DBMS Models and implementation
Instructor: Sharma Chakravarthy
Project 3: Big Data Analytics (Map/Reduce implementation)

Made available on: 4/10/2016
Due on: 5/6/2016 (11:55Pm)
Submit by: Blackboard (1 zipped folder containing all the files/sub-folders) https://elearn.uta.edu/
Weight: 15% of total
Total Points: 100

One of the advantages of cloud computing framework and the use of Map/Reduce paradigm is its ability to deal with very large data sets and one-time computations with a reasonable response time. This is done by using as many processors as needed. Typically, the map/reduce paradigm is used for these types of problems in contrast to the RDBMS approach which is better suited for storing, managing, and manipulating this data. Hadoop is a widely used open source map/reduce platform. Hadoop Map/Reduce is a software framework for writing applications which process vast amounts of data in parallel on large clusters. In this project, you will use the US census dataset and develop a program to compute a number of demographic statistics using Hadoop map/reduce paradigm. Please use the following links for a better understanding of Hadoop and Map/Reduce.

https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

1. **Dataset: consist of US census data for years 2009 to 2013.**
   i. Dataset size: 8 GB (two files and can be downloaded using the link given below)
   ii. Coverage: all of US
   iii. Large number of characteristics per person

We will only be using a few items from each record, such as id, gender, weight, age, salary, race, and state. The rest of the data will be ignored.

Column No.	Attribute name	Description
0	SERIALNO	Unique identifier. Eg. **200900000001-201399999999** the first four digits represents year. There are five years – 2009, 2010, 2011, 2012, 2013
5	STATE	State Code; 1. Alabama/AL, 2. Alaska/AK 4. Arizona/AZ, 5. Arkansas/AR, 6. California/CA, 8. Colorado/CO, 9. Connecticut/CT, 10. Delaware/DE, 11. District of Columbia/DC, 12. Florida/FL, 13. Georgia/GA, 15. Hawaii/HI, 16. Idaho/ID, 17. Illinois/IL, 18. Indiana/IN, 19. Iowa/IA, 20. Kansas/KS, 21. Kentucky/KY, 22. Louisiana/LA, 23. Maine/ME, 24. Maryland/MD, 25. Massachusetts/MA, 26. Michigan/MI,

		27. Minnesota/MN, 28. Mississippi/MS, 29. Missouri/MO, 30. Montana/MT, 31. Nebraska/NE, 32. Nevada/NV, 33. New Hampshire/NH, 34. New Jersey/NJ, 35. New Mexico/NM, 36. New York/NY, 37. North Carolina/NC, 38. North Dakota/ND, 39. Ohio/OH, 40. Oklahoma/OK, 41.Oregon/OR, 42. Pennsylvania/PA, 44. Rhode Island/RI, 45. South Carolina/SC, 46. South Dakota/SD, 47. Tennessee/TN, 48. Texas/TX, 49. Utah/UT, 50. Vermont/VT, 51. Virginia/VA, 53. Washington/WA, 54. West Virginia/WV, 55. Wisconsin/WI, 56. Wyoming/WY, 72. Puerto Rico/PR *The codes are not in sequence.*
7	PWGTP	Person's weight; 1 to 9999 weight of person (in kilogram)
8	AGEP	Age; 00. Under 1 year, 01-99. 1 to 99 years
9	CIT	Citizenship status; 1. Born in the U.S., 2. Born in Puerto Rico, Guam, the U.S. Virgin Islands, .or the Northern Marianas, 3. Born abroad of American parent(s), 4. U.S. citizen by naturalization, 5. Not a citizen of the U.S. *The codes are in sequence.*
69	SEX	Sex; Code is 1. Male, 2. Female
72	WAGP	Wages or salary income past 12 months; bbbbbb. N/A (less than 15 years old) **[Ignore]**, 000000. None **[Ignore]**, 000001-999999. $1 to 999999
97	RAC1P	Recoded detailed race code; 1. White alone, 2. Black or African American alone, 3. American Indian alone, 4. Alaska Native alone, 5. American Indian and Alaska Native tribes specified; or American .Indian or Alaska Native, not specified and no other races, 6. Asian alone, 7. Native Hawaiian and Other Pacific Islander alone, 8. Some Other Race alone, 9. Two or More Races *The codes are in sequence.*

There are 295 attributes. **All the attributes are *comma* separated**. The above table provides description of few attributes needed for your project. Use the appropriate attribute to solve the problem. You can ignore rest of the attributes. ***Also, Ignore null values.***

The above data set can be downloaded from (right click on the link below). Contains 2 files each about 4GB.

http://itlab.uta.edu/downloads/cse5331_Project3_census_datasets.zip (1.25 GB, compressed size)

2. Problem Specification:

i. You need to compute the weight trend over a few years for different states along genders using the given data. You need to develop a map/reduce solution.

For each of the following states, calculate average weight of males and females for each year and plot on a graph with year as the X-axis and avg. weight as the Y-axis. Note that this can be done for all states as well as in other ways. I have chosen 7 states to compare these trends: I) California, ii) Colorado, iii) North Dakota, iv) Texas, v) New York, vi) Maine, and vii) Florida.

ii) If you end up doing this project using different numbers of mappers and reducers, you can also visualize and understand the response time improvements when you use more number of mappers and reducers (divide and conquer)

The original data size is even larger and we are using only 5 years.

For this problem, map/reduce can be beneficially used. It will be easy to adjust the number of mapper and reducer nodes based on the data size and the Hadoop framework will do most of the work in partitioning (sharding) the data and passing intermediate output from mappers to reducers. It will also take care of failures of nodes etc.

