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

<Date>

Harvard Extension - Big Data Principles e88

Homework 7: Flume

* **Make sure you submit your solution document as a separate file in Canvas**
* **submit all your source code/ config files (if any) in a separate archive, named <LastName>\_<FirstName>\_HW7.zip**

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

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

**Problem 1: Simple Flume Setup [25 points]**

Paste your Flume config file below [5 points]

|  |
| --- |
|  |

Show content of your source and sink directories before and after the Flume run. Compare the content by showing the line count for all files ('wc -l') These can be screen shots (ideally), or a copy/paste of console text. [5 points]

|  |
| --- |
|  |

Paste the last 10 lines of the flume log file [5 points]

|  |
| --- |
|  |

Demo results of Experiment 1 - content of Flume's source and sink dirs, and relevant lines from its log file [5 points]

|  |
| --- |
|  |

Demo results of Experiment 2 - content of Flume's source and sink dirs, and relevant lines from its log file [5 points]

|  |
| --- |
|  |

**Problem 2: Memory Channel Analysis [25 points]**

Paste your web server's index.html files below; demo access to this page - screenshot [1 point]

|  |
| --- |
|  |

Show corresponding entry in the web server log file [1 point]

|  |
| --- |
|  |

Paste content of your script/program that generates requests to your web server via 'curl' at a specified rate [1 point]

|  |
| --- |
|  |

Paste your Flume config file below [2 points]

|  |
| --- |
|  |

Demo your Flume topology working - compare content of the web server log with the resulting files in the sink directory [5 points]

|  |
| --- |
|  |

Demo results of the Experiment 1 - content of the web server logs (number of lines), Flume sink dir, relevant lines from Flume's log file [5 points]

|  |
| --- |
|  |

Demo results of the Experiment 2 - content of the web server logs (number of lines), Flume sink dir, relevant lines from Flume's log file; explain the results [5 points]

|  |
| --- |
|  |

Demo results of the Experiment 3 - content of the web server logs (number of lines), Flume sink dir, relevant lines from Flume's log file - at different event generation rates; explain the results [5 points]

|  |
| --- |
|  |

**Problem 3: File Channel Analysis [25 points]**

Paste your Flume config file below [2 points]

|  |
| --- |
|  |

Demo your Flume topology working - compare content of the web server log with the resulting files in the sink directory [3 points]

|  |
| --- |
|  |

Demo results of the Experiment 1 - content of the web server logs (number of lines), Flume sink dir, relevant lines from Flume's log file and content of the 'checkpointDir' and 'dataDir'; explain the results; [5 points]

|  |
| --- |
|  |

What do you need to do to save more events? [2 points]

|  |
| --- |
|  |

Demo results of the Experiment 2 - content of the web server logs (number of lines), Flume sink dir, relevant lines from Flume's log file and content of the 'checkpointDir' and 'dataDir'; explain any difference in behavior you may see with Experiment 1 [5 points]

|  |
| --- |
|  |

Demo results of the Experiment 3 - with different event generation rates and 'capacity' settings of the channel; show relevant changes in the content of the 'checkpointDir' and 'dataDir', Flume sink dir, relevant lines from Flume's log file; explain the results [5 points]

|  |
| --- |
|  |

At what point do you start seeing lost data, if any? [3 points]

|  |
| --- |
|  |

**Problem 4: Flume with UUID interceptor [25 points]**

Paste your Flume config file below [10 points]

|  |
| --- |
|  |

Demo that the resulting events in the sink now have UUIDs [15 points]

|  |
| --- |
|  |

**Problem 5: Bonus: Flume with HDFS sink [+15 points]**

Paste your Flume config file below

|  |
| --- |
|  |

Demo your Flume topology working at a rate of 100 requests/events per second - compare content of the web server log with the resulting files in the sink HDFS directories

|  |
| --- |
|  |

Demo that correct events are stored into the correct corresponding HDFS folders/files

|  |
| --- |
|  |