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

<Date>

Harvard University Extension - Principles of Big Data Processing e88

Homework 6: **Batch Views with HBase and NoSQL**

* **Make sure you submit your solution document as a separate file in Canvas**
* **submit all your source code in a separate archive, named <LastName>\_<FirstName>\_HW6.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: AWS EMR with HBase** **[points: 25]**

show version and status info from your HBase [5 points]

|  |
| --- |
|  |

include screenshots of your 'vet\_visits\_hw6' table creation, and all commands to insert and query data. Explain the difference between 'scan' and 'get' commands [15 points]

|  |
| --- |
|  |

Show results of the 'describe' table command [5 points]

|  |
| --- |
|  |

**Problem 2: Batch Views with HBase [points: 25]**

Show your HBase tables created for the Batch Views. Explain your design. [10 points]

|  |
| --- |
|  |

Show the data you've inserted [5 points]

|  |
| --- |
|  |

Show execution and results of running your 5 queries in the HBase CLI - for your selected parameters [10 points]

|  |
| --- |
|  |

**Problem 3: [points: 25] Spark with HBase - writes**

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

|  |
| --- |
|  |

Show execution and console output of your job - include a couple of screenshots [10 points]

|  |
| --- |
|  |

Show execution of the Query 2 from the HBase CLI and results for the specified keys [10 points]

|  |
| --- |
| 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, ?? |

**Problem 4: Research of NoSQL DBs [points: 25]**

Fill in as much information about the Techniques, Functional and Non-functional Requirements of your chosen NoSQL DB you were able to identify. Use the same format as in the original Gessert's research: [points: 15]

|  |
| --- |
| https://lh4.googleusercontent.com/89-WkekSnPNvIeYV5z2fqEayNe-M1B7ocKcOYRGfhqMwEafPTKBa3pmhmTBL4Xe1CfyFHzjPPaCi3IU9vrZGVGK6qkT2UVsOxHl3AsJ7Kmh7ucPg41Ahb1KCy965xcqRwwOGfe5y |

Explain your DB decisions/choices and what information sources you have used [10 points]

|  |
| --- |
|  |

**Problem 5: Bonus: [points: 15] Spark with HBase - writes**

Show screenshots of the EMR cluster with both Spark and HBase installed [5 points]

|  |
| --- |
|  |

Show execution and console output of the Spark job running on the EMR cluster [5 points]

|  |
| --- |
|  |

Show results of the Query1 - retrieved by your job [5 points]

|  |
| --- |
|  |