You need to design and develop a map program (including a combiner if needed) and a reduce program to solve the above problem. The most important aspects of this design will be to identify the key value pairs to be output by the mapper and worked on by the reducer.

If you want to do ii) above to understand and appreciate the power of this paradigm and the ease of scaling using this paradigm, you will run the same data set:

i. On multiple mappers and to see how parallelism works
ii. Measure response time for different configurations and see how scalability is accomplished.

The input data is typically partitioned (sharded) into 64MB splits as default. The number of splits can also be configured by the user. Similarly the number of map and reduce nodes can also be configured. Remember that the number of map tasks are determined by the number of shards. Number of mappers (or mapper nodes) used determine how many map tasks are assigned to a mapper. The number of Reduce tasks can also be different from the number of reduce nodes (based on the problem needs). The number of mapper and reducer nodes need not be the same.

The purpose of this project is to understand the design and development of map/reduce programs as well as the speed up obtained by using different number of map and reduce nodes. You can compare the alternate configurations used and draw your conclusions in the report. The graph will show how weight trends vary and whether it differs significantly based on the region.

3. **Installation:**

To complete the project you will need a Hadoop installation. This can be done in one of 2 ways:

Hadoop is available at http://hadoop.apache.org/. You may install Hadoop single node cluster on your own computer. Detailed guide (both text and video) for installing Hadoop on your Linux box can be downloaded from (right click on the link)
http://itlab.uta.edu/downloads/pdf_Hadoop_Installation_Guide.zip (1.4MB)

http://itlab.uta.edu/downloads/video_Hadoop_Installation_Guide.zip (50MB)

The second option is to use Amazon Elastic Map/Reduce. Amazon EMR (Elastic Map/Reduce) is a web service provided by Amazon that uses Hadoop and distributes large datasets and processes them into multiple EC2 instances. You have to sign up for AWS. Please make sure you read carefully your AWS agreements/contracts/free use. You may have sufficient free services to complete the projects, but you should monitor and understand your use. Don't leave things running when not necessary. Note that if you exceed your monthly quota, you will get charged. Also, a credit/debit card is necessary for signing up for this service. Detailed guide for executing a simple map/reduce program on Amazon EMR can be downloaded from (right click on the link)

http://itlab.uta.edu/downloads/pdf_AWS_Hadoop_EMR_Guide.zip (600KB)

Option 2 above is preferred as you will have access to more number of mapper and reduce nodes than on your laptop. However, make sure you kill your jobs properly to avoid any extra charges! If you have any difficulty in setting up an Amazon account, please talk to the TA.

4. **Project Report:**

Please include (at least) the following sections in a **REPORT.{txt, pdf, doc}** file that you will turn in with your code:

i. **Overall Status**

Give a *brief* overview of how you implemented the major components. If you were unable to finish any portion of the project, please give details about what is completed and your understanding of what is not. (This information is useful when determining partial credit.)

ii. **Performance Measure:**
Please include all the methods that you used to measure and compare performance of the results.

iii. **File Descriptions**
List the files you have created and *briefly* explain their major functions and/or data structures.

iv. **Division of Labor**
Describe how you divided the work, i.e. which group member did what. Please also include how much time each of you spent on this project. (This has no impact on your grade whatsoever; we will only use this as feedback in planning future projects -- so be honest!)

v. **Logical errors and how you handled them**:
List at least 3 logical errors you encountered during the implementation of the project. Pick those that challenged you. This will provide us some insights into how we can improve the description and forewarn students for future assignments.

5. What to submit:

- After you are satisfied that your code does exactly what the project requires, you may turn it in for grading. Please submit your project report with your project.
- You will turn in one zipped file containing you source code as well as the report
- All of the above files should be placed in a single zipped folder named as - 'proj3_firstname_lastname_Section_final'. **Only one zipped folder should be uploaded using blackboard.**
- You can submit your zip file at most 5 times. The latest one (based on timestamp) will be used for grading. So, be careful in what you turn in and when!
- **Only one person per group should turn in the zip file!**
- Three days after the due date, the submission will be closed

6. Coding style:

Be sure to observe the following standard Java naming conventions and style. These will be used across all projects for this course; hence it is necessary that you understand and follow them correctly. You can look this up on the web. Remember the following:

i. Class names begin with an upper-case letter, as do any subsequent words in the class name.

ii. Method names begin with a lower-case letter, and any subsequent words in the method name begin with an upper-case letter.

iii. Class, instance and local variables begin with a lower-case letter, and any subsequent words in the name of that variable begin with an upper-case letter.

iv. No hardwiring of constants. Constants should be declared using all upper case identifiers with _ as separators.

v. All user prompts (if any) must be clear and understandable

vi. Give meaningful names for classes, methods, and variables even if they seem to be long. The point is that the names should be easy to understand for a new person looking at your code

vii. Your program is properly indented to make it understandable. Proper matching of if … then … else and other control structures is important and should be easily understandable

viii. Do not put multiple statements in a single line

In addition, ensure that your code is properly documented in terms of comments and other forms of documentation for generating meaningful javadoc.

7. Grading scheme:

The project will be graded using the following scheme:

1.	Correctness of the Map and Reduce code	60
2.	Correct results for multiple configurations	10
3.	Analysis of results	15
4.	Report	15

8. Class presentation on projects

Since this is the last project, each team (both members where applicable) will make a presentation of their project experience in the class on May 3rd and May 5th 2016. You will answer any questions on the project experience. Download the template slides with questions from (right click on the link). Do not modify the format and layout! Please rename the file to include your team number. Presentations will be in order of team numbers. Please send these slides to me and TA by May 2nd 2016, 7pm.

use 4331_5331_Project_Presentation..ppt slides from bb under project 3

This presentation is a requirement and counts towards class participation.