**<Last Name>, <First Name>**

<Date>

Harvard University Extension - Principles of Big Data Processing e88

Homework 4: Hadoop MapReduce

* **Make sure to submit all your source code (.java files , .py files or whatever language you are using) - in a separate archive, named <LastName>\_<FirstName>\_HW4.zip**
* **Make sure to add full result files into that archive as well**

Please identify which problems were completed. If any were incomplete, please identify where you encountered problems.

|  |
| --- |
| *for example:*  Problem 1: x% complete  Problem 2: x% complete  Problem 3: x% complete  Problem 4: x% complete  Problem 5: Bonus: x% complete |

**Problem 1: [15 points] running example wordcount job with S3 data**

Include a screenshot of your S3 bucket with the 4 input log files [5 points]

|  |
| --- |
|  |

Show command you used to run the example wordcount job with your S3 data, and the job output [5 points]

|  |
| --- |
|  |

Show the number of splits, map and reduce tasks for this job - from the job console output or log files [5 points]

|  |
| --- |
|  |

**Problem 2: unique counts** [points: 40]

Paste your source code of the MR jobs into the following area [10 points]

|  |
| --- |
|  |

Show commands you used to run your jobs in the AWS EMR; include screenshots of the console output and the summary of each job (see example) [10 points]

|  |
| --- |
| Example excerpt of a MR job console output:  >>>  Map-Reduce Framework  Map input records=975000  Map output records=975000  Map output bytes=15600000  Map output materialized bytes=484  Input split bytes=640  Combine input records=975000  Combine output records=20  Reduce input groups=5  Reduce shuffle bytes=484  Reduce input records=20  Reduce output records=5  Spilled Records=40  Shuffled Maps =12  Failed Shuffles=0  Merged Map outputs=12  GC time elapsed (ms)=1124  CPU time spent (ms)=30280  Physical memory (bytes) snapshot=2970128384  Virtual memory (bytes) snapshot=27007504384  Total committed heap usage (bytes)=2598895616  Shuffle Errors …. |

Show how many part-r-xxxxx result files you've got for each job - explain why the specific number? [5 points]

|  |
| --- |
|  |

Show results of your jobs for the specified keys - from the relevant result files [15 points]

|  |
| --- |
| Query 1:  <date\_hour>, <url\_count>  2019-09-12:13, ?? 2019-09-12:14, ?? 2019-09-12:15, ?? 2019-09-12:16, ?? 2019-09-12:17, ??  Query 2  <date:hour:url>, unique\_user\_count  2019-09-12:02:http://example.com/?url=003, ?? 2019-09-12:02:http://example.com/?url=004, ?? 2019-09-12:02:http://example.com/?url=005, ?? 2019-09-12:02:http://example.com/?url=006, ??  Query 3  <date:hour:url>, event\_count  2019-09-12:02:http://example.com/?url=003, ?? 2019-09-12:02:http://example.com/?url=004, ?? 2019-09-12:02:http://example.com/?url=005, ?? 2019-09-12:02:http://example.com/?url=006, ?? |

**Problem 3: time range queries** [points: 30]

Paste your source code into the following area [15 points]

|  |
| --- |
|  |

Include screenshots of the console output and the job summary executed in the AWS EMR [10 points]

|  |
| --- |
|  |

Show results of your query for the specified keys [5 points]

|  |
| --- |
| <date,hour,country>, url\_count  2019-09-13:19,IQ, ?? 2019-09-13:19,IR, ?? 2019-09-13:19,IS, ?? 2019-09-13:19,IT, ?? 2019-09-13:19,JE, ?? |

**Problem 4: [15 points] MR Jobs configurations and UI**

show command you run to execute your jobs with different reduce job settings - and console output of each job [5 points]

|  |
| --- |
|  |

compare results of each job - explain the difference [5 points]

|  |
| --- |
|  |

include screenshots of the EMR ResourceManager UI with each job' statistics - explain differences if any [5 points]

|  |
| --- |
|  |

**Problem 5: Bonus: Top N queries** [15 points]

Paste your source code into the following area [5 points]

|  |
| --- |
|  |

What are the main differences with the Problem 3 and 4 implementation? [5 points]

|  |
| --- |
|  |

Show results of your query [5 points]

|  |
| --- |
|  |