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

<Date>

Harvard University Extension - Principles of Big Data Processing e88

Homework 2: Vertical and Horizontal Scaling , Shared State Management

This document is a template for your solutions submission. You are free to add additional information in this submission if you would like. Extra screenshots and extra documentation are appreciated. Screenshots must always be viewable. If a screenshot is too blurry or chopped off in a key area you will not receive full credit for it.

**Make sure to also submit all your source code (.java files , .py files or whatever language you are using) - in a separate archive, named <LastName>\_<FirstName>\_HW2.zip**

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: |

**Problem 1: CPU Analysis** [points: 25]

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

|  |
| --- |
|  |

Provide your table or graphs demonstrating the results of running this code with 2, 4, and 16 threads on a 4 CPU machine: [5 points]

|  |
| --- |
|  |

What can you summarize about the results? [3 points]

|  |
| --- |
|  |

Provide your table or graphs demonstrating the results of running this code with 2, 4, and 16 threads on an 8 CPU machine: [5 points]

|  |
| --- |
|  |

What can you summarize about the results? How does a 4 CPU machine compare to an 8 CPU machine in this exercise? [2 points]

|  |
| --- |
|  |

**Problem 2: I/O Analysis** [points: 25]

Paste your source code into the following area. Make sure you clarify what you did to programmatically create an I/O intensive process. [10 points]

|  |
| --- |
|  |

Provide your table or graphs demonstrating the results of running this code with 2, 4 and 16 threads on a 4 CPU machine: [5 points]

|  |
| --- |
|  |

What can you summarize about the results? [3 points]

|  |
| --- |
|  |

Provide your table or graphs demonstrating the results of running this code with 2, 4 and 16 threads on an 8 CPU machine: [5 points]

|  |
| --- |
|  |

What can you summarize about the results? How does a 4 CPU machine compare to an 8 CPU machine in this exercise? [2 points]

|  |
| --- |
|  |

**Problem 3: unique counts** [points: 25]

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

|  |
| --- |
|  |

Explain your choice of the data structures for shared state management [5 points]

|  |
| --- |
|  |

What are the results of your queries for the following specified keys? [10 points]

The expected output for the first value is provided for your reference.

|  |
| --- |
| **Query 1:**  **<date\_hour>, <url\_count>**  2019-09-12:13, 185  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, 1  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, 1  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 4: time range queries** [points: 25]

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

|  |
| --- |
|  |

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

|  |
| --- |
|  |

What are the results of your query for the specified keys ? [5 points]

The expected output for the first value is provided for your reference.

|  |
| --- |
| <date,hour,country>, url\_count  2018-09-13:19,IQ, 1  2018-09-13:19,IR, ??  2018-09-13:19,IS, ??  2018-09-13:19,IT, ??  2018-09-13:19,JE, ?? |

**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]

|  |
| --- |
|  |

What are the results of your query? [5 points] The expected 5 values for 9/12 are provided, please fill in the values for avg TTFB and the URLs for 9/13 and 9/14.

|  |
| --- |
| Date URL Average\_TTFB  9/12/19 <http://example.com/?url=114> 0.393101408  9/12/19 <http://example.com/?url=101> 0.402545  9/12/19 <http://example.com/?url=133> 0.413317187  9/12/19 <http://example.com/?url=033> 0.418867857  9/12/19 <http://example.com/?url=157> 0.419289394  9/13/19  9/13/19  9/13/19  9/13/19  9/13/19  9/14/19  9/14/19  9/14/19  9/14/19  9/14/19 |