test things on a small test file first and check the query output file from the HDFS website to make sure your answer is correct, before running it on a larger dataset.

**Notes:**

1. You can consider using a Combiner in your MapReduce jobs to reduce the IO costs. However, make sure you understand when you can use a Combiner and when you cannot. Not all MapReduce jobs can be optimized with Combiner. Combiner could deliver incorrect results in some cases. For example, if you are doing a Map only job, Combiner will not have any effect.
2. You can compare the performance of both the Reduce-side join and replicated (Map-side) join in your MapReduce job to determine the most optimized option.
3. Consider optimizing your solution by minimizing the use of multiple MapReduce jobs, when possible. For example, if you can finish the task with a Map Only job, then you should skip the reducer to save I/O costs.

**Task a**

Report all Facebook users (name, and hobby) whose Nationality is the same as your own Nationality (pick one, e.g., “American” or “Indian”, etc.).

**Task b**

Find the top 10 popular Facebook pages, namely, those that got the most accesses based on your AccessLog dataset compared to all other pages. Return their Id, Name and Nationality.

**Task c**

Write job(s) that reports for each country, how many of its citizens have a Facebook page

**Task d**

For each Facebook page owner p2, compute the “connectedness factor” of p2. That is, for each person p2 in your dataset, report the owner p2’s name, and the number of people p1 that are listing p2 as their friend. For page owners that aren't listed as anybody's friend, i.e., they are not in the Friends file, return a count of zero.

**Task e**

Determine which people have favorites. That is, for each Facebook page owner, determine how many total accesses to Facebook pages they have made (as reported in the AccessLog) and how many distinct Facebook pages they have accessed in total.

**Task f**

Identify people p1 that have declared someone as their friend p2 yet who have never accessed their respective friend p2’s Facebook page and return the return the PersonID and Name of the person p1. (This may indicate that they don’t care enough to find out any news about their friend -- at least not via Facebook).

**Task g**

Identify all "disconnected" people (and return their personID and Name) that have not accessed the Facebook site for 14 days or longer (i.e., meaning no entries in the AccessLog exist in the last 14 days).

**Task h**

Report all people with a Facebook who are more ‘popular’ than the average person on the Facebook site, namely, those who have more friendship relationships than the average number of friend relationships per person across all people in the site.

1. For instance, if each team member has done the project independently, and then only at the end you pulled the best of the material together, you need to say so. Or, if one team member helped and taught the second team member how to do it, and then the 2nd team member succeeded to do some of the queries (even if with some guidance), please report this. If you have closely collaborated and done the same amount of effort working side by side helping each other, also please explain. [↑](#footnote-ref-1)