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

<Date>

Harvard Extension - Big Data Principles e88

Homework 9: Streaming Processing with Spark

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

|  |
| --- |
| Problem 1:  Problem 2:  Problem 3:  Problem 4:  Bonus Problem 5: |

**Problem 1: [25 points]**

Paste most relevant source code of your Spark Streaming job into the following area.[5 points]

|  |
| --- |
|  |

Demo your Kafka producer sending events to the Kafka cluster [10 points]

|  |
| --- |
|  |

Show your Spark job's output for a few batch windows, with the results of the job - counts of the number of clicks per URL per batch and running total [10 points]

|  |
| --- |
|  |

**Problem 2: [25 points]**

Paste the most relevant source code of your job into the following area [10 points]

|  |
| --- |
|  |

Show a few events that your Kafka producer is sending to the Kafka cluster [5 points]

|  |
| --- |
|  |

Show your job's console output, as well as the results of the job [10 points]

|  |
| --- |
|  |

**Problem 3: [35 points]**

Paste the most relevant source code of your job [5 points]

|  |
| --- |
|  |

Show a few events that your Kafka producer is sending to the Kafka cluster [5 points]

|  |
| --- |
|  |

Show your job's console output, as well as the results of the job - number of unique users per 30 sec windows using Spark aggregation methods [5 points]

|  |
| --- |
|  |

Show your job's console output, as well as the results of the job - number of unique users per 30 sec windows using HyperLogLog algorithm; run HLL job with a few (2-3) different accuracy settings [10 points]

|  |
| --- |
|  |

Compare and explain the results of HLL with different accuracy settings and the regular aggregation results [5 points]

|  |
| --- |
|  |

Compare the execution flow (number of jobs) for HLL with different accuracy settings - explain the results. [5 points]

|  |
| --- |
|  |

**Problem 4: [15 points]**

Draw the state of the counters array after adding s4 element. Show all steps of your calculations [5 points]

|  |
| --- |
|  |

Draw the state of the counters array after adding s5 element. Show all steps of your calculations [5 points]

|  |
| --- |
|  |

Explain the results and what could be done to improve (if needed) them [5 points]

|  |
| --- |
|  |

**Bonus Problem 4: [+20 points]**

Paste most relevant source code of your job [10 points]

|  |
| --- |
|  |

Show a few events that your Kafka producer is sending to the Kafka cluster [2 points]

|  |
| --- |
|  |

Show results of the job - demonstrate how events are de-duped within the window, but not de-duped outside of the window [8 points]

|  |
| --- |
|  |