**Submission**

**The submission you need to generate for each task consists of:**

**A report documenting what you have done with justification and explanation. The report should address all criteria in the grading rubric and in the submission details described in each task. The report will be no longer than 4-5 pages, 11 point font.**

**A video demonstration of the use of your system to answer the required questions. For the queries in Group 2 and Question 3.2, it will suffice to illustrate the results for a small subset of queries. The video demo should be no longer than 5 minutes.**

**Further details are provided in the instructions for each task. For your Task 2 report, you will also be asked to include a general comparison of the stacks used in each task (Hadoop, Storm, Spark).**

*PDF Report*

*You must submit your report in PDF format. Your report should be no longer than 4-5 pages, 11 point font. Your report should include the following:*

*Give a brief overview of how you extracted and cleaned the data.*

*Give a brief overview of how you integrated each system.*

*What approaches and algorithms did you use to answer each question?*

*What are the results of each question? Use only the provided subset for questions from Group 2 and Question 3.2.*

*What system- or application-level optimizations (if any) did you employ?*

*Give your opinion about whether the results make sense and are useful in any way.*

*Video Demonstration Link*

*In your report, you will also need to submit a link to a video demonstration of your approach. Your video should be no more than 5 minutes long. Your video should include the following:*

*Ingesting and analyzing data for each question*

*Displaying/querying the results for each question*

*Record Video Demonstration*

*You can use Personal Capture from Illinois MediaSpace (AKA Kaltura) to record your video demonstration. To learn to use Personal Capture, see the instructions here. Or you can choose to use other recording software to capture your video demonstration.*

*Your video should be uploaded to Illinois MediaSpace by following the steps below:*

*Upload your video to Illinois MediaSpace. See instructions.*

*Publish your video to obtain a shareable link. See Instructions.*

*Include the shareable link for the video demonstration in your report.*

*Submit Task 1*

*Evaluation*

*Your peers will evaluate your submission based on the Task 1 Rubric. This assignment is worth 50 points. The evaluation period will begin immediately after you submit your assignment. You must evaluate at least 3 of your peers' submissions. We suggest you evaluate 5 of your peers.*

*Evaluate Task*

System integration: Is the description of the system integration clear enough?

Speed/efficiency: Is the system efficient? Does it produce results quickly? What optimizations are used?

Approaches: Are the techniques used to solve each question reasonable?

Quality of results: Are the results comprehensive, intuitive, and accurate?

Project Report: Report is informative and covers all questions in detail. Report clearly shows all of the following: 1) How data was extracted and cleaned; 2) What approaches were used to solve each question; 3) What results were generated.

Project Video: Video is informative and demonstrates all features/properties in detail. Video clearly shows both of the following: 1) Ingesting and analyzing data in the systems; 2) Querying results.

Speed/Efficiency: System is responsive and quick to generate results. Implementation utilizes 2 or more optimizations to increase speed and efficiency.

System Integration: Integration of systems is sufficient to generate results and is not redundant, and is efficient in all respects.

Quality of Results: Results are correct and include important fields for all questions.

Extracted and cleaned the data

I reviewed the three question groups and found that all results could be accurately and comprehensively obtained using the Airline On-Time Performance Data (<https://www.transtats.bts.gov/DatabaseInfo.asp?DB_ID=120>).

The data was available from the ESB Snapshot as zip files under the read-only folders, one per year-month. The data extraction was performed as four, 5-year batches for cost optimization. Each batch was run on a single **free-tier** EC2 instance with access to both the ESB Snapshot and a locally Writable ESB Volume. The data extraction process was bash commands which mounted the snapshot then placed the unzipped files into a folder named *raw*. It was found that two files were unusable (HTML not CSV) and I confirmed on Piazza that neither those files nor the data which would have been contained within them were required for this project.

I did spot checks of the unzipped files to inspect the schema and its consistency. Two of the fields (*OriginCityName*, *DestCityName*) used quoted identifiers (double quotes [ASCII 34]). Those fields were non-atomic (violate first normal form [1NF]) and redundant (*OriginCityName* + ", " + *OriginState*, *DestCityName* + ", " + *DestState*). There also existed two different record schemas. The record schema before 1991-Oct differed from the later record schema by having twenty additional NULLable fields but as such they were still compatible schemas.

File record cleaning was limited to the elimination of quoted identifiers which was only done as an optimization. It was performed in parallel across the same four EC2 instances. The process was bash commands using sed to first find and replace, comma space, with hyphen, then to find and replace, double quotes, with nothing.

All cleaned files were uploaded to a single S3 bucket under a folder named *scrubbed*.

In Athena, I defined a table named *scrubbed* over the folder using *org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe* which had the quoted identifiers not been removed, it would od had to use a slower deserializer (*OpenCSVSerDe*).

I decided to use a single table on which all queries would be performed against so that answers are more intuitive by the nature of results are transferable (all from a single source). To ensure answers are accurate, I removed data for any flights which were cancelled or diverted.

What is the departure delay of a canceled flight?

What is the arrival delay of a diverted flight?

Will diverted/cancelled flights count as flights to/from an airport?

The *completedflights* table was based on the *scrubbed* table. It was partitioned by year as a performance optimization. It was trimmed to only eleven columns also as a performance optimization. The actual and Computer Reservation System (CRS) arrival/departure columns were converted to timestamps to make the SQL more intuitive.

Integrated each system

Results of each question

System-level or application-level optimizations

Opinion about the results

Video Demonstration Link