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

<Date>

Harvard University Extension - Principles of Big Data Processing e88

Homework 5: **Spark Batch processing and Parquet**

* **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>\_HW5.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 B1: Bonus: x% complete |

**Problem 1: unique counts** [points: 40]

Paste source code of your Spark 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 of the summary of each job (see example) [10 points]

|  |
| --- |
|  |

show execution stages and input/output sizes for each job [10 points]

|  |
| --- |
|  |

Show results of your jobs for the specified keys [10 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 2: Spark with Parquet [points: 20]**

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

|  |
| --- |
|  |

Include screenshots of your job execution and listing of the resulting Parquet files [5 points]

|  |
| --- |
|  |

Show execution and results of the parquet-tools 'meta' command. Show info about the Row groups and compression type from the meta info [10 points]

|  |
| --- |
|  |

**Problem 3: Spark Dataframe - time range queries [points: 20]**

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

|  |
| --- |
|  |

Include screenshots of your job execution [5 points]

|  |
| --- |
|  |

Show results of your query for the specified keys [10 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: SparkSQL [points: 20]**

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

|  |
| --- |
|  |

Include screenshots of your job execution and results for hour: 2019-09-12:14 [5 points]

|  |
| --- |
|  |

Show execution plan of the query and demo/explain whether it took advantage of the Parquet columnar format (or not) [10 points]

|  |
| --- |
|  |

**Problem B1: Bonus: Top N queries** [15 points]

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

|  |
| --- |
|  |

Explain your solution approach [5 points]

|  |
| --- |
|  |

Show results of your query [5 points]

|  |
| --- |
|  |